D3.1 State of the art report on affective technologies

Date: 11/02/2013

VERSION CONTROL

Project: ITEA2 Empathic Products	
Acces: **Restricted to the Consortium/ Public**	Status: **Final V1.0.1**
Editor: E. Marilly – (ALBLF) Contributors: Lille 1 (M. Bilasco), INABENSA (E. Zubillaga), VTT (S.M. Mäkelä, M. Lindholm, J. Miettinen, J. Plomp,), SoftKinetic (I. Ravyse), LORIA (S. Cruz-Lara, A. Denis), Delicode (J. Tuomisto), ALBLF (E. Marilly, A. Gonguet), TECNALIA (M. Montejo), FADO (F. Reis), TSP (T. Zaharia), City Passenger (V. Renardias), Noldus (N. van der Aa), VicarVision (P. Lewinski), U. Hasselt (J. Van den Bergh)	
Document:	
Reviewed: S. Cruz-Lara (LORIA)	
Approved:	
Date: **2013/02/11**	
Comments: **This document is the State of the art report on affective technologies.**	

This document contains **97** pages in edition **1**

Contents

1. Executive summary

This document is a state of the art on affective technologies. Three sections have been defined: sensing technologies, data fusion and interaction modelling and behaviour modelling.
A palette of sensors as large as possible has been considered, including (mono and multi-camera) vision-based, audio/speech systems, textual information, touch sensors, physiological and biosignals (HR/HRV, GSR, EEG).
As key technologies, let us cite:

- Facial expression analysers, gesture recognition and interpretation, human body pose estimation, physiological affective wearables
- Speech recognition, emotion interpretation from audio cues,
- Textual analysers,
- Emotion analysis from interaction devices (keyboard, mouse, focus of attention on screen....)

2. Introduction

Affective technologies are trying to assign computers the human-like capabilities of observation, interpretation, and generation of affect features. This document tries to give a state of the art on affective technologies used in the Empathic Products project. Three sections have been defined: sensing technologies, data fusion and interaction modelling and behaviour modelling. The document includes an annex listing the available technologies provided by the project partners.

3. Sensing technologies for affective computing

Whole set of independent affective analysers, associated with the analysis, the interpretation and/or the understanding of signals provided by individual sensors.

3.1 Video and depth sensor Analyzers

3.1.1 Facial expression analysers and interpretation

3.1.1.1 Facial expression

Facial expression recognition aims to detect different expression types caused by combination of facial muscle movements. After the pioneering work of [Ekm78], it becomes a hot research topic in pattern recognition and computer vision domains.
Automatic facial expression recognition algorithms consist of face detection, facial feature extraction, pre and post normalizations and finally classification steps. The six basic classes (angers, happiness, sadness, surprise, fear and disgust) are widely used by researchers.
Figure 1 presents common facial expression recognition framework.

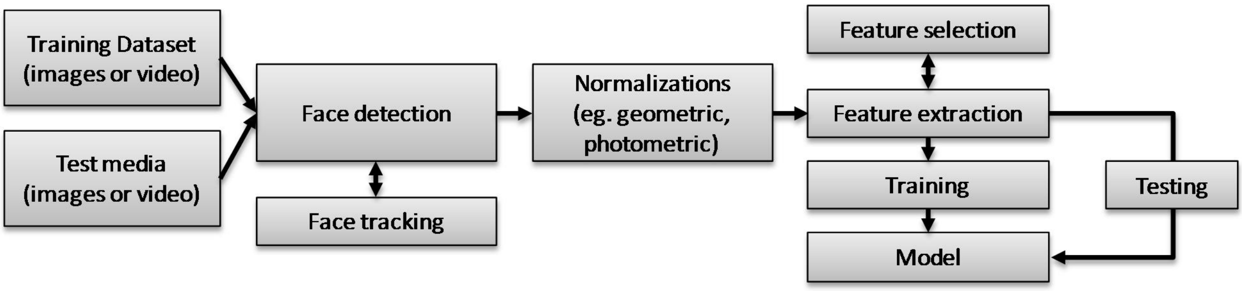

Training Dataset (images or video)
Test media (images or video)
Face detection
Face tracking
Normalizations (eg. geometric, photometric)
Feature selection
Feature extraction
Training
Model
Testing

Figure 1. Generalized facial expression recognition framework

In the training phase, face detection provides elimination of the background from the face. In the case of video files, a face tracker also supports face detection. In the next step, normalization takes place where all the images are normalized using geometric and photometric normalization techniques.

3.1.1.1.1 Normalizations

Geometric techniques provide DOF (depth of freedom) corrections such as roll yaw and pitch corrections as seen on Figure 2 [Mur09]. In addition a similar distance metric (e.g. eye distance, eye to mouth distance) is used to normalize the face to a unique size.

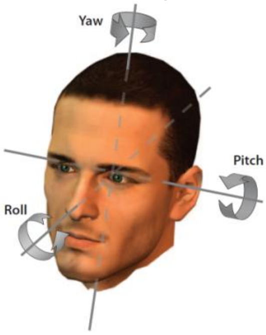


Figure 2 Three DOF for egocentric rotation angles [Murphy-Chutorian 2009]

Photometric techniques aim to eliminate light and illumination related defects from the facial area (e.g. Histogram equalization). Due to the non-uniform distribution of the visible light on the facial area, infrared are also employed. Figure 3 and Figure 4 shows a combination of geometric and photometric normalization step used by [Dan12]

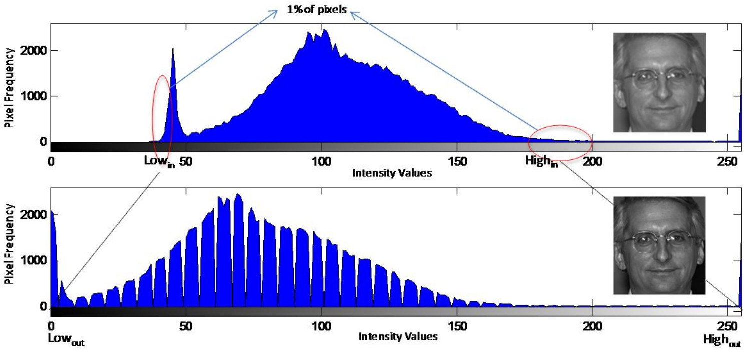


Figure 3 Contrast stretching on face image

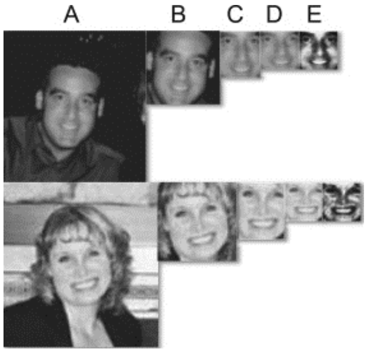


Figure 4 Preprocessing steps for the sample images from GENKI dataset. (A) Original image. (B) OpenCV face detection output. (C) Rotating and reshaping the face dimensions considering IPD distance. (D) Normalizing face size to 50×50. (E) Histogram equalization.

3.1.1.1.2 Feature Extraction & Feature selection

Facial expressive features can be extracted by using:

- Geometric feature extraction methods (e.g. Deformable models, Active Shape Models (ASM))
- Texture extraction methods (e.g. Gabor wavelet transforms, Local binary patterns LBP)
- Motion flow algorithms (e.g. optical flow)
- Pixel intensities (e.g. gray level pixel values)

In addition, vertical and horizontal projection functions and second order derivatives are widely used for the facial feature extraction process.

Feature selection is a dimension reduction technique that selects a subset of relevant information from the original data by eliminating redundant information. PCA and Nonnegative Matrix Factorization (NMF) [Zil09] and Adaboost [Fre96] are used for dimensionality reduction. Feature subset selection method is also used by [Dor11] for facial expression recognition.

3.1.1.1.3 Facial Expression Classification

Efficient and non-intrusive detection of emotions are an important step towards natural human machine interaction systems. In the literature, Bayesian networks, neural networks, support vector machines are widely used classifiers in facial expression recognition domain. The following classifiers are commonly used for facial expression recognition researches:

- Neural networks
- Support vector machines (linear and RBF kernels)
- Hidden markov models
- Bayesian networks
- kNN

The following table shows general features of the aforementioned feature extraction and classification methodologies.

Table 1 Summary of facial expression recognition systems

Facial feature	Classification	Pros	Cons
Active Appearance Models [Yeo09]	Hausdorff distance + KNN	A prior knowledge is learned through observation of both shape and texture variations in a given training set	His processing time, extensive texture data storage requirements
Active Shape Models[Coo95]	-	Produce small set of facial features. Resistant to	Incomplete or loss of a facial point may lead to less accurate

		illumination differences.	results.
Gabor wavelets[Bas08] [Kot08]	LVQ (Learning vector quantization) SVM, MLP	Insensitive to light changes and individual face differences.	Computationally complex and requires support from feature selection and dimension reduction algorithms.
Local Binary Patterns (LBP)[Val11], [Zao07]	kNN	Insensitive to light changes	Binary data computation is sensitive to noise. Difficult for large-scale datasets and requires dimension reduction
Optical flow[Yac96]		Small set of features	Sensitive to rapid illumination changes and non-rigid facial movement.
Active Appearance Models [Ash07]	SVM+linear kernel	Good generalization ability	
Pixel intensities[Fas04]	PLSA	direct clustering aspects in an unsupervised manner	
Motion history[Val06]	KNN, SNoW	Overcomes problem of motion self occlusions	
Haar features [Whi06]	Adaboost		limited discriminant capability, they can only represent the regular rectangular shapes
Frontal and Profile facial points, [Pan04]	Rule based classifier	Automatic AU coding	Not real time

3.1.1.1.4 Expression presentation

Detected expressions can be presented in several different graphical presentations. Figure 5, Figure 6 and Figure 7 presents graph based presentations of facial expressions.

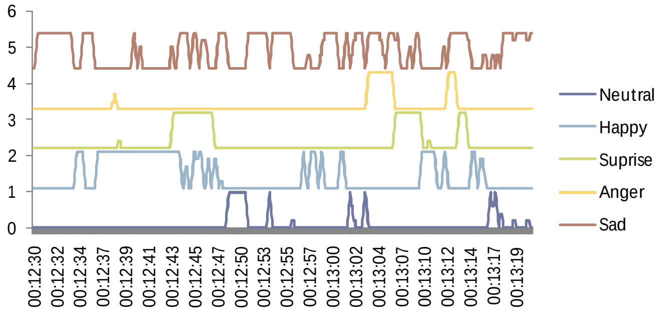


Figure 5 Emotion graph of a video session

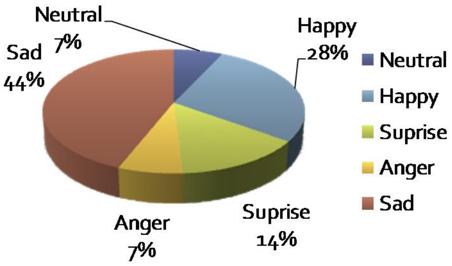


Figure 6 Cumulative emotion pie chart

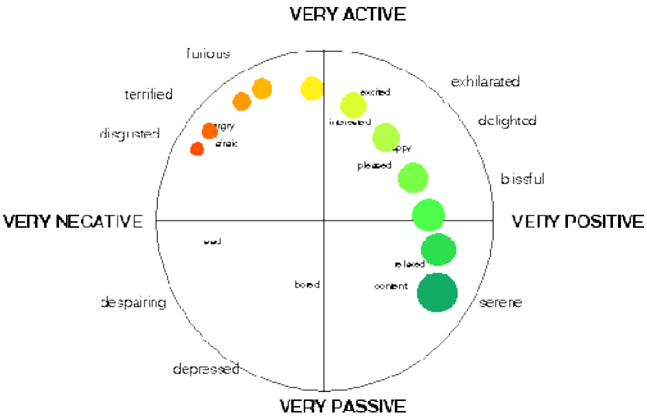


Figure 7: Example of a FEELTRACE [Cow03] display during a tracking session for recording how perceived emotion develops over time

3.1.1.1.5 Facial Expression datasets

The following table lists available facial expression datasets.

Table 2 Facial expression databases

Name	#of subjects	#of images	Face Pose	Expressions
FERET	10465	14051	Mixed	Mixed
Cohn-Kanade AU-Coded Facial Expression Database	100	486	Frontal	AU combinations
MMI Database	43	1280 videos	Frontal	AU combinations
Japanese Female Facial Expression (JAFFE) Database	10	213 images	Frontal	7 expressions
Belfast Naturalistic Database	125	>250 video	Mixed	Mixed
CAS-PEAL Database	1040	30900	Mixed	Mixed
FG-NET Database	19	399 video	Frontal	7 emotions
Bosphorus DB	105	4666 images	Pitch, yaw rotations	AU coded 35 expressions per subject

Figure 8 presents sample AU4 samples from Bosphorus Database

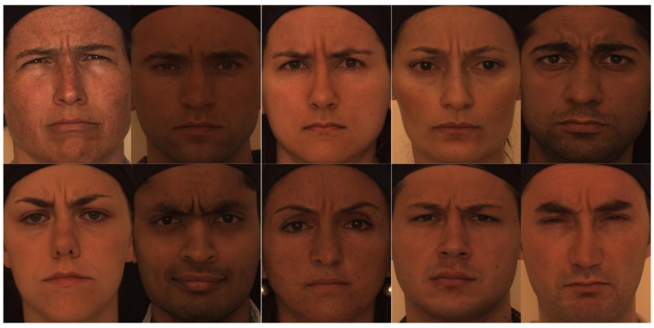


Figure 8 AU4 samples from Bosphorus dataset.

3.1.1.2 Noldus Approach

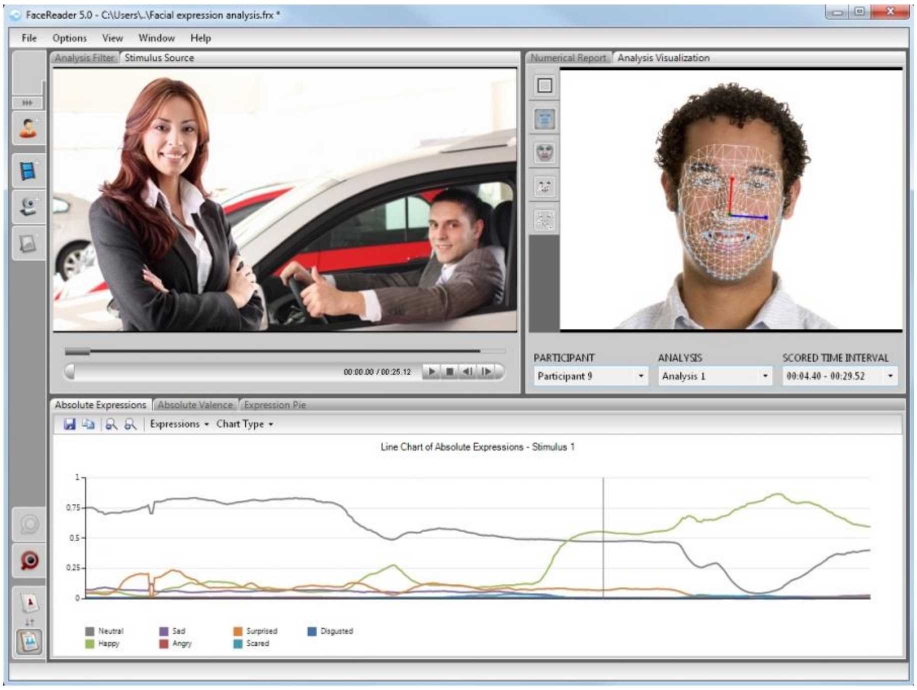
3.1.1.2.1 What a face can tell

Apart from being the means to identify other members of the species, the human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to regulate the conversation by gazing or nodding, and to interpret what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression [Vi04]. Personality, attractiveness, age and gender can also be seen from someone's face. Thus the face is a multi-signal sender/receiver capable of tremendous flexibility and specificity. In turn, automating the analysis of facial signals would be highly beneficial for fields as diverse as security, behavioral science, medicine, communication, education, and human-machine interaction.

3.1.1.2.2 How can FaceReading help?

In security contexts, apart from their relevance for person spotting and identification, facial signals play a crucial role in establishing or detracting from credibility. In medicine, facial signals are the direct means to identify when specific mental processes are occurring. In education, pupils' facial expressions inform the teacher of the need to adjust the instructional message. As far as natural interfaces between humans and machines (computers, robots, cars, etc.) are concerned, facial signals provide a way to communicate basic information about needs and demands to the machine. Where the user is looking (i.e., gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g., a wink) can be associated with certain commands (e.g., a mouse click) offering an alternative to traditional keyboard and mouse commands. The human ability to read emotions from someone's facial expressions is the basis of facial affect processing that can lead to expanding interfaces with emotional communication and, in turn, to obtaining a more flexible, adaptable, and natural interaction between humans and machines.

3.1.1.2.3 FaceReader 5.0



FaceReader 5.0 - C:\Users\..\Facial expression analysis.frx *

PARTICIPANT: Participant 9

ANALYSIS: Analysis 1

SCORED TIME INTERVAL: 00:04.40 - 00:29.52

00:00.00 / 00:25.12

Figure 9: *FaceReader 5.0 – analysis mode*

FaceReader is a program for facial analysis (Figure 9). It can detect emotional expressions in the face. It can identify six basic emotions: happy, sad, angry, surprised, scared, disgusted and a neutral state. Additionally, it can detect facial states (left and right eye open or closed, mouth open or closed and eyebrows raised, neutral or lowered), the test participant’s global gaze direction and track the head orientation. FaceReader can also indicate the person's gender, age, ethnicity, the amount of facial hair (beard and/or moustache) and whether the person is wearing glasses or not. The software can also identify the subject.

FaceReader data can be imported into The Observer XT, the leading software package for the collection, analysis and presentation of observational data from Noldus int.. This enables to integrate FaceReader data with other data, such as manually logged events, physiological data and eye tracker data and to analyse the full context. For instance, what user interface is the test participant looking at and what part triggers an emotion.

3.1.1.2.3.1 FaceReader 5.0 applications

FaceReader can be used in a wide range of research areas:

- **Psychology** — How do people respond to particular stimuli, e.g. in cognitive research.
- **Education** — Observing students' facial expressions can support the development of educational tools.
- **Human-computer interaction** — Facial expressions can provide valuable information about user experience.
- **Usability testing** — Emotional expressions can indicate the ease of use and efficiency of user interfaces.
- **Market research** — How do people respond to a new commercial's design?
- **Consumer behaviour** — How do participants in a sensory panel react to a presentation?

3.1.1.2.3.2 How does FaceReader 5.0 work?

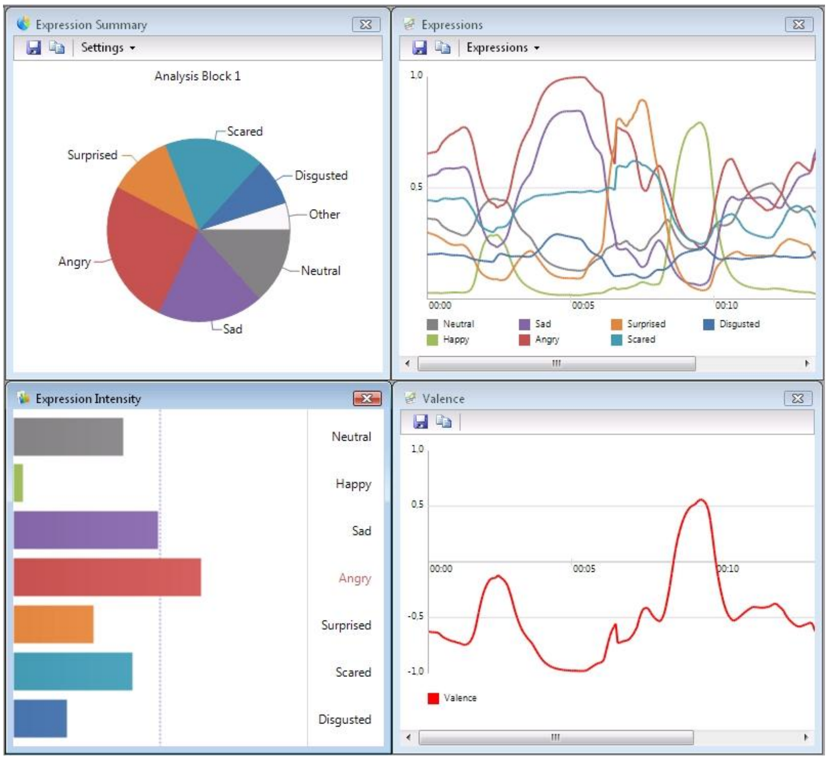


Figure 10: *FaceReader 5.0 – output, graphic presentation of the results.*

The main challenge in the analysis of facial expressions is how to deal with variance in pose/ orientation and lighting of the face. The solution that FaceReader uses is to classify faces in three consecutive steps. Please see reference [Ku05] - [Ku08].

3.1.1.2.3.2.1 Face finding

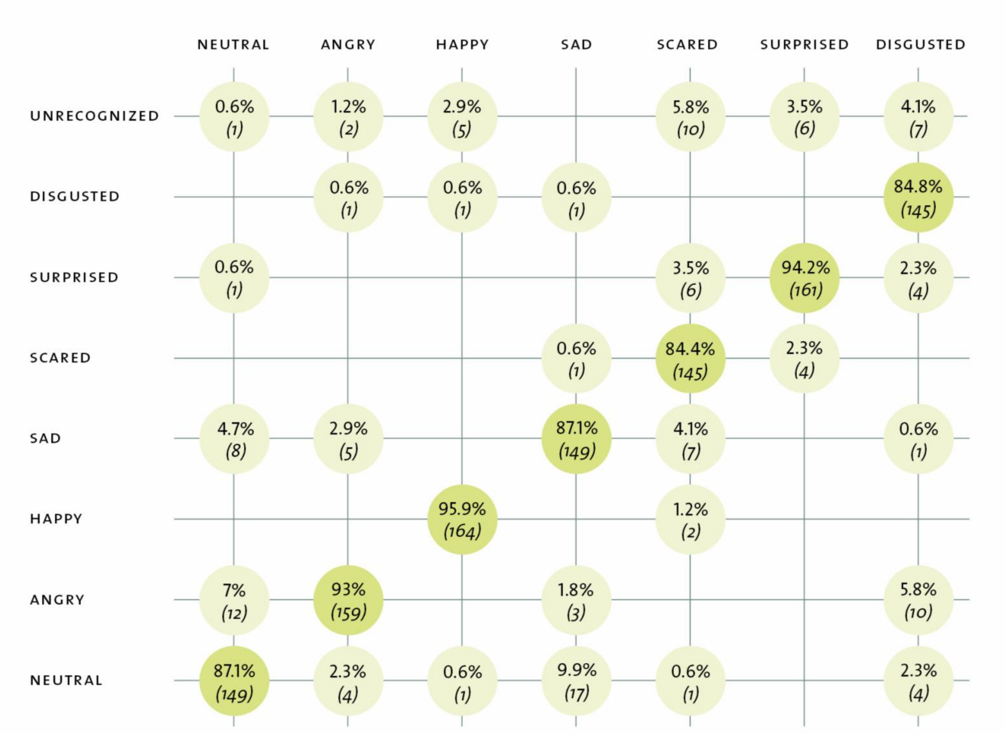
The position of the face in an image is found using a method called the Viola Jones cascaded classifier algorithm, which was developed for finding the face in images [Vi04].

3.1.1.2.3.2.2 Face modelling

In this step, a model-based method is used, called the Active Appearance Model (AAM) [Co00], to synthesize an artificial face model, which describes the location of 500 key points in the face and the facial texture of the area entangled by these points. The model uses a database of annotated images and calculates the main sources of variation found in the images. Principal Component Analysis compression is used to reduce the model dimensionality. New faces can then be described as deviations from the mean face, using a vector.

3.1.1.2.3.2.3 Face classification

The actual classification of the facial expressions is done by training an artificial neural network [Bi95], which takes the above vector as input. As training material over 10000 manually annotated images were used. The network was trained to classify the six basic or universal emotions described by Ekman [Ek70] happy, sad, angry, surprised, scared, disgusted and a neutral state. FaceReader can recognize facial expressions with an accuracy of 90%. For some emotions, the accuracy is higher, for others lower (Figure 11). The results might be presented in graphic form (Figure 10)



	NEUTRAL	ANGRY	HAPPY	SAD	SCARED	SURPRISED	DISGUSTED
UNRECOGNIZED	0.6% (1)	1.2% (2)	2.9% (5)		5.8% (10)	3.5% (6)	4.1% (7)
DISGUSTED		0.6% (1)	0.6% (1)	0.6% (1)			84.8% (145)
SURPRISED	0.6% (1)				3.5% (6)	94.2% (161)	2.3% (4)
SCARED				0.6% (1)	84.4% (145)	2.3% (4)	
SAD	4.7% (8)	2.9% (5)		87.1% (149)	4.1% (7)		0.6% (1)
HAPPY			95.9% (164)		1.2% (2)		
ANGRY	7% (12)	93% (159)		1.8% (3)			5.8% (10)
NEUTRAL	87.1% (149)	2.3% (4)	0.6% (1)	9.9% (17)	0.6% (1)		2.3% (4)

Figure 11: *Horizontally: emotional expressions scored manually by the annotators of the Radboud Faces Database .[Bi11] [La10]. Vertically: expressions scored by FaceReader (v. 4)* .[Ku05] [Hy08].

3.1.1.2.3.3 Limitations of the current version

The current FaceReader version has a number of limitations. It is good to keep these in mind when you start working with FaceReader.

- FaceReader is currently not trained to work with very young children, below the age of 3.
- FaceReader 5 is not yet trained for analysis of children from East Asia and South-East Asia. FaceReader 5 works well with other children and East Asian and South-east Asian adults.
- Pose, movement and rotation of the test person are limited. The test person should stand or sit and look frontally into the camera (angle < 40°)
- FaceReader requires strict light conditions.
- The face should not be partially hidden, for instance by a hat or very heavy facial hair. It is also very difficult to classify a person's facial expressions when he/she is eating, because the person's hand covers part of the face when he/she puts food in the mouth and the muscles in the face move.

3.1.1.2.3.4 Add-on modules

You can extend the base functionality of FaceReader with a number of add-on modules:

3.1.1.2.3.4.1 Project Analysis module

This module allows you to make a selection out of your participants and carry out a group analysis. You can for example analyze the facial expressions of all male and all female participants.

With this module it is also possible to mark events in the analysis, for example to mark the occurrence of a stimulus like a smell or sound. In addition, you can synchronize and visualize the video tests participants have watched. You can mark parts of interests within these videos and compare the data associated with these video fragments with each other, or with other parts of the analysis.

3.1.1.2.3.4.2 Action Unit Module

With the Action Unit Module you can analyze the intensity of a number of Action Units from the Facial Action Coding Scheme. FaceReader can analyze 19 Action Units, that are most common in facial expressions.

3.1.1.2.3.4.3 Improved API

The API (Application Programming Interface) which was already available for FaceReader 4, has been improved. The API allows you to send and receive information to/from other applications. It is now possible to (1) send both state and detailed log data simultaneously; (2) send out additional information such as image quality, coordinates of a number of key points in the face and action units; (3) send information about stimuli and event markers to other applications and (4) receive triggers from other applications.

3.1.1.2.3.5 System Requirements

Prior to installation of FaceReader the computer must meet the minimal system requirements for running the software.

3.1.1.2.3.5.1 Operating system - Microsoft Windows

FaceReader has been thoroughly tested using a US English version of Windows 7 (32 or 64 bit Professional edition). Like any software package, it is possible that minor differences in the operating systems of certain local language versions may affect how well FaceReader runs.

3.1.1.2.3.5.2 Computer – Workstation

Below are the specification o the workstation on which the FaceReader 5.0 was tested. It is possible to use it on a hardware that has similar (or better) components:

Technical specifications Dell Precision™ T3500 Workstation (or its successor)

- Processor: Intel®Xeon Quad Core CPU, 2.8 GHz
- Internal memory: 6 GB

Technical specifications Dell Precision™ M4600 laptop (or its successor)

- Processor: Intel®i7 Quad Core CPU, 2.2 GHz
- Internal memory: 4 GB

Professional workstation is recommended to be used. It is possible to buy consumer-range computers with a high processor speed and plenty of memory, but in order to remain competitive regarding price, the manufacturers often economize on the underlying system architecture. That means those computers are suitable for home use, but not for running professional scientific software. You should select a computer which is intended for professional use or labeled by the manufacturer as a workstation.

3.1.1.2.3.5.3 Camera

CCD webcam with a resolution of at least 640 x 480 pixels. We strongly recommend that you use a high-quality webcam. Simple web cams are not suitable. You can also use an IP camera or convert a webcam into an IP camera. If you choose the latter option, you need a program like webcamXP.

3.1.1.2.3.5.4 Internet connection

If you are using an IP camera or converted a webcam into an IP camera and access it with internet, you need a fast internet connection for these set-ups to work. In the case of a converted webcam, the internet connection for both computers must be fast.

3.1.2 Gaze detection & head orientation

3.1.2.1 Gaze detection

Gaze detection is to locate the position where a user is looking. When we want to exploit eye movements in human-technology interaction, we are interested in the focus of the gaze. Thus, we should know the user's perceived image at each point in time.

Eye tracking has been referred as having "promising" potential to enhance human computer interaction already for about 20 years. Nowadays the eye-gaze tracking has become more popular in human-computer interfaces and it has been shown to be useful in diverse applications. The eye-gaze tracking is the process of measuring either the point of gaze or the motion of an eye relative to the head. Eye tracking is for measuring eye positions and eye movement. One of the recent improvements, consequent of increasing computing power and improving camera optics, is that eye trackers are moving towards using large-field-of-view cameras (Ver2002, LC2012, Tobii2012) instead of focusing on the camera image of the eye only.

There are many areas that benefit from eye tracking systems. Specific applications include these systems in language reading, music reading, human activity recognition, the perception of advertising, playing of sport, HCI (especially for handicap people suffering from diseases), medical research and other areas [Orman2011] .

One of the most promising applications of eye tracking research is in the field of automotive design. Research is currently underway to integrate eye tracking cameras into automobiles. The goal of this endeavour is to provide the vehicle with the capacity to assess in real-time the visual behaviour of the driver. The National Highway Traffic Safety Administration (NHTSA) estimates that drowsiness is the primary causal factor in 100,000 police-reported accidents per year. Another NHTSA study suggests that 80% of collisions occur within three seconds of a distraction. By equipping automobiles with the ability to monitor drowsiness, inattention, and cognitive engagement driving safety could be dramatically enhanced. Lexus claims to have equipped its LS 460 with the first driver monitor system in 2006, providing a warning if the driver takes his or her eye off the road [Orman2011].

Current eye trackers require a calibration routine to be performed before they are able to detect user's point of gaze. Through calibration, the tracker is taught the individual characteristics of each user's eyes. How the eyes are positioned when different parts of the screen are being looked at. The calibration is performed by requesting the user to follow the reference points appearing on the screen, in five to 17 (Don2005) different positions. Some techniques have managed to decrease the number of points needed to two. Most trackers need to be calibrated at the beginning of each session, since the accuracy of the calibration usually decrease during the session, often the routine has to be done repeatedly every few minutes. The need for calibration is one of the issue that should be given extra attention. Some trackers (Tobii2012) support persistent calibration, in which case the calibration has to be performed only once, when the tracker is used for the first time.

Moving eyes is a natural process, adding a little conscious effort we can develop a new means of input technique, that can interpret effectively and quickly. Eye gaze interaction can provide a convenient and natural addition to user computer dialogues providing us the fastest means of communication with an eye blink. It is a reasonable addition to computer interaction as it is convenient in situations where it is important to use the hands for other tasks, contributing in human life for multitasking. It is particularly beneficial for the larger screen workspaces, virtual environments and controlling distant devices that can only perceived by the senses. An important side benefit is that eye position indicates the area of user's attention.

Eye and gaze tracking have a long history, but it is only relatively recently that gaze trackers have become sufficiently robust for use outside laboratories [Hansen2009]. The expensiveness of eye tracking devices from the fact that the volume of devices purchased is marginal at the moment, leaving the price dominated by development costs.

3.1.2.1.1 Hierarchy of systems

Duchowski provides a hierarchy in which he divides "eye tracking systems" into diagnostic and interactive systems as depicted in Figure 1 [DUCHO2002]. Diagnostic applications analyse eye movements to gather information on the person's cognitive processes and attention when performing a specific task, such as web page browsing, reading a newspaper or driving. This analysis is usually carried out a posteriori, and therefore the eye movements do not have an effect on the scene observed. Diagnostic applications are widely employed in different fields, such usability research, marketing and psychology.

In interactive systems, gaze is used in real time. Duchowski divides interactive systems into selective and gaze contingent systems. Selective systems are defined as those in which the point of gaze is used analogously to the mouse, as a pointing device. Gaze contingent systems exploit the user's gaze for rendering complex displays, which further divides between the screen based and model-based according to the technique used to accomplish the rendering. Interactive applications permit users to control the position of the mouse on the screen and the activation of items by their gaze, thus allowing highly impaired users with controlled eye movement to interact with the computer and their environment.

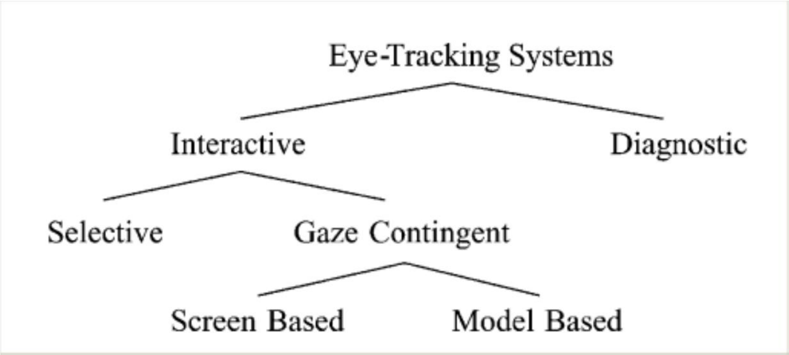


Figure 12. Hierarchy of eye tracking systems [DUCHO2002].

3.1.2.1.2 Visual gaze estimation

In order to detect and track users' gaze, it is necessary to employ a gaze tracking device which is able to determine the fixation point of a user on a screen from the position of her/his eye and her/his head pose. Earlier gaze trackers were very intrusive. In addition to require the user to be totally static, they were in direct contact with him by sticking a reflective white dot directly onto the eye or attaching a number of electrodes around the eye.

3.1.2.1.3 Intrusive systems

Earlier gaze trackers were very intrusive. In addition to require the user to be totally static, they were in direct contact with him by sticking a reflective white dot directly onto the eye or attaching a number of electrodes around the eye. Nowadays, the most accurate gaze tracking systems consist of head mounted devices that allow detecting the direction of the gaze without having to cope with the pose of the user's head. These trackers are also intrusive; they consist of devices composed of 3 cameras mounted on a padded headband (2 eye cameras to allow binocular eye tracking with built-in light sources, and 1 camera to allow accurate tracking of the user's point of gaze).

3.1.2.1.4 Non-intrusive systems

Non-intrusive gaze tracking systems usually require a static camera recording the face of the user and detecting the direction of their gaze with respect to a known position. A basic gaze tracking system is composed of a static camera, a display device and software to provide an interface between them. The precision of the system can be improved in different ways, such as adding a specific light source (e.g. an infrared beam) in order to create reflections on the eye and produce more accurate tracking information.

Figure 12 shows a single-camera gaze tracker configuration, based the Pupil-Centre/Corneal-Reflection (PCCR) method to determine the gaze direction. The video camera is located below the computer screen, and monitors the subject's eyes. No attachment to the head is required, but the head still needs to be motionless.
A small low power infrared light emitting diode (LED) is embedded in the infrared camera and directed towards the eye. The LED generates the corneal reflection and causes the Bright pupil effect, which gives an enhanced image of the pupil. The centers of both the pupil and the corneal reflection are identified and located, and trigonometric calculations allow projecting the gaze point onto the image.

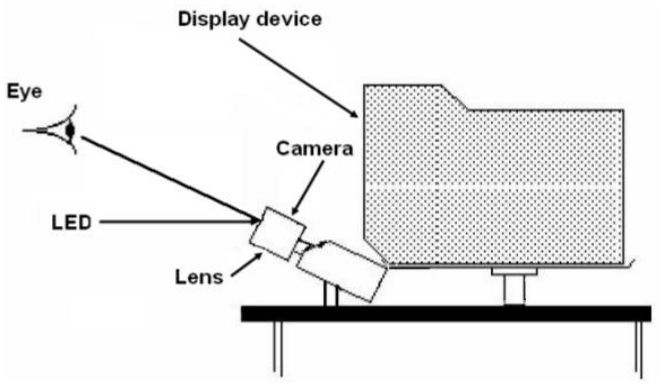


Figure 13:Schema of a gaze tracking system

3.1.2.1.5 Eye Tracking techniques

There are several techniques that can be used for monitoring eye movements. The techniques can be divided into three classes [Hyrsky2006]:

- those using electrooculography (EOG) techniques, which measure difference in electric skin around the eye,
- those requiring physical connection to the eye and the ones using
- non-contact camera-based methods.

In eye-based interaction, researchers at the moment almost invariably use camera-based techniques. If head movements are allowed, tracking only one element, for example pupil, is clearly not enough for obtaining the point of gaze. Most available remote eye gaze trackers have two characteristics that hinder them being widely used as the important computer input devices for human computer interaction. First, they have to be calibrated for each user individually; second, they have low tolerance for head movement and require the users to hold their heads unnaturally still [Zhu2007].

Many distinguishing features of the eye can be used to infer point-of-regard, such as corneal reflections (known as Purkinje images), the iris-sclera boundary, and the apparent pupil shape [Ducho2003].

Most commercial eye-tracking systems available today measure point-of-regard by the “corneal-reflection/pupil-centre” method [Gold2003]. These kinds of trackers usually consist of a standard desktop computer with an infrared camera mounted beneath (or next to) a display monitor, with image processing software to locate and identify the features of the eye used for tracking. In operation, infrared light from an LED embedded in the infrared camera is first directed into the eye to create strong reflections in target eye features to make them easier to track (infrared light is used to avoid dazzling the user with visible light). The light enters the retina and a large proportion of it is reflected back, making the pupil appear as a bright, well defined disc (known as the “bright pupil” effect). The corneal reflection (or first Purkinje image) is also generated by the infrared light, appearing as a small, but sharp glint (see Figure 2).

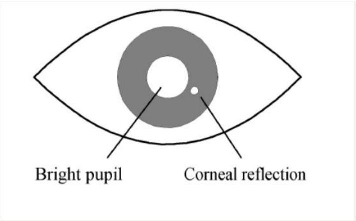


Figure 14. Corneal reflection and bright pupil as seen in the infrared camera image [Poole2005],.

Once the image processing software has identified the centre of the pupil and the location of the corneal reflection, the vector between them is measured, and, with further trigonometric calculations, point-of-regard can be found. Although it is possible to determine approximate point-of-regard by the corneal reflection alone (as shown in Figure 3), by tracking both features eye movements can, critically, be disassociated from head movements [Ducho2003], [Jac2003]).

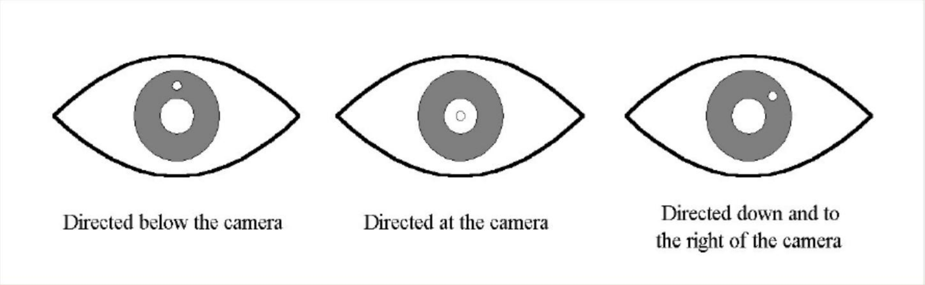

Directed below the camera
Directed at the camera
Directed down and to the right of the camera

Figure 15. Corneal reflection position changing according to point of regard [Poole2005],

Video-based eye trackers need to be fine-tuned to the particularities of each person's eye movements by a "calibration" process. This calibration works by displaying a dot on the screen, and if the eye fixes for longer than a certain threshold time and within a certain area, the system records that pupil-centre/corneal-reflection relationship as corresponding to a specific x, y coordinate on the screen. This is repeated over a 9 to 13 point grid-pattern to gain an accurate calibration over the whole screen [Gold2003].

Research in eye detection and tracking focuses on two areas: eye localization in the image and gaze estimation. There are three aspects of eye detection. One is to detect the existence of eyes, another is to accurately interpret eye positions in the images, and finally, for video images, the detected eyes are tracked from frame to frame. The eye position is commonly measured using the pupil or iris center. Gaze estimation is using the detected eyes in the images to estimate and track where a person is looking in 3D or, alternatively, determining the 3D line of sight. In the subsequent discussion, we will use the terms eye detection and gaze tracking to differentiate them, where eye detection represents eye localization in the image while gaze tracking means estimating gaze paths. They focuses on eye detection and gaze tracking in video-based eye trackers. A general overview of the components of eye and gaze trackers is shown in Figure 4. Video-oculography systems obtain information from one or more cameras (*Image data*). The eye location in the image is detected and is either used directly in the application or subsequently tracked over frames. Based on the information obtained from the eye region and possibly head pose, the direction of gaze can be estimated. This information is then used by gaze-based applications e.g. moving the cursor on the screen [Hansen2009].

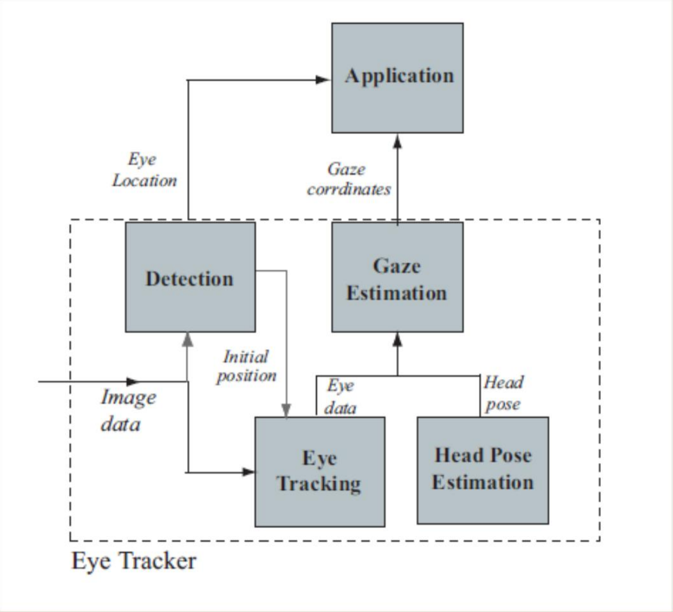


Figure 16. Components of video-based eye detection and gaze tracking [Hansen2009].

However, gaze tracking technologies are still not useful for a large part of society. New commercial applications, such as in video games and the automotive industry, would attract more companies and general interest in these systems, but several technical obstacles still need to be overcome. For instance, the image-processing task is still problematic in outdoor scenarios, in which rapid light variations can occur. In addition, the head position constraints of these systems considerably reduce the potential applications of this technology. The accuracy of gaze tracking systems is, to a large extent, compromised by head position since any head movement requires the system to readjust to preserve accuracy; i.e., gaze estimation accuracy can vary as the head moves. Gaze estimation is defined as the function of converting the image processing results (image features) into gaze (gaze direction/gazed point) using a mathematical equation. The usual procedure in any gaze tracking session is to first perform a calibration of the user. The calibration consists of asking the user to fixate on a set of known points on the screen. Calibration adapts the gaze estimation function to the user's position and system configuration. The mathematical method (gaze estimation function) used determines the dependence of the system accuracy on head position, i.e. how the accuracy varies as the head moves [Villa2008].

Eye trackers necessarily measure the rotation of the eye with respect to the measuring system. If the measuring system is head mounted then eye-in-head angles are measured. If the measuring system is table mounted, as with scleral search coils or table mounted camera ("remote") systems, then gaze angles are measured. In many applications, the head position is fixed using a bite bar, a forehead support or something similar, so that eye position and gaze are the same. In other cases, the head is free to move, and head movement is measured with systems such as magnetic or video based head trackers. For head-mounted trackers, head position and direction are added to eye-in-head direction to determine gaze direction. For table-mounted systems, such as search coils, head direction is subtracted from gaze direction to determine eye-in-head position.

Head mounted devices are a good option when accurate gaze detection is needed because they allow relatively free head movements. Most head mounted trackers use a second video camera, which captures the scene view and allows tracking objects of interest. Raudonis et al algorithm uses monocular video camera that is mounted on the glasses and directed to the user's eye. The user's eye is illuminated with one IR light diode. Head orientation in two directions is measured using accelerometer, which is attached to the side of the glasses. The signals from the accelerometer are used for the head pose compensation. The hardware of the proposed eye tracking system is shown in Figure 5 [RAUDO2012].

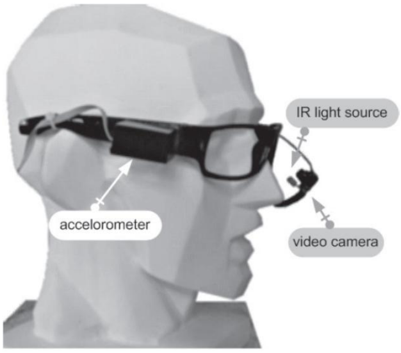


Figure 17 Proposed monocular eye tracking system [RAUDO2012] .

It is obvious, that the designed gaze tracking system must satisfy strong requirements, that come from the users of the target group: the system must be stable and work steadily in different lightning conditions; the user should be able to calibrate and recalibrate the system easily and independently; the system should be portable, flexible and as miniature as possible, i.e., easy to mount and move with electrical wheelchair; it should be possible to use the system for emailing, internet browsing, writing, gaming and etc. Most of the reviewed systems are applicable to certain cases, but still struggle with three main tasks: accurate pupil detection, compensation of natural head motions and "Midas touch" problem [RAUDO2012].

Zhu et all introduce two novel techniques to improve the existing gaze tracking techniques. First, a simple 3-D gaze tracking technique is proposed to estimate the 3-D direction of the gaze. Different from the existing 3-D techniques, the proposed 3-D gaze tracking technique can estimate the optic axis of the eye without the need to know any user-dependent parameters about the eyeball. Hence, the 3-D direction of the gaze can be estimated in a way allowing more easy implementation, improving the robustness and accuracy of the gaze estimation simultaneously. Second, they introduced a novel 2-D mapping-based gaze estimation technique to allow free head movements and minimize the calibration procedure to only one time for a new individual. A dynamic head compensation model is proposed to compensate for the head movements so that whenever the head moves, the gaze mapping function at a new 3-D head position can be updated automatically. Hence, accurate gaze information can always be estimated as the head moves. Therefore, by using their proposed gaze tracking techniques, a more robust, accurate, comfortable and useful eye gaze tracking system can be built [Zhu2007].

Hyrskykari writes that the inaccuracy of eye trackers derives not only from technology issues but also from the physical structure of the eye. She divides at least the following three sources [Hyrsky2006]:

- Measurement in accuracies – the accuracy of the measured point of gaze depends on the eye tracking device and the success of the calibration performed.
- Drift from calibration – inaccuracy also originates from imprecise compensation for head movements and from change in the size or shape of the measured characteristics of the eye.
- nature of the eye - even if no error is caused by the other two factors, the he or she can focus the visual attention without moving the eyes.

Most eye-tracking devices give a gaze position accuracy of 0.5˚–1.0˚, which equals to approximately 0.5cm – 1.0 cm on a monitor when the distance between the monitor and the subject is 60 cm [AOKI2009].

3.1.2.1.6 Applications

Typical application domains for eye trackers are [Drewes2010]:

• **Market research and advertising testing**

The perhaps biggest field in terms of money is the use of eye trackers for market research. When designing posters for an advertisement campaign the marketing research department likes to test the materials. They present the posters to potential clients whose eyes are tracked to get answers to questions like “Did the person look at the product?”, “How much time did the gaze spend on the company’s logo?” and so on. With a portable eye tracker it is also possible to send people to a supermarket to find out which products the people notice and how much influence the form or colour or positioning of the product has on being noticed. See Figure 6 for examples of eye trackers in marketing research [Drewes2010].

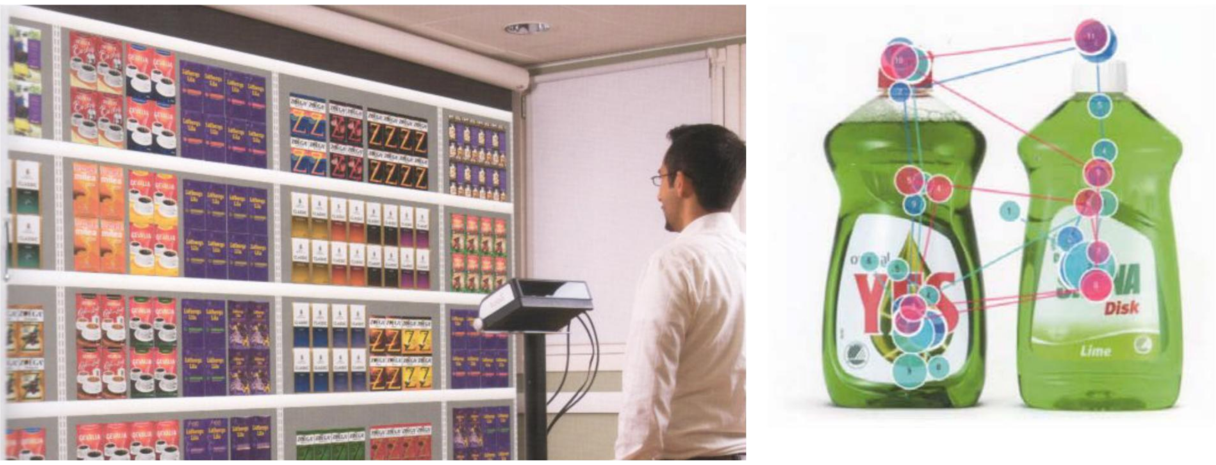


Figure 18. Advertisement for Tobii eye trackers showing a possible application for their product [Tobii2012].

• Usability research

Another field of commercial interest is usability testing. When offering a new device to somebody whose eyes are tracked, it is easy to see where the gaze moves in the expectation to find the control for solving the given task.

With the rise of the World Wide Web as a commercial platform the usability of web pages became an important topic. The user interfaces provided by web pages are very often the only contact of the company and the client and if only few percent of the users are irritated and do not click the “I order now” button the company loses sales. See [Pan2004] for a scientific study with the title “The determinants of web page viewing behaviour: an eye-tracking study”. Some examples are illustrated in Figure 7.

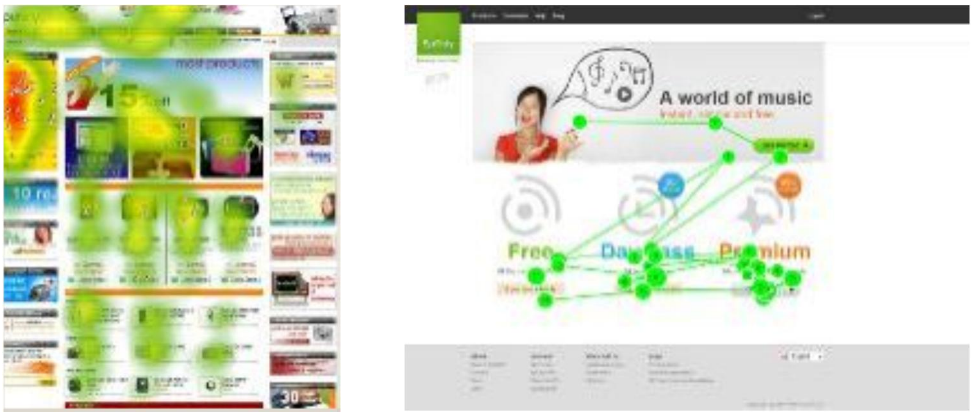


Figure 19. This heat map on the right side shows that most attention is paid to the menu on the left, which is also where most mouse clicks are deployed. The boxes to the right and the left of the page are practically ignored. The gaze plot visualization on the left side shows the search pattern of a user trying to sign up for a subscription and the options considered before making a decision. [Tobii2012].

• Eye control for accessibility

Another field for eye trackers is accessibility. Quadriplegics and people with diseases, causing a loss of control over the muscles, use eye trackers to interact with the world. Such eye tracker systems provide an eye-typing interface with a text-to-speech output. In addition, other types of eye-control, for instance directing the wheel 2 Overview and Related Work 17 chair or switching on the TV, are common in this field. See [Majaranta2002] for a scientific retrospective overview on eye typing. See [Hornof2004 for EyeDraw, a software program that enables children with severe mobility impairments to use an eye tracker to draw pictures with their eyes [Figure 8]

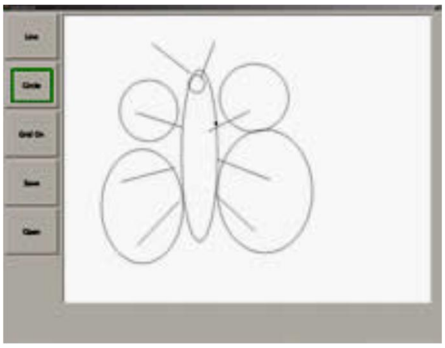
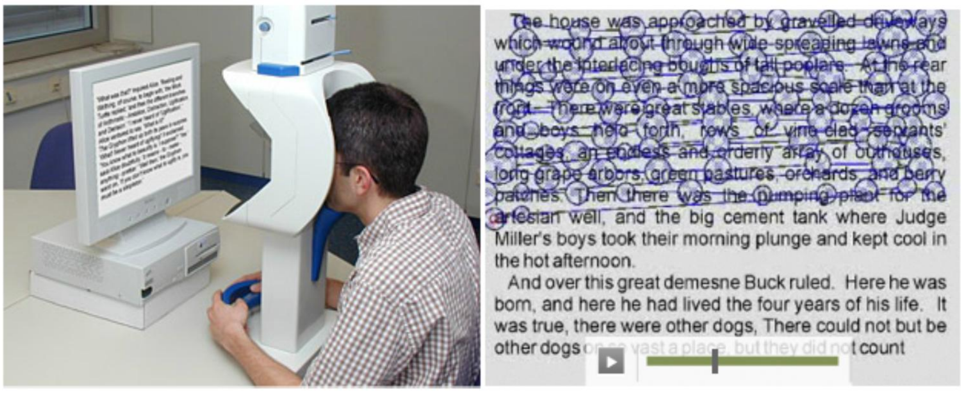


Figure 20The picture is an early version of EyeDraw and a drawing created by one of its developers. The challenge is to provide an intuitive and plausible way for users to intentionally place shapes on the canvas. This requires the program to distinguish between when a user is "looking" and when they are "drawing." [EyeDraw2012].

Psychology and vision research

Eye trackers are a valuable tool for research on vision, perception, cognition, and psychology. One of the first things psychologists did after eye tracking became available was studying the reading process. See the publication list at {Deubel2012] for examples of current psychology research on eye movements. The reading process can be studied and analysed using gaze tracking solution which is depicted in Figure 9.



The house was approached by gravelled driveways which wound about through wide spreading lawns and under the interlacing boughs of tall poplars. At the rear things were on even a more spacious scale than at the front. There were great stables, where a dozen grooms and boys held forth, rows of vine clad servants' cottages, an endless and orderly array of outhouses, long grape arbors, green pastures, orchards, and berry patches. Then there was the pumping plant for the artesian well, and the big cement tank where Judge Miller's boys took their morning plunge and kept cool in the hot afternoon.

And over this great demesne Buck ruled. Here he was born, and here he had lived the four years of his life. It was true, there were other dogs, There could not but be other dogs on so vast a place, but they did not count

Figure 21. Eye and gaze tracking solutions can assist in monitoring on-screen eye movements in reading, following verbal directions in finding objects, and saccades/anti-saccades based on read or verbal instructions for both the qualitative and quantitative observation of the behavioural areas [smivision2012].

• Medical research, diagnostics and rehabilitation

Eye trackers are also a valuable tool for medical research and diagnostics to examine the function of the eyes especially on people with eye injuries or brain damage causing partial loss of vision. In these cases eye tracking is used to monitor the success of physical rehabilitation [Drewes2010]:

• Gaze interaction and car assistant systems

Future applications for eye-tracking technologies are gaze interaction for regular users. A TV set could switch on by itself when somebody is looking at it. The car industry does research in this field too, with the aim of developing assistant systems for cars. For example, an eye tracker in the car could warn the driver when she or he falls asleep while driving the car. This field of application for eye tracking is not yet commercially available and needs further research [Drewes2010].

• Application of gaze detection to immersion measurement in communications

The gaze direction reveals information on the users' intentions and about their focus of attention [Toet, A. (2006); "Gaze Directed Displays as an Enabling Technology for Attention Aware Systems"; Computers in Human Behavior 22(4), 615-647]. Because the feeling of immersion in a communication/interactive application is a gradual process, it seems reasonable to associate it to a change in visual attention. The eye movement consists in fixations and saccades, and in [Land, M.F. (2006); Eye movements and the control of actions in everyday life; Progress in Retinal and Eye Research 25, 296–324] we learn that users capture information during the fixations. Measuring attention will then consist in measuring fixations, what suggests two observations: fixations duration and number of fixations.

In [Jenett, C., Cox, A., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A. (2008); Measuring and defining the experience of immersion in games; International Journal of Human-Computer Studies

66(9), 641–661], the change in gaze movement during an immersive task is investigated. The authors say that the number of fixations per second should decrease for a more immersive task, as the attention will be more focused. On the contrary the number of fixations per second should increase if the user is more distracted…

3.1.2.1.7 Commercial Eye Tracker

The number of available systems is continuously increasing. There is a list of available systems on web pages [COCAIN2012].

Commercial eye-tracking systems that are used for controlling a computer or as communication aids by people with disabilities:

- Alea Technologies Gmbh: Intelligaze IG-30
- DynaVox Technologies: EyeMax System
- Eye Response Technologies: ERICA
- EyeTech Digital Systems: EyeTech TM3, TM4, VT1 and VT2
- H.K. EyeCan: VisionKey (5+, 6V/H, 7)
- Eye-Com
- HumanElektronik GmbH: SeeTech
- LC Technologies: The Eyegaze Communication System, Eyegaze Edge and Eyegaze Edge Tablet
- Metrovision: VISIOBOARD
- Opportunity Foundation of America: EagleEyes
- PRC (Prentke Romich Company): ECOpoint
- TechnoWorks CO.,LTD.: TE-9100 Nursing System for Enhancing Patients' Self-support
- Tobii Technology: Tobii C8, C12, CEye, MyTobii P10, D10
- Utechzone: Spring

Eye trackers for eye movement research, analysis and evaluation. They are more academic systems as follows:

- AmTech GmbH, Compact Intergrated Pupillograph (CIP), Pupillograhic Sleepiness Test (PST), table mounted, monocular, video based systems
- Applied Science Laboratories, ASL, eye tracking and pupillometry systems, both IROG (limbus tracker) and VOG (video) based systems, both head mounted and remote tracking, also mobile tracking!
- Arrington Research, ViewPoint EyeTracker, both remote and head mounted, video based
- Cambridge Research Systems Ltd., MR-Eyetracker, a low-cost, contact-free eyetracker for fMRI & MEG
- Chronos Vision eye tracking devices are used in e.g. neuroscience, ophthalmology, refractive surgery or clinical research. The classic Chronos Eye Tracker was deployed on the International Space Station (ISS) in early 2004 and is in continuous use for the study of eye and head coordination during long-term stays in the weightlessness of spaceflight.
- CLS ProFakt Ltd, offers eye tracking services, analysis software and an integrated virtual shopping with eye-tracking tool for FMCG manufacturers
- easyGaze(R), a low-cost high fidelity eye-tracker for research and training enhancement
- EL-MAR Inc., VISION 2000, portable head mounted video based eye-tracking systems
- Ergoneers Dikablis, Soft- and hardwaresuite D-Lab & Dikablis for planning, performing and analyzing eye-tracking and behavioral experiments; fully automated gaze-data analysis in any environment without any restrictions in head and body movement for a motion range of 500m with Dikablis Wireless eye-tracking system.
- Eye-Com wearable eye tracking and head tracking for clinical and human factors research.
- EyeTech Digital Systems, EyeTech TM3 Eye Tracker Add-on, Research Package, and MegaTracker with free API with full access to raw gaze data and eye metrics
- EyeTracking, Inc., technology developed by Marshall & CERF, San Diego State University

- Fourward Technologies, Inc., Dual-Purkinje-Image (DPI) Eye tracker, mainly for research purposes
- ILAB, eye movement analysis software, works with a number of common eye trackers by ASL, ISCAN, and SMI, reads also CORTEX files
- Interactive Minds, Eye tracking software and tools
- Interactive Systems Labs, Model-based face and gaze tracking (from video image), Carnegie Mellon University
- Iota AB, EyeTrace Systems, head mounted, binocular, video and IR based eye trace systems
- ISCAN, Eye & Target Tracking Instrumentation, head mounted and remote eye tracking systems, single and multible target video tracking systems
- LC Technologies Inc., a remote video based eyegaze development system for human factors research
- Mangold International, MangoldVision for lightweight, portable eye tracking, solutions for both remote and head-mounted eye tracking. Software for data recording and analysis.
- Metrovision, MonEOG: Electro-oculography (EOG) potential measurement based gaze tracking, MonVOG1&2: video-oculography (VOG) based gaze tracking
- Mirametrix, Portable, remote, USB based eye tracking for academic and market research with the S1 eye tracker and easy to use open standard API
- NAC Image Technology, NAC EMR-8 eye path tracking (IROG based)
- Ober Consulting Poland: JAZZ-novo, portable multisensor system with IR based eye-tracker (1 kHz temporal resolution), head rotation and tilt measurement, blood pulse monitoring, voice recording and optional video context recording, designed to study human interaction with environment.
- Ober Consulting Poland: Saccadometer, portable eye movement laboratory for study on saccadic reactions using multiple diagnostic experiments, integrated stimulation and eye movement measurement and recording system, head mounted, IR based (1 kHz temporal resolution).
- Optomotor Laboratories, Express-Eye, a stand-alone eye tracker with saccade analysis, and FixTrain, a small hand held device for daily training of saccadic eye movement control
- Primelec, D. Florin, Angle-Meter NT, a digitally controlled scleral search coil system for the linear detection of 3D angular eye and head movements
- Seeing Machines, faceLAB, a 3D head position and eye-gaze direction tracking system (VOG based)
- SensoMotoric Instruments GmbH, Remote (RED), head mounted (HED) and Hi-Speed eye and gaze tracking for research and applied science, open programming interface and comprehensive stimulus/analysis software.
- Skalar Medical BV, head mounted Chronos and IRIS eye trackers, Scleral Search Coil Systems
- Smart Eye AB, eye tracking analysis based on any standard camera(s), analog or digital
- SR Research Ltd, EyeLink II, video based, head mounted eye tracking system
- Synthetic Environments, Inc., EyeTalk integrates voice recognition and eye-tracking
- TestUsability, EyeCatcher system measures eye scanning and mouse clicking, a helmet fitted with cameras, optics and a microphone
- Thomas RECORDING GmbH, Eye-Tracking-System (ET-49) system, constructed for neuro-scientific purposes and enables a laboratory to correlate the monkey's eye position
- Tobii Technology, Tobii T60 and T120 Eye Trackers - both integrated into a 17" TFT monitor, and Tobii X120 Eye Tracker - a standalone eye tracking unit designed for eye tracking studies relative to any surface.

Probably the most popular eye tracker manufacture is Tobii [Tobii2012]. The Tobii eye trackers allow free head movements in the front of the device and can be found at psychology departments and marketing research companies. The temporal resolution of Tobii X120 is Hz. The spatial resolution for Tobii has not been cited, but the accuracy of gaze position is reported to be 0.5 degrees. It only works when the participant is close to the monitor or device. It can be used in psychology studies, using, web and software usability studies, online marketing studies, human computer interaction research and eye based computer interaction.

FaceLab 5 is flexible, mobile and complete eye and facial tracking solution system. It can detect eyelid, lip, eyebrow and even the size of pupil. It can be used in studies, like flight and driving simulations, and for general purpose tracking like experiment recording, data visualization, analysing and reporting. The accuracy of gaze position is reported to be 0.5-1° rotational error, depending on selected field-of-view [facelab2012]. It is more academic system. FaceLab 5 is shown in Figure 11.

The Mirametrix S2 Eye Tracker is capable of determining where someone is looking on a computer screen in real time. Its algorithms track both eyes, offering a point of gaze measurement that is accurate within a 0.5 to 1 degree range. The eye tracker is flexible under a number of different conditions, like wearing glasses, head tilting, excessive blinking and head shaking. It is also very quick to reacquire a test subjects eyes when they move out of view. The S2 Eye Tracker is priced $5,000, which includes the eye tracker hardware, software and API [mirametrix2012]. Mirametrix S2 Eye Tracker is in Figure 12.

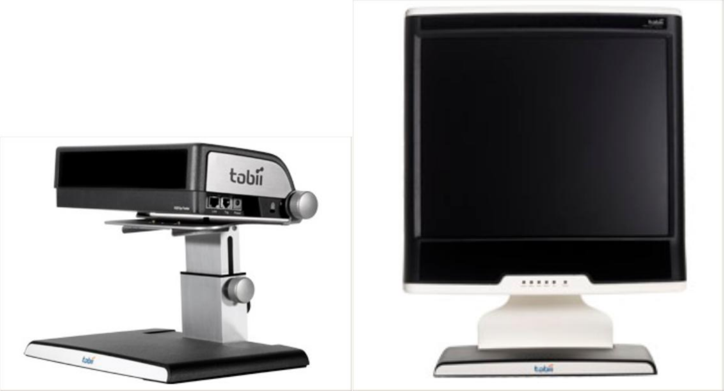


Figure 22. The Tobii X120 (stand-alone) and Tobii T60 (integrated into display) [Tobii2012].

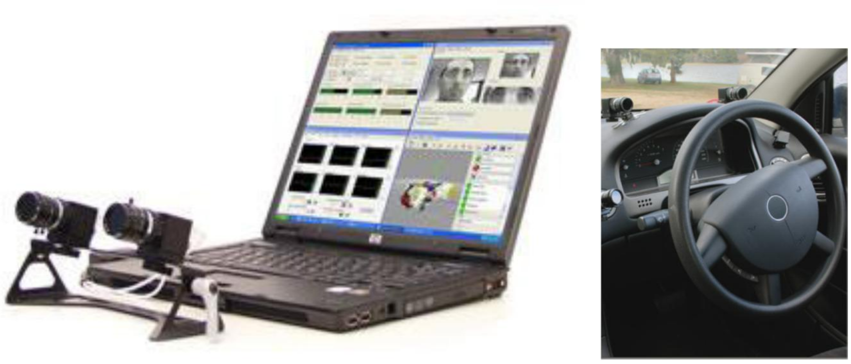


Figure 23. faceLAB is capable of providing head and eye tracking data for research into human behavior in real vehicles, cabins and cockpits, both indoors or outdoors. [facelab2012].

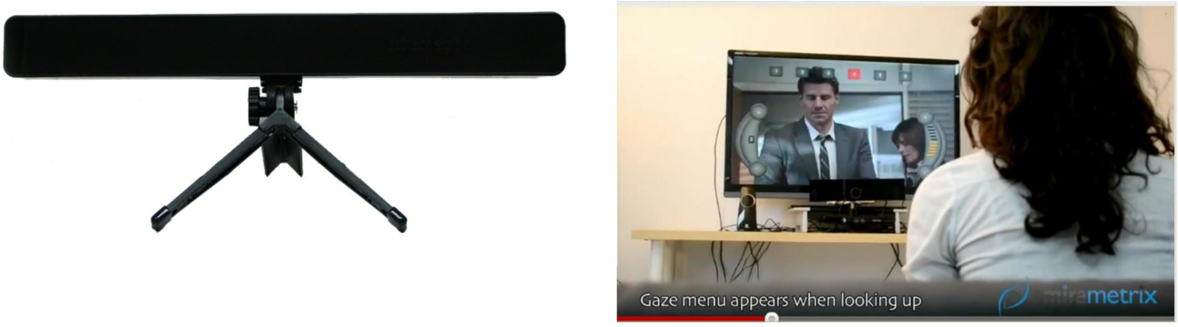


Figure 24. Mirametrix S2 Eye Tracker. Demonstration how to interact with technology in your living room [mirametrix2012].

3.1.2.2 Head orientation

The head pose estimation consists of locating a person's head and estimating its orientation in a space using the 3 degrees of freedom (see Figure 13) which are :

- Tilt (Pitch) : corresponds to a bottom/up head movement, around the x axis.
- Pan (Yaw) : corresponds to a right/left head movement, around the y axis.
- Roll (Slant) : corresponds to a profile head movement, around the z axis.

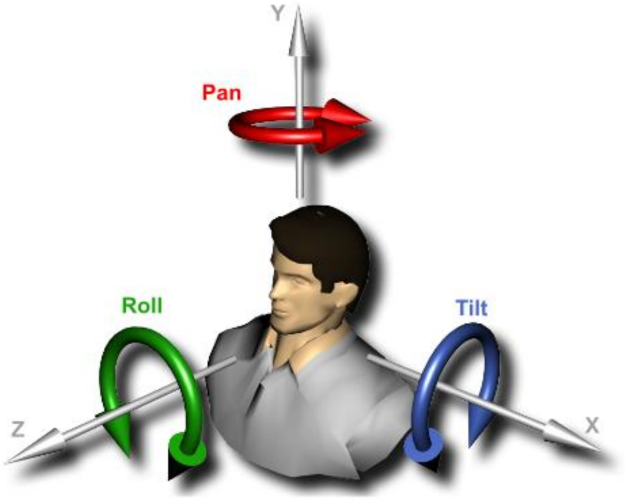


Figure 25 : Head degrees of freedom model for head pose estimation

Head pose estimation is intrinsically linked with visual gaze estimation. Head pose provides a coarse indication of gaze that can be estimated in situations when the eyes of a person are not visible (like low-resolution imagery, or in the presence of eye-occluding objects like sunglasses).
Physiological investigations have demonstrated that a person's prediction of gaze comes from a combination of both head pose and eye direction [La04]. The authors established that an observer's interpretation of gaze is skewed in the direction of the target's head.
A graphic example of this effect was demonstrated in the 19th-century drawing shown in Figure 14 [Wo24]. In this sketch, two views of a head are presented at different orientations, but the eyes are drawn in an identical configuration in both.

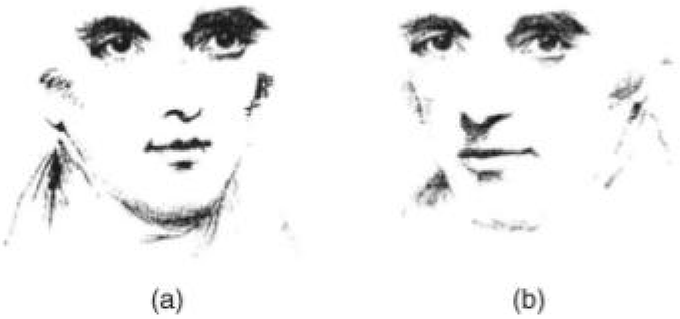


Figure 26 : Wollaston illusion: Although the eyes are the same in both images, the perceived gaze direction is dictated by the orientation of the head [Wo].

Creating robust and accurate head pose estimation is a classic problem in computer vision. It has been widely used in many applications such as video conferencing, driver monitoring or human computer interaction. Moreover, for many pattern recognition applications, it is necessary to estimate coarse head pose. Like performing face recognition or facial expression analysis under varying poses.

Head-pose estimation approaches can provide a discrete or continuous pose estimate. When the system builds a model that will provide a discrete pose, there is no way to derive a reliable continuous estimate from the result. This allows only coarse head pose estimates. Therefore, the number of fixed poses must be large enough to sufficiently sample the continuous pose space.

Many approaches are proposed to deal with head pose estimation. They can be organized in different ways. In [Mu09] authors consider the fundamental approach that underlies each techniques implementation. These techniques can be organized in seven categories with regard to their functional approach:

- Geometric methods: use the location of facial features such as eyes, mouth and nose and geometrically determine pose from there relative configuration.
- Flexible models: build a nonrigid model which is fit to the image such that it conforms to the facial structure of each individual (AAM).
- Template matching methods: consist on comparing images or filtered images to a set of training examples and find the most similar.
- Classification methods: formulate the head pose estimation as a pattern classification problem.
- Regression methods: use regression tools to develop a functional mapping from the image or feature data to a head pose measurement.
- Manifold embedding methods: produce a low dimensional representation of the original facial features and then learn a mapping from the low dimentional manifold to the angles.
- Tracking methods: use temporal information brought by the sequence of frames in a video.

Geometric methods – Wang and al. [Wa07] propose a method using the inner and outer corners of each eye and the corners of the mouth. They assume that if the three lines between the outer eye corners, the inner
eye corners, and the mouth corners are deviated from parallel, then there is a perspective distortion. The vanishing point is used to estimate the 3D orientation of the parallel lines.
In [Pan05] authors propose an algorithm to obtain 3-D head pose information based on the relations of projections and the positions changing of seven facial points.

Flexible models – A flexible model that has evolved for head pose estimation is the Active Appearance Model (AAM) [Co01], which learns the primary modes of variation in facial shape and texture from a 2D perspective. Further works with AAMs have introduced modifications that expand theirutility to driver head pose estimation [Ba04]. Accurately matching a deformable face model to image sequences with large amounts of head movement is still challenging task [An10].

Template matching methods – Beymer [Ba94] represent faces with templates from multiple model views that cover different poses and use normalized cross-correlation at multiple image resolutions to find the best match in the data base of people. Authors in [Ja99] investigate for filters that highlight oriented features such as the vertical lines of the nose and horizontal orientation of the mouth. They use Gabor filters to discriminate pose.

Classification methods – Several works have used a range of classifiers such as SVM [Hu98] and recently [Da11]. Isarun and al. [Is11] uses random trees beside SVM. In [Li01] Kernel Principal Component Analysis (KPCA) is used to learn a non-linear subspace for each range of view, then a test face is classified into one of the facial views using Kernel Support Vector Classifier (KSVC). Also, classification is achieved in [Be08] using an ensemble of randomized ferns and in [Zh07] Naive Bayes classifier is applied to estimate head pose.

Regression methods – Several regressors are possible such as Convex Regularized Sparse Regression (CRSR) [Hao11] and Gaussian Progress Regression (GPR) [Ra08]. Murad and al. [Mu12] proposed a method based on Partial Least Squares (PLS) Regression to estimate head pose. Support Vector Regressors (SVRs) are used to train Localised Gradient Orientation (LGO) histogram computed on detected facial region to estimate driver's head pose in [Mu07].

Manifold embedding methods – In [Be10] author study the incorporation of continuous pose angle information into one or more stage of the manifold learning process such as Neighborhood Preserving Embedding (NPE) and Locality Preserving Projection (LPP). Dong [Do11] proposed Supervised Local

Subspace Learning (SL2) to learn a local linear model where the mapping from the input data to the embedded space was learned using a Generalized Regression Neural Network regression (GRNN). In [Xi10] author proposed the K-manifold clustering method, integrating manifold embedding and clustering.

Tracking methods – [Ba08] evaluate two tracking methods : Annealed Particle Filtering et Partitioned Sampling in the context of estimation 3D human pose. In [Ch11], a pedestrian tracker is applied to the heads video to infer head pose labels from walking direction and automatically aggregate ground truth head pose labels. Ba and al. [Ba11] aims recognition of people's visual focus of attention by using a tracking system based on particle filtering techniques.

3.1.3 Gesture recognition and interpretation

First attempts to integrate hand gestures in Human-Computer Interaction (HCI) resulted in sensors and some special mechanical devices that must be worn by user to measure specific data of the hand [Qua90]. This group of devices may be composed by acoustic or inertial trackers, magnetic sensors, or special gloves such as the CyberGlove[A].

Figure 27 - The CyberGlove from the Immersion Corporation (from [A]).

Although those invasive interaction techniques provide very accurate results and are useful especially for cooperative people, such cumbersome and expensive tools tend to impose a non common burden in our everyday life. Besides, the use of gloves limited the number of natural postures. Moreover, computer vision techniques are less intrusive in our context. They provide a more natural interaction since no device occurs in the interaction process. Furthermore, vision-based interface carries the advantage of considering the hand as the only device used during the interaction between human and computer. In our work, we are focusing on computer vision approaches.

A complete hand gesture recognition system is composed of three layers: hand detection, hand tracking and gesture recognition (see figure 16).

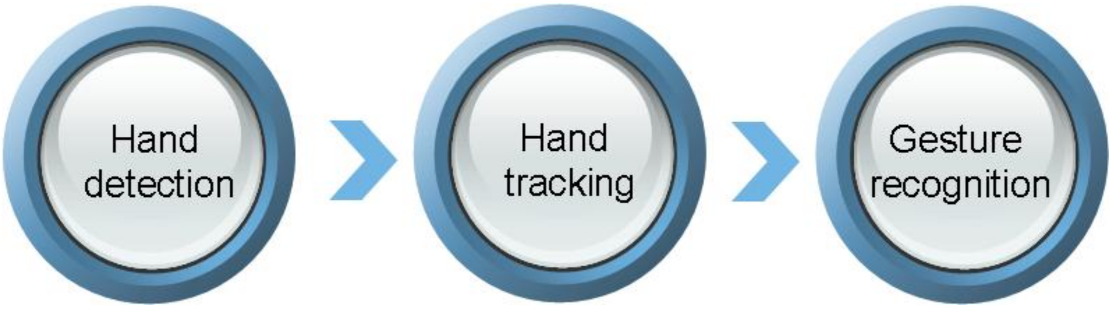


Figure 28- The steps of hand gesture recognition system

The detection layer consists on the extraction of hand region and position. The tracking layer performs temporal data association between successive frames in the video sequence. Finally, the last layer uses previous spatio-temporal information to categorize gestures into the appropriate classes. Based on the features used to represent the hand, one dichotomy to differentiate vision-based hand gesture recognition algorithms is appearance-based approach versus 3D hand model-based approach [Pav97].

3.1.3.1 Hand Gesture/Posture recognition from CCD camera

3.1.3.1.1 Hand detection

Hand detection is a key step in hand gesture recognition system. Depending on the approach used for the hand detection task, the result can be a simple box surrounding the hand or a blob representing the segmented hand. Unlike the task of face detection, limited researches have been done to detect the hand in the image due to its high deformable nature. Most of the approaches have been concentrated on background subtraction, thresholding, or skin detection. This section presents a small review of hand detection approaches.

3.1.3.1.1.1 Background subtraction

The background subtraction technique is widely used in computer vision systems. It consists in subtracting a reference image from the scene to obtain the foreground. This image can represent a special known background in the case of hand detection in static images, or the first frame in the case of video sequence.

In order to track the finger of the hand, Conseil et al. [Con05] detect the hand in the image by subtracting the background that constitutes a reference image taken in the initialization of the system. To eliminate the noise in the binary image obtained after the subtraction, they apply a median filter. Then, they extract the connected component corresponding to the hand. However, their approach is very sensitive to the shadows that can exist with the hand.

Shaker et al. [Sha07] used a background subtraction scheme to separate potential hand pixels from other pixels. First, they converted captured images into grayscale images. Then, they perform a threshold on image pixels resulted from the subtraction between the static background and the current image.

While the subtraction between two images is low computationally expensive, it is quite sensible to luminosity change. Mostly, the background subtraction is based on the assumption of a static background and camera. This assumption can generate the detection of other object if the camera moves or if the background changes by putting new statics objects. This technique is insufficient if the user is completely present in the captured image since it assumes that the hand is the only object existing in the image. In this case, it may be interesting to combine it with another approach to detect the hand.

3.1.3.1.1.2 Threshold-based approaches

These approaches are based on some constraint assumptions to facilitate the hand detection task. The idea is to make the hand's or marker's color or intensity quite different from the rest of the scene. The hand is then segmented in the image using a simple thresholding.

In some work on gesture recognition, the user has to wear gloves with a special color in order to segment the hand easily. In the context of human computer interaction, Kang-Hyun et al. [Kun98] detect the hand regions in the color image by a simple thresholding in hue information. In their work, the user has to wear a green glove and a red one on each hand. Similarly in [Sta95], the user must wear distinct colored gloves on each hand; a yellow glove for the right hand and an orange glove for the left one.

Another attempt to detect the hand based on thresholding is to use a special marker worn by the hand user. In [Jep98], when the user needs to perform an order with a dynamic gesture, he must pick up a “phicon” (physical icon) that has a distinctive color which facilitate the detection and tracking process. In the same way in [Loc02], users have to wear a wristband in order to determinate the localization, the orientation and the scale of the hand region.

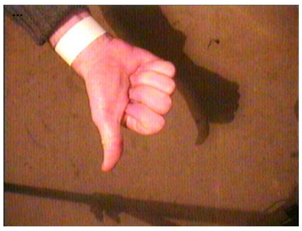


Figure 29 - The wristband used in [Loc02]

Since the colored gloves and markers are applied only on colored image, Hamada et al. [Ham02] proposed to detect the hand silhouette in grey-level image assuming that the hand has a different intensity in the particular environment used.

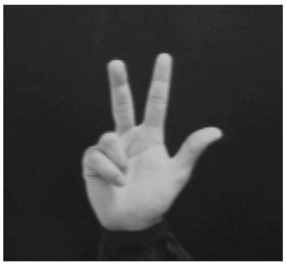


Figure 30 - The detection of the hand in front of a black background

In their work, the hand is the only object presented in the image in front of a black background. The user has to wear dark clothes so that the hand region can be extracted easily since it is brighter than the background and the clothes.

These simple approaches based on special environment assumption are used in order to obtain a real-time processing. In spite of their efficiency and their speed, these techniques may not be applied in a real situation since they impose constraints that may not be accepted by many users in the context of a natural human computer interaction.

3.1.3.1.1.3 Skin detection

Since the human skin is characterized by a special color that is clearly recognized by human, the use of skin color as a cue to detect hands and faces has gained strong popularity. The final objective of skin color detection is defining decision rules that allow the discrimination between skin-color pixels and non-skin-color pixels. In this section, we quote the different ways to model the skin color distribution.

3.1.3.1.1.3.1 Colorspace skin clustering

These techniques define explicitly the skin color as skin clusters in the appropriate color space. Then, the defined rules giving a rough skin sample regions depend on the chosen colorspace.

In [Gon10], the samples pixels initialization was performed by defining explicitly skin regions. The boundaries skin clusters were defined through a number of rules in RGB colorspace [Kov03]. To differentiate between the noise information and skin color regions resulting from this operation, they keep only the sample pixels belonging to the greatest area. To generate the specific model chosen for the skin color, they used the YCbCr colorspace to model it as a normal distribution while Y component was rejected from the process to avoid shadow problems.

The main difficulty with the use of 3D chromaticity space, like RGB colorspace, for skin detection is that it doesn't separate the luminance components from the chrominance components which can affect the performance of the used rules in the case of bad illumination conditions. To solve this problem, a specific algorithm performing a pre-treatment in the image can be used to eliminate the influence of non-standard illumination [Fun98]. A comparative study in [Kov03] proved that the use of the algorithm Grey World [Buc80] with RGB colorspace has shown better results.

The main advantage of colorspace skin clustering methods is the simplicity which allows constructing a very rapid classifier based on defined skin detection rules. Besides, there is no need to perform a learning stage.

3.1.3.1.1.3.2 Non-parametric skin distribution modeling

The principle of non-parametric skin modeling is that the estimation of skin color distribution doesn't require an explicit model or rules of the skin color. It is based on training data.

Mostly, a histogram is used to detect skin pixels. In [Kol04], the authors used a normalized RGB histogram for hand detection and segmentation. Nerveless, their algorithm relies on two assumptions: first, the background reference area does not contain another exposed skin body parts of the same person whose hand is to be tracked. Second, the reference background does not contain wooden objects; otherwise, they will be classified as a skin color. The histogram will deviate if these two assumptions are not considered, which lead to a fast degradation of segmentation results.

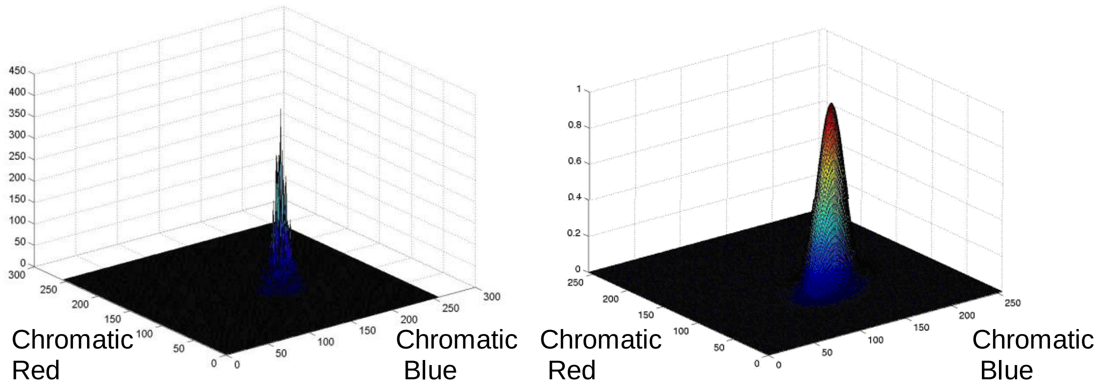
A color-histogram is constructed in [Sub95] to estimate the skin color distribution based on the normalized RGB colorspace. The system is trained on the sample skin patches (a small square region in the image) selected from an image. The result of hand segmentation with the constructed histogram is composed of the selected rectangular patches.

In spite of the speed of non-parametric skin distribution modeling techniques in the training and the usage, their performance is strongly related to the used database. To achieve an important rate of detection, they must use a rich database composed of many kind of color skin. Furthermore, histogram-based techniques require a large storage space.

3.1.3.1.1.3.3 Parametric skin distribution modeling

These techniques are motivated by the limitations of non-parametric skin distribution modeling methods. They consist on modeling a compact skin representation that is required for certain applications. They model skin color densities in a parametric form such as the Gaussian model or the mixture Gaussian model.

Lai et al. [Lai06] detect skin pixels in images by fitting a Gaussian model on the image in the colorspace YCbCr. 45.000 pixels were extracted from 90 face images from different ethnicities to construct the Gaussian model fitted to this data only on the chrominance components Cb and Cr (see figure 19). Although this method was proved only on faces image, it can be applied also on hands since they have almost the same color.



Different people skin color disttibution

Fitting skin color into a Gaussian distribution

a)

b)

Figure 31 - (a) Skin colour distributions of the data (45000 pixels) in Cb-Cr space, (b) A Gaussian model fitted on the data shown in (a).

Assuming that the simple Gaussian model is insufficient to obtain an accurate segmentation of the hand, Gejgus et al. [Gej04] construct a Gaussian mixture model to detect efficiently skin pixels. In their approach, the user must wear black clothes and stay in front of a black background in order to facilitate the hand segmentation task.

A Gaussian mixture model with the restricted EM algorithm was used in [Zhu00] to segment hands. Their approach is based on Bayes decision theory to classify pixels into skin and background pixels.

Similarly to non-parametric skin distribution modeling methods, these techniques require a large database containing skin color images from different ethnicities and taken under different luminance conditions, otherwise they can fail in classifying skin pixels that don't verify the same conditions as in the training database.

3.1.3.1.1.4 Other hand detection approaches

In addition to the classical approaches mentioned previously, there are other techniques to detect the hand in the image.

Motivated by the success of the Viola-Jones object detection algorithm [Vio01] for face detection, kolsch et al. [Tur04] proposed a robustness analysis of this algorithm for hand appearance detection. For this purpose, only six different postures were used to recognize the hand. They concluded that only about 15° of rotations can be detected knowing that the training database must contain some rotated images.

Ong et al. [Ong04] proposed a new approach capable of both detecting and classifying the hand postures. They perform the learning step via Boosting algorithm to improve the accuracy of their approach. The 634 weak classifiers used in the Boosting algorithm consist on a simple detector based on image block differences calculated using an integral image [Ong04]. 5013 hand images of different people were selected for training while 2509 hand images were kept for testing. The detection error in the test was 0.2%. However, their approach result only in the localization of the hand and it can not perform the segmentation task. Besides, it requires the collection of a large database which slowed the training processing.

Considering that the hand is characterized by special geometric properties, Caglar et al. [Cag06] proposed a hand detection technique that is efficient even with the skin color variation, the change of illumination conditions, and shadow. Their approach detects only open hands based on the different properties of the fingers when the hand is open. They identified first the parallel lines presented in the image resulted from the Burns algorithm [Cag06].

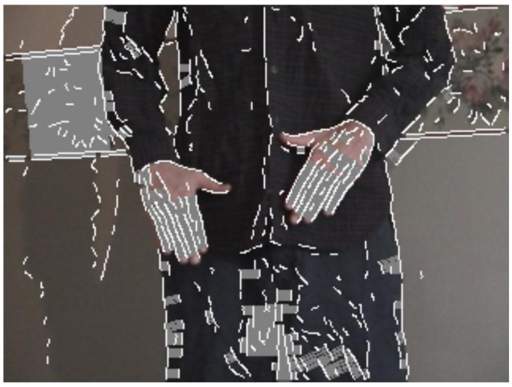


Figure 32 - Finding lines and areas that satisfy the criteria to be fingers

To cope with different open hand orientations, they used 8 models for fingertips. Although their approach is independent from the skin's color, it can show false detection when the hand is situated in a very complex background since the algorithm is based on the edge and lines extracted from the image.

3.1.3.1.2 Tracking

The aim of the tracking over time is to track the position of an object positions in consecutive frames of the video. In the context of gesture recognition, the tracking task can generate the trajectory of the hand over time by locating its position. The difficulties of this analysis are related to the complexity of the scene and the tracked objects or to the partial or total occultation of the object. In general, this one can be represented using different ways. The most commonly representation employed for tracking are points, which can be the centroid of detected objects or a set of points around the object, primitive geometric shapes such as rectangle and ellipse, silhouette or contour and finally skeletal models. The selected model to represent object shape limits the type of motion or deformation it can undergo, then the choice of tracking algorithms can be done according to this model. In this section, we are going to discuss the principal techniques of tracking used in the literature.

3.1.3.1.2.1 Active contour

The active contours or "snakes" were introduced first by Kass et al [Kas88]. They are defined as a set of points representing an initial curve. The curve's topology can constantly change until reaching a final position. This curve evolves iteratively depending on two types of forces: internal forces and external forces corresponding respectively to an internal energy and an external energy. The internal energy measures the regularity of the curve while the external energy is derived from the given image.

Kasprzak et al. [Kas06] proposed an approach for hand sign interpretation based on active contour for tracking. First, they apply a color-based skin pixel approach to detect the hand in the image. For the initialisation of the active contour, the binary image containing skin-color pixels is used as an input data. This image is divided into 3x3 or 4x4 blocks in order to find control points of the active contour. Those points are defined by the left upper corner of blocks which contains a significant amount of skin pixels in the input data. From this initialisation, two active contours are computed corresponding to the palm area and the total area of the hand including fingers. In their work, they used a new form of the snake called "Gradient Vector Flow" (GVF) snake proposed by Xu et al. [Xu98]. The external force of this snake takes the form of gradient vector flow (GVF) fields. Although this force permits the convergence into boundary concavities, it requires more computation time and it is very sensitive to noise.

The disadvantage of the active contour is that the computation time required for adjusting the curve between two frames can not allow a real time tracking. In addition, the set of their initial parameters must be done thoroughly and sometimes it requires an expert‘s intervention. Even the initialization of the active contour can be resolved in an automatic way [Li08] [Ge02] [Liu05] instead of some geometric curves (circle, rectangle, ellipse), the update of this contour in the video is mostly based on the contour of the previous frame. While moving to a new position, the active contour can fall into local minima if there is large difference between the position or the form of the object in successive images.

3.1.3.1.2.2 Condensation algorithm

The Condensation Algorithm (Conditional Density Propagation) [Isa98] is a Bayesian filtering method for motion tracking. It uses a random sampling in order to model an arbitrary probability density function. The algorithm permits the approximation of the curve described by the data by using many weighted samples composed of states and their corresponding weights. Every weight is proportional to the probability that its state is predicted by the input data. In the particular case of tracking, a state corresponds generally to the location or velocity of the tracked object.

The Condensation algorithm was used in [Mam01] to track simultaneously both hands represented by two rectangular windows. They used 100 samples points distributed on the rectangular windows surrounding the hand. They showed that their method is robust against errors in observations due to clutter and occlusions.

The Condensation algorithm is a particle filter that is widely used in computer vision applications for its highly robust tracking of object motion. The precision of the samples to model the observed data is proportional to the number of samples.

3.1.3.1.2.3 Mean Shift

The Mean Shift tracking algorithm is an iterative procedure that localizes the object in successive frames by comparing the visual features of the tracked object and those of its neighborhood. The color of the hand is the most used features for the tracking task with Mean Shift algorithm. In this case, the histogram of the original object and histogram of candidate regions in the image are compared. The selected region must verify the maximum of correlation between both histograms.

In [Bra98], a new version of the mean shift algorithm called Continuously Adaptive Mean Shift (CAMSHIFT) was developed and applied to face tracking. The tracking task was performed using the color histogram of the skin. At each iteration, a search window is resized and moved to a new position until convergence. The convergence is reached when the window’s center coincides with the center of mass. The proposed algorithm can also be used to track any object of interest in the video sequence.

Based only on the color histogram, the performance of the mean shift algorithm can be affected by a cluttered background. In [Sha04], a combination of mean shift and particle filtering showed more efficiency.

3.1.3.1.2.4 The Kalman filter

The kalman filter is a statistical approach that can be defined as a set of mathematical equations. These equations implement a predictor-corrector type estimator. It has been used extensively for tracking in interactive computer graphics since it allows describing the behaviour of a dynamic system by the evolution of a set of variables called state variables. The kalman filter aims to correct the trajectory of the predicted model by combining the real observation and the information provided by the model in order to minimize the error between the true state and the filtered state.

Coogan and al. [Coo06] used kalman filter model to track the four corner of each bounding box around the face and two hands. At each frame, they check if there is any overlap between the bounding boxes and they compute the number of detected skin objects: if the number of detected skin objects decrease without overlap detection, they conclude that one or more skin objects were missed in the frame.

In [Ram03], the kalman filter is adopted for hand tracking in order to recognize dynamic hand gestures. The states tracked over time are some control points that represent the B-Spline coordinates. The tracker used a discrete time motion model [Ram03].

Since the standard kalman filter is limited to linear movement, Stenger et al. [Ste01] used a new method for hand tracking to estimate the pose of 3D hand model constructed from truncated quadrics in front of a dark background. Their approach is based on a nonlinear version of the Kalman filter called Unscented Kalman Filter (UKF) proposed by [Jul95] which is an alternative to the Extended Kalman Filtering (EKF) [Lju79].

3.1.3.1.3 Hand gestures recognition

The most important differences in the vision based hand gesture recognition approaches arise depending on whether a 3D hand model or an image appearance of the hand is used. The first solution tends to create a three-dimensional model of the user hand which will be used for the recognition task. The second one tries to calculate hand features directly from the 2D image. The strengths and weaknesses of both techniques are discussed in this section.

3.1.3.1.3.1 3D hand model-based approaches

The 3D hand model-based approaches are high-level approaches since they require the extraction of high-level features from the images. They are based on an articulated 3D hand model based on some kinematics parameters to describe the shape and the movement of the hand. These approaches assume that human gestures are some actions in our 3d space that require involving the characterization of their spatial properties to describe their gestures.

In their work, Ueda and al. [Ued03] estimate joint angles of the hand using three dimensional matching between hand model and voxel model. The latter is reconstructed by collecting the multi-viewpoint silhouette images obtained by several cameras.

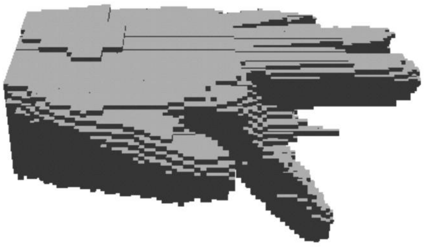


Figure 33 - Voxel model used in [Ued03]

Ying et al. [Yin01] present a model-based approach for capturing natural articulated hand motion. In their cardboard hand model based on 3D geometric structures, the hand articulation is represented by its joint angles.

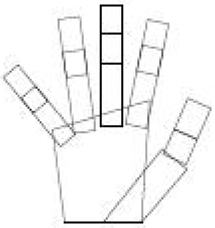


Figure 34 - Cardboard hand model employed in [Yin01]

To model the constraints of hand configuration, they suggest a learning approach using a large set of hand motion data collected using CyberGlove. In order to represent hand articulations in a lower dimensional space, the joint angle space is projected to the configuration space by applying the Principal Components Analysis (PCA) technique. Based on this representation of hand articulation, they employed a sequential Monte Carlo tracking algorithm based on importance sampling [Yin01]. Although this algorithm produces good results, it is view-dependent and it is unable to handle the hand global motions such as rotation.

In [Hea96], the authors use a 3D deformable Point Distribution Model (PDM) to represent the hand as a surface mech. This model is built via the statistical technique PCA from a collection of 3D training images. The landmark coordinate data, which represent the required training information, is extracted semi-automatically from the training database. The tracking is achieved by finding the best matching between the image and the model proposed. Their approach allowed a real-time processing (10 frames/second). Nerveless, it is not scale invariant and it is not able to handle the occlusion problem.

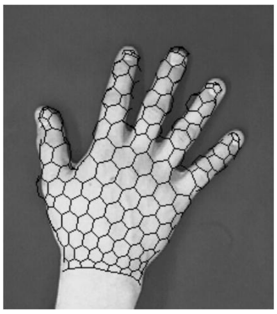


Figure 35 - The 3D hand PDM used in [Hea96]

Taking into account many parameters related to the hand, 3D hand-model based approaches offer a rich description that allows classifying a wide range of hand gestures. However, as the human hand is an articulated deformable object with many DOF [Che08], a very rich database is required to cover most of the characteristic shapes of the hand. For example in the work of Ying et al. [Yin01], a set of more than 30 000 joint angle measurements were collected to form the training database in order to perform various natural finger motions.

Most of these approaches assume that only few updates are necessary between frames as the 3D model configuration at the previous frame is known. Therefore, model-based approaches are strongly depending on the model initialization which brings many researches to adjust the hand model manually or semi-automatically [Hea96] in the first frame. Some authors tend to use gloves [Lee95] in order to facilitate the feature extraction process. However, using gloves for gesture recognition is unpractical since the gloves limit natural moving abilities of the user's hand.

3.1.3.1.3.2 Appearance-based approaches

These approaches are also called view-based approaches since they try to infer hand gestures directly from the visual image observed. Generally, hand gestures recognition by these approaches is performed by a comparison between the parameters defined by the 2D image features extracted to model the appearance of the hand and the 2D image features extracted from the image or video input.

In hand gestures analysis, we have to distinguish between two concepts that are often confused: hand posture and hand gesture (static versus dynamic). The first is defined as a static hand gesture (or configuration) which does not change during a period of time. The second is a dynamic hand movement which can be defined as a temporal trajectory of some estimated parameter over time [Jep98] or as a sequence of hand postures composed of successive hand or finger motion over a period of time [Che08].

3.1.3.1.3.2.1 Static hand gestures recognition

Static hand gestures or postures recognition is a topic of great interest on its own since it's a basic step for numerous applications like sign language communication. Moreover, it represents the basis of some hand gestures recognition techniques which consider the dynamic gesture as a connected sequence of hand postures [Che08]. Since the hand posture can be seen as a particular static configuration, hand posture recognition can refer to a pattern recognition problem. We can split hand postures recognition between 3 categories: statistical approaches, neural networks approaches and template matching approaches.

Statistical approaches:

Statistical approaches represent the hand posture by a feature vector which is viewed as a particular point in the feature space. The features selection step must be done thoroughly so that the appropriate features allow hand postures belonging to discriminated regions in the feature space. These regions, corresponding to the different postures, have to be correctly identified by the classifier.

In [Coo06], authors proposed a static gesture recognition approach based on PCA. The data extracted during the sequence is represented by features vectors. Two elements are presented in those vectors; the first one defines the group in which the static hand shape is classified, the second denotes the position of the hand in the image. They define in their work 40 static hand shape groups and 9 possible positions which represent 9 regions (or ranges) in the images. The position of the hand is classified by determining the range which contains the centre position of its bounding box. The advantage of hand position determination is that it is invariant to the distance between the user and the camera. Nevertheless, the hand shape classifier requires a training phase for each possible gesture using different samples of hands.

Den et al. [Den02] proposed a new approach based on a Principle Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) scheme to recognize 100 hand postures. They have shown an improvement in the recognition rate from 57% to 63.5% by using the Expectation-Maximization (EM) algorithm for the crude classification.

The Support Vector Machines (SVM) is used in [Ras09] to classify and recognize the different hand postures defined in the American Sign Language (ASL). The classified features are composed of two types of features: statistical features represented by seven Hu-moments and two geometric features. The hand shape group recognized contains thirteen ASL alphabets and seven numbers distributed in 3000 training samples and 2000 test samples that allowed achieving a high rate of recognition.

Another approach called the AdaBoost algorithm is widely used in the posture recognition task. The aim of this learning boosting algorithm machine is to combine many weak classifiers into a stronger one. Usually, the accuracies of these simple classifiers are only slightly better than 50%.

Just et al. [Jus06] proposed a boosting approach based on MCT (Modified Census Transform) features to classify 10 different postures. In order to augment the robustness of their classification technique, they trained one classifier for each posture and they add another class of images containing samples images of non-posture for each training task. The rate of recognition has shown a significant decrease with a complex background.

The statistical approaches are widely used for the appearance-based hand posture recognition task, for their ability to handle an unstable data and to discriminate efficiently between dependent features that might be not considered in visual analysis.

Neural networks approaches:

In addition to the discussed statistical approaches, neural networks are a special tool widely used in computer vision. Neural networks are implicitly equivalent or similar to classical statistical pattern recognition methods [Jai00]. Neural networks represent a parallel computing system composed of a large number of simple elements with many interconnections. These elements are inspired from biological nervous systems. The neural networks are characterized by the ability to learn complex nonlinear input-output relationships using sequential training procedures and to adapt themselves to the data.

Gutta et al. [Gut96] used neural networks for face and hand gesture recognition. A hybrid approach is built; it consists on a combination of connectionist networks and inductive decision trees (C4.5). Their experimental results showed the improvement of the classification performance due to their hybrid approach.

Handouyahia et al. [Han99] present a recognition approach for International Sign Language (ISL) using a three-layer feedforward neural network to train and recognize the alphabets. The input of the network is the feature vector calculated on image and composed of 146 elements including size functions and principal axis orientation while the output of the network is a vector with 25 elements. The limitation of their work is that the features used are not rotation invariant.

In spite of their accuracy and power, neural computing has some limitations. In fact, neural networks implement pattern recognition as black boxes by hiding the complexity of the system from the user. This limitation reflects the inability to explain the built model and the difficulty to extract rules used in the process. However, it is sometimes important to people who have to explain used rules to others. Besides, they need an initial learning phase that could be time consuming.

Template matching approaches:

Template matching approaches are the most intuitive, simplest and earliest approaches used in pattern recognition. They are considered as a generic operation used to measure the similarity between two entities of the same type. They consist in matching the pattern to be recognized against a stored template. In the case of images, the matching is performed by the pixel-by-pixel comparison of the candidate image and the prototype image. In that particular case, it is clear that template matching is not invariant to scale and rotation.

Elastic graph is a particular example for template matching. This matching technique was used by Triesch and Malsburg [Tri96] in order to classify hand postures against complex backgrounds. In their approach applied in grey level images, the hand posture is represented by a labelled graph connected by some nodes called jets (see figure 24).

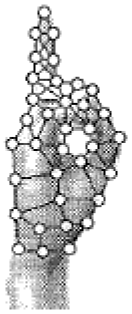


Figure 36 - Hand posture represented by a labelled graph

These nodes are extracted by Gabor-like filters. The use of elastic graph does not require hand segmentation, but it is view-dependent since only one graph is used per hand posture. Besides, it requires the extraction of a large number of nodes to perform the elastic graph matching.

3.1.3.1.3.2.2 Dynamic hand gestures recognition

Several works have been based on the recognition of static hand gestures. In recent years, there has been an increasing interest in including the dynamic characteristics of gestures considering the assumption that human gestures are a dynamic process. A psychological study on the process of temporal discrimination of gestures is showed in [Pav97].

There are many techniques developed around dynamic hand gestures recognition in recent years. Depending on the aspect used for the recognition, we can split these techniques between two categories: Recognition by modeling the dynamic or the semantics and Recognition by modeling the state.

Recognition by modeling the dynamic or the semantics:

Modeling the dynamic of hand motion is important for the tracking task and also for dynamic hand gesture recognition. It permits to reduce the complexity of gesture representing it by the trajectories parameters estimated over time.

In [Yan99], hand gestures are recognized by using motion trajectories in the context of American Sign Language (ASL). They employed a Time-Delayed Neural Network used to classify 40 complex signs. To match the motion regions across frames, they try to find a set of graph transformation operations. The matching operation between a pair of regions takes into account their similarity (in terms of area, expected position, average intensity...). The motion trajectories of the palm region are then extracted by performing an affine transformation of successive pairs. Their general method that extracts motions trajectories is not robust to partial or total occlusion that can happen on hand regions.

Black and Jepson [Jep98] proposed an extension to the Condensation algorithm to recognize temporal trajectories. The extended algorithm allowed them to recognize more complex gestures compared with the standard Condensation algorithm used first for tracking task. The recognition of dynamic gestures was performed by the probabilistic matching of models curves with input curves, while curves represented the trajectory of the hand gesture. In their work, the user must perform the gesture by the mean of a physical distinctly colored object (a “phicon”) to facilitate the location and the tracking of the hand.

In addition to the dynamic modelling techniques which analyze the motion trajectories performed by the hand, another attempt to gesture recognition is to model the semantic of gestures by defining some rules.

Cutler et al. [Cut98] used optical flow for gesture recognition. Optical flow was firstly estimated. Then, it was segmented into different blobs motion. They proposed a rule-based technique for gestures recognition. Six motion rules corresponding to six gestures were defined in their system.

The use of optical flow in gesture recognition has not been extensive, probably because it assumes that the hand is the only moving object in the video sequence. This assumption requires a very controlled setup.

Recognition by modeling the states:

Another alternative to analyze and recognize dynamic gestures consists in representing the relationships between the different gestures by a state transition diagram [kun98]. This approach assumes that hand gestures will not happen independently.

Inspired by the success of the Hidden Markov Models (HMM) in speech recognition, some authors used this approach for gesture recognition task. The first attempt of gesture recognition based on HMM was introduced by Yamato et al. [Yam92] to recognize six different tennis strokes with a rate of 90%. The Hidden Markov Models is viewed as a stochastic process where the evolution is managed by a series of states that constitute a Markov chain which is not directly observed [Mar00]. This technique has been a popular statistical tool for modelling and recognizing sequential data.

Haberdar and Albayrak [Hab05] employed Hidden Markov Models to recognize 50 different gestures from the Turkish Sign Language. They used 500 samples to train the distinct HMMs. While the recognition rate of their work achieved 95.7%, their training phase has taken too much time.

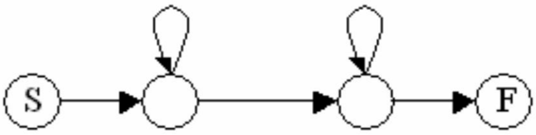


Figure 37 - The HMM topology used in [Hab05]

In [Nam96] the authors proposed a HMM-based method to recognize the space-time hand movement pattern. In their work, hand gestures are described in terms of three attributes: hand posture, hand orientation and hand movement. They modelled the hand movement as a sequence of prime gestures. A discrete HMM is constructed for each movement prime. The complex hand gestures are recognized by involving the connection or the repetition of some movements.

Even the use of HMM allows obtaining a high rate of recognition, the training step is always slow due the enumeration of all possible paths through the model.

3.1.3.1.4 Conclusion

Hand gestures recognition approaches and their principal steps preceding techniques have not been published recently and they are quite various as we have seen in this section. The previous research work of hand gesture recognition is reviewed from one dichotomy: appearance-based approaches versus 3D hand model-based approaches. In spite of the limited ability of appearance-based approaches to cover a wide range of hand gestures classes due to the viewpoint dependence and the simple features image used, these approaches allow achieving a real time performance thanks to the low computation of its 2D features calculation. 3D hand model-based approaches are ideal for realistic interaction in virtual environment but their accuracy and power are offered at the expense of the time consuming and the sensibility to the model initialization. Besides, the determination of model parameters via computer vision techniques is quite difficult due to the 3D structure complexity of high-level features.

It is interesting to involve the human perspective in the hand gesture recognition system. Taking into account all possible hand shape to detect the hand is a difficult task since the human hand has many degrees of freedom. The use of special markers or gloves and the assumption of a known background are not common in our daily life. The skin detection seems to be the most natural approach to detect the hand. Among the skin detection approaches, we have seen that the low computation cost of skin clustering method improve the system’s processing speed. However, the skin detection approaches are not sufficient to detect the hand since they can not differentiate between hands and the others objects presented in the scene and characterized by the skin color. In consequence, the cue color must be combined with another cue to perform a fast and natural hand detection approach.

3.1.4 Body expression for automatic emotion recognition

As the interest towards recognising human emotions is growing all the time the body movements and pose revealing the body expressions are less studied non verbal modality but growing interest is taken on them. It is disagreed how much body expressions and behaviour truly contribute on emotion perception, how much is controlled behaviour and how much emotion they convey, but there are evidence that the body expressions are more important factor in emotional expressions than previously thought [Sto07]. The human emotions are reflected to the body posture and movement and they can be used consciously to emphasis the verbal message (especially hands). There is a very extensive survey on affective body expressions [Kle12] that covers all different aspects from human perception to automatically recognised body expressions and emotions that they convey.

The emotions are currently modelled either with discrete emotions or in continuous space. Although there is no agreement of what are the basic emotions or discrete categories to be recognised the most famous is Ekman “basic emotions” fear, anger, sadness, happiness, disgust and surprise. When using continuous dimensions the emotional space often contains 2 or 3 dimension. In two dimensional modelling arousal and valence characterises the emotion as in [Glo11]. Arousal describes the emotion in calm/excited axel where the valence is the negative / positive dimension. The three dimensional model is often called PAD model (pleasure, arousal and dominance) as in [Gil09]. In this case dominance describes the how controlling and powerful the emotion is compared to submissive state.

There have been studies to understand how the body expressiveness maps either to discrete emotions or continuous space. In [Kle05] they studied human postures and the affective dimensions that human observers have when evaluating the emotion. They also mapped the emotion dimensions and postural expressions via low level features to affective space for creating a frame work for interpreting emotion of postures. In [Pat01] they mapped the upper body movement (head, shoulders and arms) to psychological space to model experienced affect. In automatic emotion recognition it is more common to map the emotions to discrete emotions rather than to emotional space.

Most of the research is done for acted emotions and only few papers can be found where non-acted expressions are studied. Many of the research with non-acted emotions are done with dance performances [Cam04] or gaming situations [Sil06]. Creating an emotion database is laborious and difficult task and finding the ground truth can be surprisingly difficult. When the emotions are acted the ground truth is given beforehand. In non-acted emotions the labelling of ground truth is not easy as the test subject cannot be interfered all the time and asking afterwards the memory might not be correct. The human observers are also used to detect the emotional state, but also the human observers cannot always tell or agree what the emotional state of the subject is. The technologies detecting movement varies from marker based to markerless computer vision based systems. Although the depth sensors or RGB-D sensor, that include both camera and depths sensor, are replacing plain video analysis in gesture recognition side there is not yet publication on affect computing domain that utilises depth sensors. For classifying human actions with depth sensors there are few examples found in [Sun12] and [Bin11]. In following part recent studies on emotion recognition from pose and movement is introduced.

Camera and marker based movement analysis was studied in [Kap05] in order distinguish four acted emotions: sadness, joy, anger and fear. The movement was analysed in 3D with aid of 14 markers that were positioned all over the body. Continuous features for velocity and acceleration of each point were calculated. The human observers achieved 93% detection rate as the automatic recognition yielded between 84 to 92% depending on the classifier.

Emotions in gaming were studied in [Sav11], [Sil06] and [Sav12]. De Silva & al. [Sil06] studied affective gestures of children in gaming context. They estimated the intensity of emotions that was used to adapt the game level. They recorded upper body Euler angles from 8 positions yielding 24 measurements to recognise gestures that were mapped to sadness, frustration, joyful and happy. The recognised emotions were used to model intensity and thus control the gaming parameters to reach good gaming experience.

Non-acted affective states of video game players from non-repeated body movements were studied in [Sav11].Eight emotion categories were used, frustration, anger, happiness, concentration, surprise, sadness, boredom and relief. The players had 17 acceleration sensors placed on their body as they played tennis with their friend. The best discriminative dynamic features were selected for emotion recognition including angular velocity, acceleration and frequency of right forearm, arm and hand, body directionality of spine and head, body rotation and angular speed of right forearm, arm and hand. They used Recurrent Neural Network algorithm (RNN). The automatic system accuracy yielded 57,46% as the observers resulted 61,49%.They expanded the research of emotions in gaming situation to aesthetic experience[Sav12], which loosely means comprehensive experience of movement and emotional engagement. They used the same database and methods than in [Sav11], but gave labels for emotional state: high negative, happy, concentrated and low negative classes as well as included an undecided class if the classifier was no clear winner of previous classes. Their system reached 61.1% accuracy where the observer's agreement was 61.49%.

Berhardt and Robinson studied in [Ber07] how emotions are present in everyday actions: knocking, lifting, throwing and walking. Their database contained four acted emotion classes neutral, happy, angry and sad that they classified with SVM (Support Vector Machine). They divided the each gesture for smaller primitives to understand the structure of the motions and removed the personal bias (average of all movement in database) in order to find out the elements carrying affect. They used movement of 15 body joints normalised and in local body coordinate system. The primitives are segmented through different energy levels in movement. For emotion recognising they used maximum distance of hand form body, average hand speed, acceleration and jerk and similar measures for elbow. They noticed that removing personal bias improved the recognition rate of the emotions from average rate of 50% to 81%. They further studied in [Ber09] the difference between recognising emotions from isolated gestures and more natural sequence of gestures and they noticed that isolated actions differ from actions made in sequence. They used HMM to model the action sequences by using global body speed, body-local joint positions and speed and body twist as features. The emotions were detected utilising their previous work in [Ber07]. For detecting emotions they used features mean and standard deviations of posture, joint speed, acceleration and jerk. They also tested extracting an individual movement bias that improved the recognition rate of the emotions from average rate of 52% to 81%.

Glowinski et al [Glo11] created a framework of affective behaviour of upper body movements based on visual information on acted emotions (GEMEP corpus). They used gestures accompanying utterances and in the three level analysis of the gestures. They started from measuring low-level motion features (position, speed etc.) and created a set of expressive features out of them including energy, spatial extent, smoothness/jerkiness, symmetry and leaning of the head. These features were further refined to dynamic-features to consider temporal aspects of affective gestures and they achieved 25 features vector. They use Principal Component Analysis (PCA) to reduce the dimensionality of feature vectors and it resulted retaining four components that were motion activity, temporal and spatial excursion, spatial extend and postural symmetry and motion discontinuity and jerkiness. They determined the optimal number of clusters (four) and compared it to ground truth of GEMEP corpus valence arousal space. The one of the main result of their work is that *”bodily expressions of different emotions may be dynamic rather that categorical in nature – concerning movement quality rather that specific type of gestures”* [Glo9]

3.1.4.1 Movement analysis

It has been shown by neuroscience and psychology that body movements are an important modality of emotional communication, notably head and hands movements. Only a few theories on gesture-based emotion recognition have emerged. The reason is certainly that in the past, most emotion recognition systems were based on classification into “universally” shared emotional categories. Such a classification is hardly possible when it comes to gestures, where the context is highly crucial. Let us observe that there is no equivalent of the FACS [Ekm02] for gestures.

Works have been done on coarse-grained posture. For example, in [Mot03], S. Mota and R. W. Picard use posture sequences and HMM to predict the interest of children during a learning task on a computer.

Other studies relate to body kinematics, following psychology showing that the dynamics play an important role in motion interpretation. In [Kap05], Kapur et al. use the velocities and acceleration of body joints shaping a skeleton to recognize 4 emotions (sadness, joy, anger, fear). The movement features are given as input of SVM or decision trees for the machine learning. These works followed the studies of D. Bernhardt and P. Robinson in [Ber07].

Work focused more specifically on upper-body movements. In [Bal04], T. Balomenos and al. combine facial features and particular hand movements and use them with HMM to classify image into 6 emotions. In [Gun07], H. Gunes and M. Piccardi use the variations of centroid, rotation, width, height and area for the hands, head and shoulders' regions as inputs of SVM to classify the images into 12 emotional categories.

In [Mar06], J. C. Martin and al. define a model of gesture expressiveness with 6 characterizations:

- Spatial Extent
- Temporal Extent
- Power
- Fluidity
- Repetition
- Global Activity

A first test is made to verify if these dimensions are recognized by human observers (if a change in a single dimension can be recognized and correctly attributed by users.). The spatial extent and the temporal extent are better recognized than the other features.

In a second test, is tested the hypothesis that combining parameters in such a way that they reflect a given communicative intent will result in a more believable overall impression of the agent. Three qualities are chosen: abrupt, sluggish and vigorous, each determined by a particular configuration of the paramaters of gestures defined above. It is asked to classify 4 instantiations of each of these qualities, from the most appropriate to the least appropriate with respect to the expressive intent. The abrupt et vigorous qualities are rightly classified.

In [Glo11], D. Glowinsky and al. build a three-modules model for a minimal representation of gesture based on hands and head movements.

- A first module extracts the “low-level” characteristics: head and hands' position and velocity.
- A second module builds 25 “high-level” features related to gesture expressiveness:
    - the energy (overall activity); for energy, dynamic features are computed relating the entire series composing the gesture
    - the spatial extent
    - the regularity, the smoothness of the gesture
    - the symmetry; A. Merhabian showed in [Mer07] that asymmetry in arm position can reveal a relax attitude or a high social position regarding the environment
    - the head movement.
- A third module performs the dimension reduction by Principal Component Analysis

Then, each eigenvector is used to gather the “close” gesture features. The different clusterings are compared to the emotional annotations of the training corpus, (the GEMEP corpus [GEM11], annotated following the “Valence-Arousal” space). It appears that most portrayals are clearly associated with one of the eigenvector.

3.1.5 Group behaviour and people tracking

Different public screens and digital signage systems in retail and advertising as well as digital installations are getting very common these days. Some of these systems are passive but some of them hold already different interaction methods. The user can either interact by gestures of their body with the content or the systems can be interactive with personal devices using for example Mobile phones with NFC (Near Field Communication) technologies. As the future systems will contain more and more interaction different methods to understand behaviour of the people are required. Depth sensing technologies can provide valuable tool for natural interaction and creating systems more intention aware.

With the computer vision and depth sensing technologies we can also provide intuitive interactions methods via gestures recognitions and automatically understanding more of the people behaviour or emotions as well as use this information in UI design of public displays [Mul10], [Bey11]. The multiuser public displays will challenge the natural gesture interfaces to find solutions to which the control is given. The interaction can be provided by “hot spot” of the main user, by selecting “who was first” or the system can divide to multiple applications that each user can control [Des09]. Also interactive public displays have to be aware of user or the user’s state of mind towards the interaction.

The audience funnel models [Mul10] describes the audience behaviour front of public displays. At the first stage the user is only passing the display, but in second stage they start to react to it by glancing, smiling or approaching towards it [Mul10]. Detecting this implicit behaviour of the user will enable developing more emphatic interactive screens. In this stage the interactive system should encourage the user for further interaction to engage the user. This will require understanding the actions and movements of users in space as well as users emotional state.

In intelligent spaces we can track the position of persons and know if they are approaching an interactive screen. Also from the tracking abilities we can derive intelligent conclusions of people intentions, depending on application context. For example detecting if a person is entering or leaving the building or store or actions taken by the users may distinguish between employees or visitor. This information can provide tools to create personal and profiled information systems.

When we understand better the people behaviour and movement (flow) in space we can tailor the content in digital displays. Computer vision and depth sensing technologies can provide valuable information that can be utilised in smart spaces including tracking people, understanding their actions and (emotional) behaviour and gestures for natural interaction.

People segmentation and tracking is a widely researched subject which has been started in context of security and surveillance. As the camera based algorithm research and development has been going on for long time there are quite extensive surveys available in [Yil06] and [Agg97]. Recent advance in depth sensing technologies has enhanced the algorithm development on that field as well. With depth sensors the people tracking is done on either from top-view or from the oblique angles [Jun12] (different side views). The depth data from top view is acquired either placing the depth sensor on the ceiling [Her11] or performing an affine transformation to the point cloud [Han08], [Alb12]. When only people counting and tracking is require the ceiling position for the senor is most likely the suitable position, but if any other information is needed (identity or people behaviour) the field of view from the ceiling is quite limited.

Herdandez et al. [Her11] calculated people entering and leaving the space with fixed camera and changing illumination conditions. Depth sensor was placed to ceiling to avoid occlusions. First they calculate the background model as well as estimate of different artefacts on depth image to be averaged as background. Background subtraction is used for calculating the region of interest. Remaining foreground pixels are presenting the passing people as “blobs”. The blobs are tracked with extended Kalman filter in four state vector representing X,Y and Z coordinates as well as motion in XY plane. The trajectory of each object is evaluated and decision made for in/out counters. They also use soft biometric method based on back projection of the head area to re-identify people. The system detection rate was greater than 95% with no false detection.

In [Han08] they tracked people with depth sensors placed on walls. They compared two methods: kernel density and Mixture of Gaussian (MoG) for creating background model and end up choosing the MoG since it requires less computation. Noise from the image was removed using statistical method over the time. Clusters for people tracking were creating an overview of depth image and projected the points in to the ground plane creating a flat-map. Clusters were tracked with the mixture of Gaussian clusters as the parameters for estimated with Expectation Maximisation (EM). The Gaussian Mixture Model included one mixture for background and one or more for the foreground. Their methods utilising statistical analysis do not require any threshold setting. They achieved 98 % accuracy in normal walking conditions. The rapid movements i.e. running and jumping caused errors.

Information from depth and RGB cameras are often fused to improve the accuracy and being new features in the system. In [Jun12] were created Home-Used human tracking systems that identified the person as well. The moving pixels are detected from background and used as seeds for clustering algorithm for object segmentation. The object is detected as human if it is stable and high enough based on predefined thresholds. If a human is detected a human visual signature based on RGB camera is calculated (only of the standing person). The visual signature contains the colour histogram and texture information of head, torso and legs which are identified from a standing person. The person can be re-identified based on this signature when he appears into to the scene again. The person in scene is tracked based on depth continuity and appearance similarity. They used RGB-D and a camera to different modules of systems and had two input channels that were fused to get the best results. Object labelling reach the 95 % in realistic environment and the persons could be re-identifying with accuracy of 80%. The tracker fusing information from depth and visual analysis achieved an accuracy rate of 96%.

Albiol et al [Alb12] created a system for tracking and re-identifying people with a RGB-D camera. Their work concentrates on re-identification and they have a good inside to different identification methods. The people segmentation and tracking was based on transforming the depth information to the height map (along z axis). When the required coordinate transformation is performed the height map is used to first subtract the background and then detect “blobs” from intensity of the height map. Tracking of blobs is done with the aid of distance matrix (containing Eucledian distances and predicted track) and if the current objects are below threshold they are detected and tracked as corresponding object. The identification is based on colour mean and variance as a function of height.

The depth sensors are a good tool for tracking people as they are more robust for changing illumination and lack of contrast as well as improve the performance of people tracking in some difficult cases. Depth sensors do have their problems as they generate artefacts and other reflection errors in some cases. In [Han08] they noted that bad configuration of the sensors caused errors or the sensors interfered with each other. In [Jun12] they noted that when using plain depth data from height map the segmentation and tracking will fail in situation where occluded people have similar depths. If the identification or labelling of the tracked people is required it would be convenient to use the RGB data as well. The depth sensor based people trackers are rarely evaluated in heavily populated environments (e.g. in entrance of malls) which is understandable due to the laborious work creating the ground data and verifying the results. Authors of [Alb12] did test their system in a supermarket, but the main emphasis was in re-identification of people.

In 1995, J. Kennedy and R. Eberhart, had been inspired by bird flocking and fish schooling to implement their Particule Swarm Optimization algorithm. They said that the trajectory of a bird (or a fish) was determined by its own memory of the best position he had occupied in the past (personal component) and its capacity to benefit of the movement of the other (birds or fishes) to know where to go. This second element relates to a social component.

A lot of recent studies focus on group activities, notably on the establishment of hierarchy. Body movements of the protagonists can help to detect this establishment.

It is the topic of [Glo10], where D. Glowinsky et al. work on the establishment of leadership (dominance) in a string quartett.
By following the head movements (very important in music to indicate the start or to drive a change of speed) of the 4 musicians while they are playing in different situations (rehearsal, rehearsal with change between the first and second violin, rehearsal with the metronome, concert situation, and rehearsal with the over-expressive constraint), they analyse the complexity of 5 time series: the 4 head's velocities' series, and the series of the area of the polygone built by the 4 heads. This complexity is given by the measure of the Samle Entropy introduced in [Ric00] by Richman and Randall Moorman, based on the probability that two sequences similar for a given number of points of the series remain similar when adding a new point.
They find that the complexity of the first violin's (leader) series is globally more correlated to the complexity of areas' series than it is correlated to complexity of the other musician's series.

3.1.6 Posture recognition

Body expressions perform many communicative functions. They may replace speech during dialogue, or when speech is not used at all. They may regulate the flow and rhythm of interaction, maintain attention, add emphasis or clarity to speech, help characterize and help speakers access and formulate speech. Birdwhistell’s analysis of nonverbal activity [Bir66] that accompanies verbal behavior led him to postulate the existence of kinesic markers. These nonverbal behaviors mark a specific oral language behavior. Kinesic markers are characterized by gross shifts in postural behavior, involving the whole body, indicating or marking a sequence of point(s) of view expressed by the speaker (e.g. the shift from leaning back when listening to leaning forward when speaking). The positions of the head, limbs, and body have been studied to forecast information to a listener, such as length of utterance, change in argument strategy or viewpoint. The observation that postural shifts mark new stages of interaction or topic shifts, particularly at the beginning or ending of speech segments, has been made by several researchers [Bul77; Eri75; Sch73].

Prior work has shown that forecasting upcoming components of speech through posture and gesture is a crucial function in social interaction. Montepare and Zebrowitz [Mon93] and Janssen et al. [Jan08] demonstrate that various emotions such as sadness, anger, and happiness can be accurately identified by a person’s gait. James discovered the importance of leaning direction, openness of the body, and head position for discriminating between several affective states. Wallbott and Scherer [Wal86] found arm, shoulder and head position afford the ability to distinguish between 14 emotions. Coulson [Coulson 2004] attempted to ground basic emotions into low-level static features that describe the configuration of posture. Picard’s studies examine non-acted postures with multimodal system to model a more complete description of the body, attempting to recognize discrete levels of interest [Kap04] and self-reported frustration [Kap07]. Of the input examined (facial expressions, body postures, and game state information), the highest recognition accuracy was obtained for posture. Kleinsmith et. Al [Kle11] demonstrated automated recognition of affective stage base on postures in game situations. They used an invasive whole-body exoskeleton to obtain the non-acted postures. Garber-Barron and Si [Gar12] were able to discern 4 different kinds of emotions (triumph, frustration, defeat, and concentration) based on body pose and movement with an accuracy of 66,5%.

3.1.7 Depth Sensing Solutions

3.1.7.1 Time-of-Flight Depth Sensing Technology

The 3D gesture recognition technology introduces a new paradigm for communication and human-machine interaction because it brings real-time processing of 3D images to practical use. Human vision, by means of eyes, builds up naturally the images that we “see” in 3D whereas for a regular camera this information needs to be inferred from an RGB signal representing the scene in 2D. The SoftKinetic system is based on a consumer ready time-of-flight ToF depth sensing camera (DS311 and DS325 DepthSense™ Camera), whereas the Microsoft system (Kinect) uses a structured light based depth sensor. The base principle of the used continuous ToF imaging is to measure the phase difference between out-going modulated infrared light and captured IR waves after they have been reflected by the measured scene. As ToF is a new technology, it has a whole lot of remaining challenges.

The Depth Sense 311 (DS311) is a time-of-flight (TOF) 3D camera. It is suitable for short (15cm to 100cm) as well as long range tracking (150cm to 400cm). It is an ideal product for indoor use - home entertainment use, serious gaming, advertising campaigns, healthcare and more. The DepthSense 325 (DS325) is designed as a PC peripheral, small in size and with powerful features. It has both 3D and 2D sensors which will enable you to create high-quality finger and hand tracking applications for touch free environments. The DS325 tracks with great precision hands or objects from 15 to 100 cm. It is powered by a single USB cable which gives you great flexibility for deployment.

	DS311 	DS325
Range Distance	Short 15cm-1.5m & Long 1.5m-4m+	Short 15cm-1m
Depth Sensor Resolution	QQVGA (160x120)	QVGA (320x240)
Color Sensor Resolution	VGA (640x480)	HD (720p)
Depth Field of View	57.3° x 42.0° x 73.8°	74°x 58°x 87° (H x V x D)
Depth Illumination Source	LED	Diffused Laser
Power & Data Connection	USB 2.0 and External Power	Single USB 2.0
Camera Size	24cm(w) 4cm(h) 5cm(d)	10.5cm(w) 3.0cm(h) 2.3cm(d)

3.1.7.2 Commercially available gesture recognition iisu^{TM} by SoftKinetic

The 3D gesture market has come finally on the verge of maturity in 2012, with Intel pushing the perceptual computing initiative including the SoftKinetic Close interaction solution for laptops. According to many analysts, this growth is related to the market adoptation of 3D smart television and mobile devices http://www.researchandmarkets.com/research/a54005/global_touchless_sensing_and_gesturing_market#. Users feel as if they are part of the natural interface to a gesture-based application since they do not have to use any controller nor marker to get a much more immersive and natural experience with the digital equipment around them. This has raised the interest of many developers: many controller-based interfaces are currently being replaced by hands-free human gesture recognition components, and new user experiences are being created.

Extracting depth information offers an array of possibilities to capture users' movements in real-time. The mission of Softkinetic Software is indeed to transform the way people interact with the digital world, and to provide the standard middleware platform for next generation human computer interactions. Softkinetic Software delivers a commercial 3D body real-time gesture recognition software and tool, bundled in the software development kit (SDK) iisu™ (the interface is you). iisu™ is a gesture recognition middleware that enables communication between depth-sensing cameras and end-user applications. iisu^{TM} is a complete platform for natural gesture development and deployment. iisu™ helps application developers to add gesture recognition functionality to their algorithms, or to define customized gesture sets. It allows developers to build immersive and intuitive applications in record time.

- iisu's far interaction library provides multi-user tracking of up to four players, captures the user's movements in real-time with volumetric information, individual body parts and full body skeleton extraction. iisu also identifies the main user and computes statistical data, including user height, body and torso orientations, center of mass, chest and pelvis direction, etc.
- Iisu's close interaction library provides position and orientation tracking of 2 hands palm, individual finger tips tracking and poses recognition. Multiple mechanisms to catch, hold, move, release and click on objects such as two-finger pinching, full hand grasping, open and close are seamlessly recognized independently of the hand orientation and position. These close interaction methods are also included in Intel's perceptual computing offer on http://intel.com/software/perceptual .

All SoftKinetic branded depth-sensing solutions are commercially available from the website www.softkinetic.com.

Applications and services have been successfully introduced to companies in the fields of Digital Entertainment, Serious Games, Digital rehabilitation and training, Interactive Marketing and Consumer Electronics. www.iisu.com provides links to demonstration videos. Softkinetic Software is worldwide cooperating with major ICT, as Intel and TI, and entertainment companies and with several Universities to develop powerful human-computer interaction tools, and to strengthen its image analysis technologies.

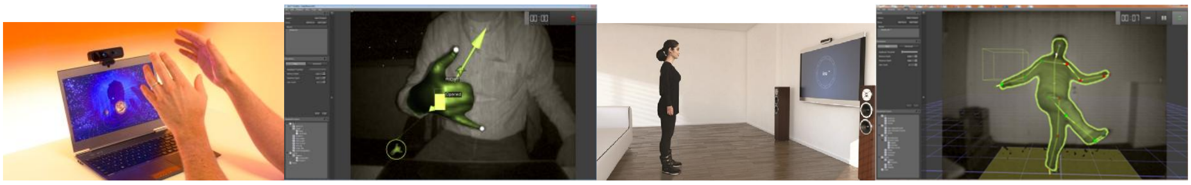


Figure 38 – Close and far range depth sensing solutions by SoftKinetic

3.1.7.3 Depth sensor based gesture recognition, especially differences in far and close field gestures (VTT)

3.1.8 Physiological affective wearables

The term affective wearable was coined by Rosalind Picard [Pic97] as a supplementary element in her theory of affective computing. An affective wearable is system that recognizes affective patterns by means of wearable sensing devices and tools. The most popular choice to realize an affective wearable is through the use of physiological sensors attached to the body or embedded in clothing. Bodily signals are collected via the sensors and analysed to identify patterns associated with emotional states. Affective wearables establish a relationship with the wearers and their preferences though perception of emotional states instead of traditional computer interfaces.

Affective wearables are appealing to researchers in many scientific areas because they enable unobtrusive monitoring of emotional states thereby facilitating the study of emotional factors in technological fields such as pervasive computing, user modelling and HCI. They also allow interdisciplinary collaboration where emotions are studied in their natural milieu with the minimal need for human intervention.

Thanks to recent advances in sensing interfaces, communications and information processing methods, affective wearables have become a burgeoning industrial area.

3.1.8.1 Commercially available affective wearables

Sensors are an important component in the design of an affective wearable. They should not only be robust and reliable but, ideally, they should also be designed to operate wirelessly for long periods of time, provide real-time data, be worn comfortably and minimize disruption of the wearer's regular activities. Some physiological sensing devices suitable for emotion detection include:

	Bodymedia Inc. offers the Sensewear system (http://www.sensewear.com) to record data for GSR, T, Heat flux and MOT. The sensewear is implemented as an armband with local storage capability which is mainly used to provide feedback about weight and metabolic disorders by estimating energy consumption and health behaviour. Sensewear measures skin temperature, Galvanic Skin Response (GSR), Heat Flux and motion via a 3-Axis Accelerometer.
	The MiniClinic Wrist-unit by Telcomed (http://www.telcomed.ie/wristwatch.html) allows monitoring of HR, T, Heart rhythm, Breathing rate, and ECG. The waterproof wristwatch can measure, store and wirelessly transmit (at certain intervals) physiological data up to 300 meters using radiofrequency. The MiniCinic is designed to work in conjunction with or support tele-care systems and is thus equipped with an alarm button to emit a distress alarm signal.
	Alive Technologies' Alive Heart Monitor (http://www.alivetec.com/pdf/heartmonitor_handout.pdf) is a Blue tooth-based wireless artefact used to monitor heart activity with the aim of supporting training and rehabilitation programs. The AHM provides real-time measurements and local storage for ECG, HR, activity, and body position by means of two electrodes and 2-3 axis accelerometers. The system could also be complemented with GPS tracking.
	The Lifeshirt was designed by VivonoeticsInc (http://vivonoetics.com/products/sensors/lifeshirt/) to enable non-invasive 24 hours monitoring of vital signs. As the name implies, this device takes the form of a sleeveless shirt that can be worn even in bed and for long periods of time. LifeShirt allows measurement of all aspects of respiratory function using two Respiratory Inductive Plethysmography (RIP) bands with the ability to observe and analyze this data in the context of other parameters such as EEG, ECG/EKG, HRV (Heart Rate variability), Temperature, Posture, Activity, Blood Pressure, Electronic Diary inputs and other additional signals like GSR. Data can be analyzed in real time with one second delay

	The wireless M-Multi Biofeedback from Bio-medical Instruments Inc. (http://bio-medical.com/products/wireless-m-multi-biofeedback-system.html) consists of a set of electrodes that provide information for skin conductance (GSR), temperature (T) and pulse (P) and internal 3D acceleration sensors that measure motility (MOT). Portability is given by Bluetooth connectivity and hook-and-loop patch which can be used to attach the acquisition module to a hook-and-loop strap on the body. The M-Multi Biofeedback can interact with Biofeedback 2000 X-pert Software to provide different kinds of therapies including Threshold value training where a user sets physical thresholds which if achieved are rewarded with a positive stimulus in the form of sounds, music or fairy tales, Audio feedback where changes in sensor readings are fed back acoustically, and Volume feedback which aims at interactively reducing blood flow in the temporal artery

3.1.8.2 Research on Affective Wearables and related systems

Although research on affective wearables as defined by Picard is scarce there are few examples that can be mentioned. Note that although the development of wearable sensors is different from the implementation of an affective wearable the two areas era interrelated and therefore some relevant examples of wearable sensors are also mentioned.

- “Conductor’s Jacket” - Marrin and Picard, M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 470, October 1998.The conductor’s jacket is an array of sensors embedded into a specially adapted shirt designed to appropriately recognize and respond to expressive features in musical gestures. The jacket collects physiological and motion information from musicians in order to better understand how their gestures and physiology are modulated while they perform.The Conductor´s jacket is equipped with sensors to measure electromyography, respiration, temperature, skin conductance, heart rate and also body position. Using these signals thirty expressive gestures were identified which could be transformed into real-time expressive shaping of MIDI music.
- “Affective DJ” - Healy, Picard and Dabek, Proceedings of the 1998 Workshop on Perceptual User Interfaces, November 1998.The Affective DJ is a wearable computer that recognizes affective and behavioural information using physiological signals and establishes a relationship between emotional states and musical preferences. This affective wearable comprises a skin conductance sensor located on the palm of the hand and a palm pilot PDA to process data. The Affective DJ recognizes affective states by means of measuring changes in the skin conductance and then adjusts song selection based on the measured emotional state.
- “StartleCam” - Healy, Second International Symposium on wearable computers October 1998. Using measures associated with changes in the skin conductance the StartleCam wearable system looks for patterns of startle response and automatically collects pictures for situations that catch the user’s attention. The StartleCam comprises a digital camera, physiological sensors to collect data for skin conductance, and a wireless connection to the internet. The method to indentify the startle response is based on a series of signal processing algorithms including convolution and digital filtering. This approach allows the StartleCam to react to the user’s real-time psychophysiological reactions with selective camera recording.
- “An integrated telemedicine platform for the assessment of affective physiological states” - Katsis, Ganiatsas and Fotiadis, Diagnostic Pathology, 1 (16), 2006. This system called AUBADE utilizes Facial Electromyogram (EMG), Respiration, Electrodermal Activity (GSR) and Electrocardiogram (ECG) to recognize affective states in real time under extreme stress conditions, e.g. driving a racing car or suffering from neurological or psychological disorders. Once an emotion is detected AUBADE provides an animated facial representation of the emotion being experienced. Most importantly, AUBADE makes use of a wearable module to collect and transmit physiological data from sensors. The wearable is composed of a mask containing sixteen EMG textile fireproof sensors, a three-lead ECG and Respiration sensors located on the thorax of the driver and a EDA textile and fireproof sensor placed inside a drivers glove.
- “HandWave Bluetooth Skin Conductance Sensor” – Strauss, Reynolds, Hughes, Park, McDarby and Picard, The 1st International Conference on Affective Computing and Intelligent

Interaction, Beijing, China, October 2005. Handwave is a wearable skin conductance sensor designed for affective computing applications. It takes the form of a wristband and is based on Bluetooth technology and standard battery power.

- “Monitoring Stress and Heart Health with a Phone and Wearable Computer” - Picard and Du, Motorola Offspring Journal, 1, November 2002. Comprising of a sensing chest strap and wristwatch, a processing module (Motohub), a mobile phone and a PC computer this wearable system measures real-time stress levels using HR variability. Motohub hub uses RF transmissions emitted by the sensors and communicates physiological data through a serial cable to the mobile phone. The phone sends instantaneous HR over to a PC where data is processed and then returned in the form of HR entropy chart where stressful episodes can be indentified. The aim of this wearable system is to provide continuous feedback about the wearer’s stress levels and cardiovascular activity.
- “EREC I and II - Emotion Recognition System” - Peter, Ebert, and Beikirch, Physiological Sensing for Affective Computing. In: Tao J. and Tan T.: Affective Information Processing, pp. 303-322. Springer Verlag Berlin Heidelberg, 2008. The ERECs integrate GSR, Skin Temperature, HR and external temperature sensing into a common glove. Physiological data are pre-processed (checked for errors) by an internal module located on the glove and then sent onto a base unit using ISM-band (RF) transmitter. The ERECs also feature a boxed GUI with lights to indicate sensor and system state and push buttons to mark special events.
- “Sensor Sleeve” - Randell, Anderson, Muller, Moore, Brock, Baurley, The Sensor Sleeve: Sensing Affective Gestures, Workshop on On-Body Sensing, Osaka, Japan, October 2005. This approach for an affective wearable comprises touch and pressure sensors realized through textile circuitry integrated into the sleeve of a garment along with a microprocessor and Bluetooth transmitters. The purpose of this artefact is to enable recognition of expressive gestures like embracing, squeezing/pressing and stroking known to carry emotional content.
- “Multi-sensorial wearable” - Katsis, Katertsidis, Ganiatsas, and Fotiadis, Toward Emotion Recognition in Car-Racing Drivers:A Biosignal Processing Approach, IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, 38(3), May 2008. The authors of this paper introduce a multisensorial wearable to measure high stress, low stress, disappointment, and euphoria in car-racing drivers. The wearable collects data from 4 physiological signals using sensors distributed across the driver’s body: A balaclava containing textile EMG sensors, ECG and respiration sensors located on the chest, and textile GSR sensors embedded inside a driver’s glove. A data acquisition and wireless communication module placed inside the car collects, filters, preprocesses and transmits data onto a PC where information is classified into emotional categories.
- “Wearable Sensor Glove” – Lee, Yoon, Lee, Lee, Wearable EDA Sensor Gloves using Conducting Fabric and Embedded System, 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, August 30-September 2006. This is a prototype of a wearable sensing device designed to measure Skin Conductance Level and Skin conductance response. Signals are acquired through a glove made of conducting fabric that is connected to a measurement system via fabric lines. The measurement system filters the signals before they are pre-processed by an embedded system and subsequently the transmitted to a host PC using Bluetooth technology.
- “MARSIAN Modular Autonomous Recorder System For Measurement Of Autonomic Nervous System Activity” - Dittmar, Meffre, De Oliveira, Gehin, Delhomme, Wearable Medical Devices Using Textile and Flexible Technologies for Ambulatory Monitoring, 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, September 2003. This project relates to the realization of an affective wearable in the form of a glove and a shirt that provide real-time measurement for skin temperature, skin blood flow, skin potential, skin resistance and respiration. MARSIAN comprises 4 modules: A smart glove with textile sensors, a shirt with ECG and respiration sensors, a wrist unit to collect, preprocess and transmits data using an RF link, and a PC-based data logger. Power supply is configurable and it also features an event marker to support real life experimentation.
- “Wearable Systems for Service based on Physiological Signals” – Ryoo, Kim and Lee., 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2005. The authors describe a wearable system to interact with a pervasive environment using physiological and emotional information The system is composed of 3 modules. The physiological data sensing module collects, filters, amplifies and converts and wirelessly transmits data from 3 sensors measuring Photoplethysmography (PPG), galvanic skin Response (GSR), and Skin temperature (T). The human status awareness system uses a Wearable Personal Station (PDA) to extract statistical features from physiological data and

use such features to derive emotional states. The service management system uses emotional information to activate services from the environment.

- “Wireless Sensor Network for Wearable Physiological Monitoring” – Pandian, Safeer, Gupta, Shakunthala, Sundersheshu and Padaki, Journal Of Networks, 3(5), May 2008. This approach to remote physiological monitoring is based on a network of sensor nodes that measure Electrocardiogram (ECG), heart rate (HR), blood pressure (BP), body temperature (T), GSR, Oxygen saturation in blood (SaO2), respiratory rate, EMG, EEG and three axis movement of the subject measured using an accelerometer. The sensor network is connected to a wearable data acquisition system (located also on the vest) through wires woven into the fabric. The data acquisition module receives, collects, processes, stores and transmits physiological data using a RF link.
- ”WEALTHY (Wearable Health Care System)” - Paradiso, Loriga, and Taccini “Wearable Health Care System for Vital Signs Monitoring”, Medicon 2004. WEALTHY is a piece of clothing made of piezoresistive and metal-based yarn designed to measure physiological signals and biomechanical activity. Data for ECG, EMG, and respiration is acquired by a Portable Patient Unit (PPT) through embedded conductive tracks. The PPT performs signal conditioning, A/D conversion, pre-processing and GPRS-based transmission. Wealthy is currently undergoing pre-commercialization tests at Smartex, Italy.
- AMON (Advanced care and alert portable telemedical MONitor) - Anliker, Ward, Lukowicz, Tröster, Dolveck, Baer, Keita, Schenker, Catarsi, Coluccini, Belardinelli, Shklarski, Alon, Hirt, Schmid, and Vuskovic, AMON: A Wearable Multiparameter Medical Monitoring and Alert System, IEEE Transactions on Information Technology in Biomedicine, 8 (4), December 2004. AMON integrates a wrist-worn unit and a stationary unit at a remote location. The wearable unit continuously measures pulse, oxygen saturation (SpO2), skin temperature and activity via acceleration. It also filters data, converts raw values into medical units (the computation of blood pressure from the pressure sensor readings) and provides an initial evaluation of parameters using predefined patient-specific values. Data is subsequently transmitted to the remote unit using a GSM connection. AMON was launched as a commercial product under the name of MDKeeper by Tadiran Lifecare (later acquired by Aerotel Medical Systems). MDKeeper is no longer available.
- HealthGear - Oliver and Flores-Mangas, “HealthGear: A Real-time Wearable System for Monitoring and Analyzing Physiological Signals”. HealthGear is a prototype developed by Microsoft Research to provide wireless, uninterrupted measurement of physiological signals. Three modules integrate the Heathgear in its current state: An oxymetry sensor to measure blood oxygen levels (SpO2) and heart rate, a data transmission module and a cell phone. The data transmission module processes raw data and transmits it as a serial data stream to the cell phones using a bluetooth transmitter. Real-time bodily values are displayed on the cell phone’s screen along with the number of sleep apnea events as determined by SpO2 levels.
- Lifeminder - Ouchi, Suzuki and Doi, LifeMinder: A Wearable Healthcare Support System Using User’s Context. Lifeminder consists of two components: 1) a wearable sensor head measuring HR, Temperature, GSR and movement, 2) a wrist-worn signal processing and communications module, and 3) a PDA that displays bodily data and identifies eating behaviour. The wrist-worn module includes a microcontroller that estimates HR information from pulse waves and uses accelerometer information to recognize four different activities (“Walking”, “Running”, “Working”, and “Quiet”). Communication between the wrist-worn module and the PDA is realized through a Bluetooth connection.
- Lifeguard - Montgomery, Mundt,Thonier, Tellier, Udoh, Barker, Ricks, Giovangrandi, Davies, Cagle, Swain, Hines, Kovacs: A Personal Physiological Monitor for Extreme Environments. Lifeguard is a body-worn system for measuring physiological signals in an unobtrusive, noninvasive and easy-to-use way. Lifeguard comprises a number of wearable sensors (ECG/Respiration electrodes, a pulse oximeter, a blood Pressure Monitor, a temperature probe), a wearable data logger that includes accelerometers (CPOD, Crew Physiological Observation Device), and a computer base station where data is displayed. CPOD collects data from sensors and transmits them onto the base station using a Bluetooth or RS232 connection. Data processing algorithms on the base station derive heart rate from the received ECG waveform and respiration rate from the received respiration waveform.

3.1.8.3 Application to immersion measurement in communications

Physiological measurements have been applied to the study of user attention and emotions while watching videos or listening audio signals [Meehan.,M. (2001). Physiological reaction as an objective measure of presence in virtual environments. Doctoral Dissertation. University of North Carolina at Chapel Hill].

Two studies [Wilson, G., Sasse, M.A. (2000). Do users always know what's good for them utilising physiological responses to assess media quality. In: McDonald, S., Waern, Y., Cockton, G. (Eds.), People and Computers XIV-Usability or Else! Proceedings of HCI. Springer, Berlin, pp. 327–339], [Ward, R.D., Marsden, P.H. (2003). Physiological responses to different web-page designs. International Journal of Human–Computer Studies. 59 (1), 199–212] used skin conductance and cardiovascular measurements to understand the effects of video and audio quality decrease in videoconferences.

Some previous research show that the heart rate decreases and attention is focused or when information is captured [Lacey Lacey, J. I., & Lacey, B. C. (1970); Some autonomic-central nervous system interrelationships; In P. Black (Ed.), Physiological correlates of emotion (pp. 205–227); New York: Academic]; [Lang, A. (1990); Involuntary attention and physiological arousal evoked by structural features and emotional content in TV commercials; Communication Research; 17, 275–299]; [Mulder, G., & Mulder, L. J. (1981); Information processing and cardiovascular control; Psychophysiolog.18, 392–402]; [Turpin, G. (1986); Effects of stimulus intensity on autonomic responding: The problem of differentiating orienting and defense reflexes; Psychophysiolog.23, 1–14]; [Papillo, J.F., Shapiro, D. (1990); The cardiovascular system. In: Cacioppo,J.T., Tassinary, L.G. (Eds.); Principles of Psychophysiology; Physical, Social, and Inferential Elements; Cambridge University Press. Cambridge. pp. 456–512].

A more recent study [Mandryk, R. L., Inkepn, K. M., and Calvert, W. (2006); Using psychophysiological techniques to measure user experience with entertainment technologies; Behavior & Information Technology; 25(2), 141-158] demonstrates a significant correlation between GSR response and feeling of fun for adults playing videogames.

3.2 Sentiment Analysis from Textual Information

In this section we present a brief state of the art of sentiment analysis and emotion detection from text, these fields are relatively recent but are growing fast. We refer the reader to several complete surveys [Pan08; Liu12] for sentiment analysis and [Cal10] for emotion detection.

3.2.1.1 Some difficulties of Sentiment Analysis

The task of determining the sentiment (positive, negative or neutral) of a sentence or a document is considered to be a hard task. The primary reason is that there are many phenomena that influence the polarity of a sentiment. [Pol04; Wil09; Liu12] describe in details the syntactic or semantic aspects of polarity influence: negation is the most important influencer, locally as in "not good", with a longer distance "does not look very good", or subject negation "no one thinks it's good". But it is not sufficient to detect the presence of these negation words, the whole co-text needs to be taken into account as with intensifiers "not only good but amazing" (positive). The polarity also depends on the certainty of existence of an entity (a phenomenon more generally called *irrealis*), it is present in conditional clauses like "if it's not good then ..." (neutral) or "it would be good if ..." (neutral or negative). The reality of an entity is not easy to determine, *irrealis* is also present in interrogative sentences, for instance "can you give me a good advice ?" is neutral because the existence of a "good advice" is not certain whereas "does anyone know how to repair this terrible printer ?" is negative because the existence of a "terrible printer" is presupposed. All these examples show that it is by far not sufficient to detect the word "good" to mark a sentence as positive.

Syntactic phenomena alone cannot explain the polarity, there are also semantic and pragmatic phenomena at hand. For instance, it is required to disambiguate the sense of "suck" in: "this camera sucks" (negative) versus "this vacuum cleaner sucks" (positive). A problematic aspect is also that the point of view of the holder needs to be taken into account like in "Israel failed to defeat Hezbollah" (positive for the Hezbollah point of view, negative for Israel), but the holder is not mentioned directly in this example. The most difficult aspect is probably the absence of polarized items which then makes necessary to include common sense knowledge like in "this washer uses a lot of water" which does not contain any negative or positive word *per se*, but in which we can assume that "a lot of water" is negative. Another example actually found in Twitter : "You know Apple, it's been almost a week since I paid for iTunes Match. I would really like to use it. Any ETA on a fix?"; to determine the negative polarity of this text it would be required to know that it is a common desire to be able to use a product right after buying it, and that failing to do so is a negative situation.

Actually the problem is even harder if we consider the five slots of the opinion definition given in [Liu12][1] because a single sentence or document may contain several positive or negative opinions related to different entities as in "I hate cheese but I love cheesecake", different aspects of a single entity "this camera is awesome but it is too expensive" (the camera as a whole being positive while the price being negative), different opinion holders as in "I love cheese but my sister hates it", or different time as in "I used to love cheese but now I hate it". The involved difficulties are related to the general problems of natural language processing: sentence tokenization, parsing, reference resolution, word sense disambiguation, named entity recognition, semantic and temporal representation, etc. That is why the problem of opinion mining and sentiment analysis is often reduced, for instance to *(e, s, h)* like in [Kim04; Kim06], the entity, the sentiment value and the holder, but most often systems only return *(s)* a single sentiment value for a whole sentence or document.

3.2.1.2 Machine Learning for Sentiment Analysis

It is not a surprise then that the first approaches to the problem of Sentiment Analysis make use of machine learning, which avoids providing an explicit account of all the complex phenomena but rather tries to learn the expected polarity from corpora. An early approach from [Hat97] explores the learning of polarity from corpora but is restricted to adjectives. Two important papers are [Tur02] and [Pan02]. The former is using unsupervised machine learning : by defining a notion of distance between words (with Pointwise Mutual Information), reviews about products are evaluated with regards to their distance to the words "excellent" and "poor" to determine their polarity. The accuracy of this algorithm differs according to the domain (from 84% for automobile reviews to 66% for movie reviews). The strongest aspect of unsupervised learning is the absence of manual annotations of reviews. On the contrary [Pan02] makes use of the rating that is associated to a movie review (for instance a number of stars), the assumption being that the highest the rating the more positive the review. The idea is then to train supervised machine learning algorithms on the correspondance of textual data to the corresponding rating. They tried several methods (Naive Bayes, Support Vector Machine, Maximum Entropy) to find out that the three methods obtain close accuracy on the movie reviews domain (around 83% for the best). One key observation for the movie review domain is that people tend to write their reviews by holding their actual opinion until the end[2] (e.g. "I admit it's a really awful movie […] but I loved it") and as such the movie domain may not be the easiest domain to work on. Following the [Tur02] and [Pan02] papers, there are many work that approach sentiment analysis from the machine learning perspective, a detailed survey can be found in [Liu12].

A generic improvement in machine learning approaches to sentiment analysis is to perform machine learning in several steps by decomposing the problem in subproblems. For instance [Yu03] first classifies documents into factual documents and opinionated documents and in a second step classify sentences in opinionated documents into positive and negative sentences. Similarly [Wil05] performs two steps, first the sentences are classified along the polarized dimension (whether sentences are polarized or not) and then polarized sentences are classified along the dimension positive/negative. Another approach [Li10] is to separate documents regarding the presence of polarity modification, separating documents that contain non-shifted polarity ("this is good") versus documents that contain shifted polarity ("this is not good"), then in a second step training different polarity classifiers on the different kinds of documents and finally constructing a multiple classifier system. A more complex version of this approach can be found in [Pak12] in which classifier trees are used.

[1] that is an opinion is a quintuple *(e, a, s, h, t)* where *e* is the name of an entity, *a* an aspect of this entity, *s* is a sentiment about *a*, *h* is the opinion holder and *t* is the time when an opinion is expressed by *h*.

[2] this observation may lead to weighting schemes that give higher weights to the last parts of a text as is done in [Tab11].

Recent work in machine learning approaches to sentiment analysis explore successfully different kinds of learning algorithms such as Conditional Random Fields [Nak10] or autoencoders which are a kind of Neural Networks and which offer good performance on the movie reviews domain [Soc11].

Although machine learning enables to have good results without having to explicit all the complex rules involved in sentiment analysis, these approaches are not without defects. For one, they need datas, and in general a lot of datas. Actually the more features that need to be taken into account, the bigger the training corpus needs to be. The manual annotation of large corpora being labor-intensive, most approaches in supervised learning try to obtain the annotation from existing sources, movie review ratings in [Pan02], metadata from newspaper articles in [Yu03], emoticons [Rea05; Pak10] or twitter hashtags [Dav10], however we may question the quality of those annotations since they were not handcrafted for the purpose of sentiment analysis.

Another important issue of machine learning is the dependence to the training domain: topic and domain dependence, where a machine learning algorithm trained on a particular topic or domain does not perform as good in another domain (for instance an "unpredictable plot" for a movie might be positive while an "unpredictable steering" for a car might be negative, from [Tur02]), or temporal dependence where an algorithm trained on a particular span of time in a timely ordered corpus does not perform as good on other spans of time [Rea05]. In [Aue05] several methods are tried to overcome the domain-dependence of machine learning and they show that the best results can be obtained by combining small amounts of labeled data from the training domain and large amounts of unlabeled datas in the target domain. Actually, unsupervised or semi-supervised machine learning seems more adequate than purely supervised machine learning to reach domain-independence [Rea09]. Another approach [And08] is to use hybrid methods, classifiers trained on corpora and polarity lexicons. Indeed, polarity lexicons, being in general domain-independent seem to be an interesting track to follow.

3.2.1.3 Polarity Lexicons

The earliest lexicon is probably the one that was used in the *General Inquirer* system [Sto66] which contains around 12,000 lemmas annotated with rich semantic information including pleasure, arousal, and valence (around 2000 positive entries and 2300 negative entries). Another lexicon is the *Affective Norms* lexicon [Bra99] in which subjects annotated manually around 1,000 words with respect to the pleasure (happiness/unhappiness), arousal (calm/excited) and dominance (controlled/in control) that they felt when reading them. While much more limited than the General Inquirer lexicon, the Affective Norms lexicon contains scalar values (between 1 and 9) whereas the General Inquirer only contains binary information. The *OpinionFinder Subjectivity Lexicon* [Wil05] is an example of lexicon aggregation. It contains around 8,000 lemmas including subjectivity clues from [Ril03] expanded with a dictionary and a thesaurus, plus positive and negative words from the General Inquirer lexicon. Lemmas were annotated manually with regards to their subjectivity (weak or strong) and their polarity (positive, negative, both or neutral) if they originated from a lexicon without such annotations. The manual design of the OpinionFinder lexicon makes it an interesting and reliable resource.

The famous *WordNet* lexicon [Mil95] is a tremendous resource for many research work. It is composed of around 150,000 lemmas organized around sets of synonyms, aka synsets, and includes several kinds of semantic relationships between synsets such as synonymy, antonymy, hyperonymy and meronymy. However WordNet does not originally contain polarity information. One early idea to build a polarity lexicon found in [Kim04; Hu04] is to use a method we could call lexicon-crawling which consists of traversing WordNet along synonymy and antonymy relationships, starting from seed words with known polarity. This is the same approach followed by *SentiWordNet* [Esu06; Bac10] but completed with a machine learning algorithm trained on the glosses of synsets. SentiWordNet is the largest available polarity lexicon, it is a complete mapping of WordNet, in which the 150,000 lemmas are associated to a positive and a negative scalar value (between 0 and 1). However, like in WordNet, lemmas are grouped together around synsets, which makes necessary a word sense disambiguation phase before using the lexicon (see [Nav09] for a survey of the field). For instance, in the sentence "she spoke to him softly", it is required to determine whether *softly* refers to *softly_1* which is mostly negative, *softly_2* which is mostly positive or *softly_3* which is neutral. While interesting for its coverage, the necessary word sense disambiguation phase makes SentiWordNet a difficult resource to use. A very close approach which also involves lexicon crawling is found in *Q-WordNet* [Age10] which seems to offer improvements over SentiWordNet, but share the same deficiencies. An interesting survey of lexicon-related methods can be found in [Tab11], they show both that lexicon-based methods prove to be interesting for domain-independence since they manage to obtain consistent results on different domains, and that the biggest lexicons are not necessary the best lexicons.

3.2.1.4 Syntax and Valence Shifting

When analyzing the sentiment conveyed by a whole document, syntax and negation may not be of prime importance because of the redundance of a sentiment, missing one occurrence may not be as problematic as when analyzing sentiments in clauses or individual sentences. That is why bag-of-words approaches or simple negation tagging techniques as in [Das01; Pan02] are not sufficient for sentence level sentiment analysis [Wie10]. One first model of syntactic valence shifting is found in [Pol04] in which is introduced a broad category of *valence shifters* that are words or constructions that can intensify or reverse the prior polarity from the lexicon. Following this model, [Moi07] propose compositional sentiment rules that modify the prior polarity by reversing it or propagating it while traversing the syntactic parse tree. Their algorithm offers good results but parsing or word sense ambiguity are important sources of errors. A similar approach with manually designed rules but working on triplets rather than parse trees can be found in [Sha07]. The valence shifting model can cover more than negation as in [Tab11] where there are rules to handle intensification, negation and irrealis. Alternatively the syntax-semantics rules that influence polarity may be learned, as in [Cho08] in which compositional semantics is directly embedded in the learning algorithm (a modified Support Vector Machine algorithm). Syntactic features may also be embedded in classical machine learning algorithms as input features like in [Wie09; Nak10]. All these approaches are interesting in that they try to model fine-grained syntactic polarity influence, but we should note that they all are dependent on the quality of the input parses.

3.2.1.5 Sentiment Analysis and Emotion Detection

In some domains, the detection of polarity is not enough and we might want to have finer grained information by detecting the emotions conveyed by a text, typically the Ekman's emotions [Ekm72]: joy, sadness, fear, disgust, anger, surprise, interest[3]. Research work on emotions themselves is not recent, but early research on emotions focused more on how to model emotions in agents rather than how to detect emotions from the user, for instance PARRY [Col81] that models a paranoid personality involving fear, anger or mistrust, or the *Affective Reasoner* [Ell92] which enables to infer emotions from world simulations. There is a number of studies on emotion detection from other modalities than text (speech, gesture, face, etc.) and from other points of view than computer science, for further reading we refer the reader to [Cow01; Cal10]. As for emotion detection from text, it is a recent field which involves approaches very close to sentiment analysis but which has not yet reached its level of maturation.

One explanation could be that in comparison to sentiment analysis, emotion detection from text is widely considered to be more difficult. When trying machine learning on emotion detection, [Alm05] acknowledge that supervised learning techniques require a larger annotated corpus than when doing sentiment analysis, moreover they also observe that when manually annotating Ekman's emotions the inter-annotator agreement is rather low. The difficulty of the task is confirmed in [Mis05] in which human annotators achieve only 63% accuracy when annotating moods in blog posts while a Support Vector Machine algorithm reaches an average of 58% accuracy. The highly subjective aspect of emotions, sparse training data and the very high number of emotion categories used in this case (132 moods) may explain these results.

Similarly to polarity lexicons, there exists *emotion lexicons* which associate emotions, typically Ekman's ones, to words. The most known lexicon is *WordNet-Affect* [Str04] which adds emotional categories including Ekman's emotions to WordNet synsets (around 4,800 words). The General Inquirer [Sto66] also contains categories related to emotions such as pleasure, arousal or pain but is less used, it is also the case of the Affective Norms lexicon [Bra99]. The *EmoLex* lexicon [Moh10] is built with crowdsourcing and contains 2,000 entries associated to Ekman's emotions, it seems to

[3] In some domains restricted to six, five or four emotions (joy, sadness, anger, fear)

reach high agreement but its coverage is limited. There is also *EmotiNet* [Bal11] which is a knowledge base built using the ISEAR corpus, a dataset that contains around 7,700 sentences manually annotated with seven emotions [Sch94].

One of the first approaches to emotion detection [Liu03] actually uses such knowledge base, the *Open Mind Common Sense* knowledge base [Sin02] annotated with Ekman's emotions using a method similar to lexicon-crawling, and manual rules. Interestingly the authors propose to model meta-emotions in text as sequences of emotions (frustration would be the repetition of low-magnitude anger while relief would be fear followed by happiness). This approach is similar to [Lu02] in which manually designed semantic rules are applied after a semantic role labelling step. We can guess however that these manual approaches do not scale very well, and this is probably why machine learning is as popular for emotion detection than it is for sentiment analysis, and so despite important difficulties.

For instance, [Alm05] explored emotion detection in the domain of fairy tales with supervised machine learning (linear classification) but limited to the presence of an emotion and its valence. In the *Feeler* system, [Dan08] compared different classifiers that were trained on three resources: the ISEAR dataset, WordNet-Affect and the WPARD dataset which is a scalar polarity lexicon [Med05]. The authors show that on the six emotions of the Semeval evaluation task [Str07] (see next section), Vector Space Model, an unsupervised approach, obtains better results than Naive Bayes or Support Vector Machine (SVM), and reaches 49.6% of F-measure for the best emotion (joy). In addition to Wordnet-Affect and Naive Bayes classifiers, Latent Semantic Analysis (LSA) is also tried in [Str08]. The results appears lower than in [Dan08] on the same test set, but confirm that unsupervised machine learning such as LSA seems more effective than supervised techniques. In [Kim10] a thorough evaluation of different unsupervised techniques was performed on the Fairy Tales dataset, the Semeval dataset and the ISEAR dataset. They found out that Non-negative Matrix Factorization outperforms LSA or PLSA (Probabilistic LSA) in most cases, but this result is not completely consistent across domains. Nevertheless, supervised machine learning is still widely used, as shows the track 2 of the Medical NLP Challenge [Pes12] which is an evaluation on emotion detection in suicide notes: among the 13 participants whose system description is provided, 8 used SVM alone or in combination with other supervised learning algorithms.

3.2.1.6 Evaluation of Sentiment Analysis Techniques

While the Medical NLP Challenge is focused on rather specific emotions that are found in suicide notes (abuse, anger, blame, fear, guilt, hoplessness, sorrow, etc.), there is another evaluation called the SemEval Affective Text task [Str07] which targets the general emotions of the Ekman's model. It consists of evaluating systems on 1,000 news headlines annotated both with Ekman's emotions and scalar valence. This evaluation provides a perspicuous setting that enables to compare different kinds of approaches on both sentiment analysis and emotion detection. For SemEval, six systems competed (five to the valence task and three to the emotion task): UPAR7 [Cha07] which is a rule-based system combining syntactic analysis and polarity lexicons (SentiWordNet and WordNet-Affect), SICS [Sah07] which is based on words space, CLaC [And07] which is an hybrid system including generated lexicon, manually designed valence shifting rules and syntactic analysis, CLaC-NB [And07] which is a classical NaiveBayes classifier, UA [Koz07] which uses Pointwise Mutual Information in a way similar to [Tur02], and SWAT [Kat07] which is a supervised system that includes synonyms expansion. The best results were obtained by the CLaC and CLaC-NB systems on the valence task, while the results on the emotion detection task were quite low for all systems. Note that these results were outperformed in [Dan08] and then in [Kim10] with unsupervised systems that did not directly competed to the challenge.

More specifically for sentiment analysis and opinion mining, there are other evaluation tasks such as the TREC campaigns which evaluate the ability of systems to detect opinionated blog posts and their polarity [Oun08] and the NTCIR6 tasks which primarily target Asian languages but also includes an English track [Eva07]. Nevertheless some other corpora emerged as *de facto* standards for comparing results, this includes the MPQA Opinion corpus [Wil05] and the Polarity dataset [Pan04] which are often used for sentiment analysis evaluation purposes.

3.2.1.7 MultiLingual Sentiment Analysis

All the research works reviewed so far cover the sentiment analysis task for English, machine learning algorithms are trained on English corpora and lexicons are in English only. Most work in multilingual or cross-lingual sentiment analysis focus on subjectivity analysis which consists in determining whether a document or sentence is subjective or objective, but we assume that these approaches may be applied to polarity analysis as well. When performing sentiment analysis in another language several strategies are possible [Ban11], sorted here from the best to the worst cases. The ideal case is when an annotated corpus is available in the target language, or if it is possible to build such corpus automatically, then classical supervised learning methods are possible. If not, automated translation can be used: from an annotated English corpus to the target language and then use the translated corpus to train a classifier [Ban08], or vice versa by translating the target corpus into English and running an existing classifier for English [Kim06; Den08], this operation can also be performed with manually aligned corpora [Mih07]. Another approach is to build directly the lexicon for the target language, by lexicon-crawling and bootstrapping techniques. The last strategy is to translate instead the English lexicon to the target language and use the translated lexicon [Kim06; Mih07], however this method does not seem to yield the best accuracy. We refer the interested reader to [Ban10; Ban11; Liu12] for a good overview of the topic.

3.2.1.8 Conclusion

The domain of sentiment analysis and emotion detection from text is relatively recent, but despite of its youth it has spawn a rich and important literature which is reviewed in major surveys [Pan08; Liu12. Cal10]. Most of the work is related to machine learning in which finding good data, good features and good algorithms are three important issues. Work around polarity/emotion lexicons is also an important research trend. However, as [Liu12] states, machine learning methods may have been over-exploited and research should focus now on the how and why a sentiment can be conveyed by a natural language text. Nevertheless, the task is very hard as there are many phenomena at hand in sentiment analysis, including syntactic, semantic and pragmatic phenomena. There is thus a lot of room left for further improvement.

3.3 Emotion analysis from interaction devices (keyboard, mouse, focus of attention on screen....).

Human computer interaction is also studied as a source for emotions: Although significant progress has been made in the development of natural user interfaces based on speech and gestures, interaction between humans and computers predominantly occurs through keyboard and mouse. The data stream generated by the user pressing keys, moving the mouse and clicking mouse buttons contains valuable information about his behavior: it tells us how the user navigates through a software application or web site, what information is entered, or which items are selected. This way, researchers and practitioners in the field of human-computer interaction can evaluate the usability of a system. Keyboard and mouse data also allow us to assess the user's physical activity and workload. These can be used to detect activity levels that can lead to RSI (Repetitive Strain Injury). Recent research has shown that keyboard and mouse data can also be used to detect certain mental states, such as confusion, derived from certain patterns of mouse movement. There are tools that capture user behavior through keyboard and mouse, often meant for remote testing of websites. The type of data collected is usually rather coarse and insufficient for scientific research on behavior and emotions, Tools like uLog are needed, which record mouse clicks, scrolls, keystrokes, text strings, resizing of windows, pop-ups, menus selected, etc. and outputs time-stamped events for offline analysis.

By analogy, also interaction of users with other computing devices can be conceived as an input for understanding emotions of its users. In this project we will focus on this topic and study to which extent interaction of users with their TV and so called second screen devices (smartphone, tablet, laptop) can be exploited in order to detect emotions in a reliable way.

Van den Hoogen [Van08] investigated the use of multiple behavioral measures, body postures and pressure on input devices (mouse and keyboard), as indicators of users' gameplay experience, including frustration. A few more studies have demonstrated that pressure exerted on other input devices, such as a touchpad [Hel02] and buttons on a game console [Syk03], can also be used to

recognize user frustration. All of their findings suggested that the pressure applied on such input devices increases with user's frustrating experiences and difficulty levels in a game.
Most of the previously mentioned works examine the user state of being frustrated and determine the level of frustration based on subjective responses such as self-reported data or questionnaires. The users had to identify the critical incidents when they feel frustrated and rate their frustration on a five- or seven-point Likert-type scale. Based on this information, the behavioral measures were then analyzed to find a correlation with the self-reported measures with regard to frustration.
Octavia, Coninx and Luyten [Oct11] used a correlation between the finger pressure exerted on a 3D input device and the user frustration state, which was determined based on the user's physiological measures and used it to adapt the interaction. In the study some users' frustration decreased after adaptation while frustrated users were more effective during the adapted interaction.

The success or failure of applications and services is greatly determined by User Experience and thus we have to achieve better user experience by applying affective computing technologies to understand and respond to user intentions and emotions. One promising but still far from mature technology in affective computing is the keystroke dynamics. As stated in [Viz09], there may be a possibility of detecting cognitive and physical stress by monitoring keyboard interactions with the eventual goal of detecting acute or gradual changes in cognitive and physical function.

Affective computing is an emerging area of research and practice broadly defined as "computing that relates to, arises from, or deliberately influences emotion" [Pic97].

The authors of [Epp11] have a well-defined depiction of the problems in detecting emotions:

"Many approaches for detecting user emotions have been investigated, including voice intonation analysis, facial expression analysis, physiological sensors attached to the skin, and thermal imaging of the face. Although these explorations have seen varying rates of success, they still exhibit one or both of two main problems preventing wide scale use: they can be intrusive to the user, and can require specialized equipment that is expensive and not found in typical home or office environments.

Key stroke dynamics detects users' emotional states through their typing rhythms on the common computer keyboard. This is an approach from user authentication research that shows promise for emotion detection in human-computer interaction (HCI). Identifying emotional state through keystroke dynamics addresses the problems of previous methods by using standard equipment that it is also non-intrusive to the user [Epp11]."

Keystroke dynamics has been studied widely in the biometric authentication (cf. [Ban12]). However, already in the late 90's, according to [Epp11], there is "anecdotal evidence" which suggests that strong emotional states can interfere with authentication [Mon00]. This idea is probably the cornerstone of the whole keystroke dynamics based affective computing.

There has been a lot of research in this field of biometrics but no commercial products still exist. In a literature review on industrial performance monitoring in [Mai11] the authors focus on emotion identification at a single moment in time, but note that the probable usage of a biometric system, such as keystroke dynamics, would be for identifying changes in emotions over time. Commercial products that employ biometric data appear to be in the early stages of development and are utilized mostly for identification and authentication purposes.

The authors of [Mai11] then continue analysing two different publications of keystroke dynamics – based affective computing. They state that especially in [Viz09] the proposed approach is attractive because it requires no additional hardware, is unobtrusive, is adaptable to individual users, and is of very low cost. As mentioned above, no available commercial products were found that have a similar functionality, since most of the products that employ keystrokes are designed for security purposes (e.g. for user authentication).

Some Examples of Keystroke Dynamics Research:

In [Epp11] the authors determined user emotion by analysing the rhythm of their typing patterns on a standard keyboard. They then conducted a field study where they collected participants' keystrokes and their emotional states via self-reports. From this data they extracted keystroke features, and created classifiers for 15 emotional states. Top results include 2-level classifiers for confidence, hesitance, nervousness, relaxation, sadness, and tiredness with accuracies ranging from 77 to 88%. In addition, they show promise for anger and excitement, with accuracies of 84%.

In [Alh11] the author investigates the effectiveness of diagnosing users' affect through their typing behaviour in an educational context. In a field study subjects used a dialogue-based tutoring system. Eighteen dialogue features associated with subjective and objective ratings for users' emotions were collected.

Several classification techniques were assessed in diagnosing users' affect. An artificial neural network approach was chosen as it yielded the highest accuracy. To lower the error rate, a hierarchical classification was implemented to first classify user emotions based on their valence (positive or negative) and then perform a finer classification step to determining which emotions the user experienced (delighted, neutral, confused, bored, and frustrated). The hierarchical classifier was successfully able to diagnose users' emotional valence, while it was moderately able to classify users' emotional states.

Authors in [Kha10] present an attempt to recognize selected emotion categories from keyboard stroke pattern. The emotional categories considered for our analysis are neutral, positive and negative. Various classifiers are used, like Simple Logistics, SMO, Multilayer Perceptron, Random Tree, J48 and BF Tree, which is a part of WEKA tool, to analyse the selected features from keyboard stroke pattern.

These results indicate that negative emotional states are more visible and recognized better using the keyboard modalities.

Some Examples of Keystroke Dynamics Together with Other Features

In [Viz5] the research explored the relationship between exposure to stress and changes in keystroke and linguistic features. For each condition and data set there are changes in keystroke features as indicated by use of the backspace, delete, end, and arrow keys. When individual differences are addressed by normalizing the data, changes in the time per keystroke (i.e., a timing feature) and lexical diversity (i.e., a content feature) for both the cognitive and physical stress conditions as well as a change in pause length (i.e., another timing feature) in the physical stress condition become evident.

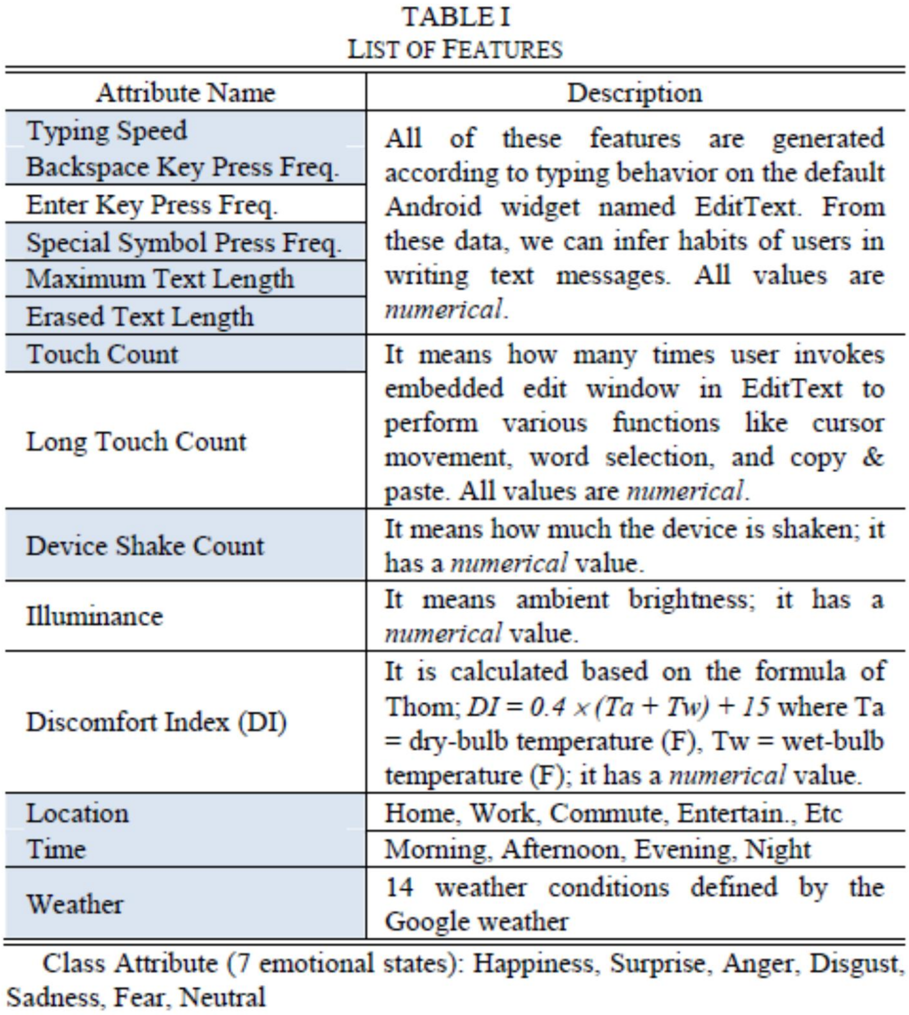


TABLE I
LIST OF FEATURES

Attribute Name	Description
Typing Speed	All of these features are generated according to typing behavior on the default Android widget named EditText. From these data, we can infer habits of users in writing text messages. All values are *numerical*.
Backspace Key Press Freq.	
Enter Key Press Freq.	
Special Symbol Press Freq.	
Maximum Text Length	
Erased Text Length	
Touch Count	It means how many times user invokes embedded edit window in EditText to perform various functions like cursor movement, word selection, and copy & paste. All values are *numerical*.
Long Touch Count	
Device Shake Count	It means how much the device is shaken; it has a *numerical* value.
Illuminance	It means ambient brightness; it has a *numerical* value.
Discomfort Index (DI)	It is calculated based on the formula of Thom; $DI = 0.4 \times (Ta + Tw) + 15$ where Ta = dry-bulb temperature (F), Tw = wet-bulb temperature (F); it has a *numerical* value.
Location	Home, Work, Commute, Entertain., Etc
Time	Morning, Afternoon, Evening, Night
Weather	14 weather conditions defined by the Google weather

Class Attribute (7 emotional states): Happiness, Surprise, Anger, Disgust, Sadness, Fear, Neutral

In [Lee12] an unobtrusive emotion recognition approach for affective social communication on mobile devices is proposed. With the development of an Android application named affective twitter client (AT Client), various real-world data and information was gathered from the Android smartphone equipped with diverse built-in sensors. By analysing these dataset, 10 features related to the emotional state of the human user were discovered (TABLE I above); these features are mainly divided into user behavioural patterns (e.g., typing speed) and the user context (e.g., location). With the training dataset including selected 10 features, Bayesian Network classifier was built and it showed classification accuracy of 67.52% on average for 7 emotional states: happiness, surprise, anger, disgust, sadness, fear, and neutral.

Other examples of affective computer in HCI with mouse, motion and gestures are introduces below.
In [Kak09] the authors have constructed a Web-based Biometric Mouse Intelligent System for Student's Emotional and Examination Process Analysis (BMIS). BMIS is an online system and allows for more physiological, psychological and behavioural data to be generated from a larger pool of students for further analysis and research. Data is accumulated in individual student modules based on the student's mouse movements, palm state and e-self-reports. The system extracts physiological and motor-behavioural parameters from mouse actions and hand characteristics, and the user fills in the psychological (e-self-reports) data, which can be used to analyse correlations with user's emotional state and labour productivity.

The standard mouse of type RX-1000 (Logitech) was modified for detection of emotional state of users and students. Four sensors were embedded into the mouse. Humidity and temperature sensors were mounted in the rearward side of mouse to control the temperature of hand and sweat of hand respectively. One force sensor was mounted in the left sidelong part of mouse to measure the force acted by the thumb. The other force sensor was mounted in the left button of the mouse to measure the force acted by the finger to this button.

The research in [Cou12] considered motion and gestures. Motion and surface gestures are, respectively, performed in 3D with a mobile device and in 2D on the touchscreen of a mobile device, like a smartphone or a mediaplayer. Sensing these gestures is now widely done on many mobile devices. Thus, 3D motion and 2D surface gestures are ideal candidates for a cheap, discreet and mobile way of expressing affective parameters.

After gathering and analyzing in situ and in the moment samples of the subjects' gestures and emotions, we found that 249 descriptors of each collected gesture, 104 of them were significantly correlated to at least one of the affective dimensions reported by subjects.

In [Sch04] the authors introduce a novel approach to human emotion recognition based on manual computer interaction. The presented methods rely on conventional graphical input devices: Firstly a standard mouse as used on desktop PCs, and secondly the interaction on touch-screens or -pads as in public information terminals, palm-top devices or tablet PCs is considered. Additionally the gain of the integration of touch pressure information is evaluated. Four discrete emotional states are classified: irritation, annoyance, reflectiveness, and neutral affect for the use in initiative tutoring, error clarification, Internet customer personalization, and others.

Even without special hardware and independent of the underlying application the affect of a user could be recognized with up to 83.2% correct assignment providing more comfort for the user compared to invasive bio measurements and avoiding extra costs for sensors.

Limitations of Keystroke Dynamics

There are some problems with keystroke dynamics based affective computing. First, no commercial products are available even though the research has been going on for well over ten years. This is mentioned, e.g., in [Mai11]. Due to lack of commercial interest the research is academic and doesn't contribute (at least at the moment!) to new commercial applications.

Generally, behavioural biometrics, like keystroke dynamics used for authentication, has less discriminatory power than conventional biometrics like fingerprint or iris. Also, in the conventional biometrics the scientific basis lies on constant underlying identifier whereas the behavioural biometrics is based on assumptions or statistical observations [Sch12].

In [Ban12] the authors list some problems of authentication with keyboard dynamics. The emotional state of the user has an impact on the typing speed. A negative state led to a 70% reduction in typing speed compared to the 83% increase in the typing speed when they are in a positive state. The effect of emotions on typing was affirmed by Epp et al. [Epp11]. Similarly, health conditions, place where a person types such as on the bed, on the table, the type and brand of computer used could also have an impact on the efficiency with which a person can be accurately classified.

This gives a clue to problems with the use of keystroke dynamics in affective computing. In [Viz09] authors note that their keystroke dynamics based solutions should be tested with varying typing abilities and keyboards, with varying physical and cognitive abilities, and in real-world stressful situations.

A layman in affective computing might think that if your hands are clumsier than normally, e.g. after even minor weightlifting, the keystroke dynamics will probably differ a lot from the normal. But is it classified as sadness, distress or clumsiness due to weightlifting?

Conclusions

Keystroke dynamics is an interesting field of affective computing. Its strength lies in the fact that no specialized hardware is needed. Since no commercial applications are available, the research in keystroke dynamics tends to be academic by nature. However, there may be a possibility of detecting cognitive and physical stress, or even gradual cognitive decline due to ageing, by monitoring keyboard interactions. In [Bal11] it is even suggested that typing sequences can be used in order to identify individuals with the early stages of dementia or Parkinson Disease. This alone makes the research of keystroke dynamics a worthwhile effort.

3.4 User activity detection/analysis

3.4.1 Situation and context awareness

The urge to understand more about the user and his/her context started with the ubiquitous computing vision that placed abundant technology in user's everyday environment [Weiser]. Context awareness was first defined by Shilit and further developed by Dey and others [Sch94] [Dey01]. Initial definitions focussed on the situation of an entity, where the user's context (or the context of the devices carried by the user) was mostly the focus of the attention. Context was understood to include location, identity, activity and time, but more recently the viewpoint has shifted towards a notion of the user being part of a process or ecology, as exemplified by Ambient Intelligence (AmI). AmI refers to a vision in which devices interact to support people in carrying out their everyday life activities and tasks in a natural way using information and intelligence that is hidden in this network of interconnected devices.

3.4.2 Location and positioning techniques

Location is by far the most used type of context. Location based services have been developed for outdoor use as GPS becomes common place in mobile devices. Indoor location is still a challenging subject, as no solution has yet become mainstream. Nevertheless, various indoor positioning solutions have been developed (e.g. based on WLAN positioning, or more coarse Bluetooth beacons) and can be applied for specific purposes (factory positioning, location based marketing). See also the sections below. As a source for user intention, location tracking provides valuable clues about the social situation (work, leisure, home) and goals (travelling to known destinations) of the user. Applications can thus advance from providing information relevant to a location, to providing information relevant to the task at hand (e.g. providing time tables of the next trains leaving from this station to home after work).

3.4.3 Positioning techniques

Many empathic applications within the smart house environment include context awareness as an essential requirement. The following techniques and technologies to support positioning in AAL applications are explained:

Indoor positioning techniques [Sinan Gezici, "*A Survey on Wireless Position Estimation*", Wireless Personal Communications, Vol. 44, Issue 3, pp. 263-282, Feb. 2008]:

- In detection by proximity, short range sensors are used to detect the passage or the position of an element close to them. The objective of these techniques is to determine spatial coordinates of the object so they can be used as a complement to refine the positioning estimated by higher-range techniques that provide less accuracy.
- In Received Signal Strength Intensity (RSSI) techniques, distances are calculated according to the strength of the signals between the fixed receiver nodes and the mobile object or person to locate. These techniques represent a good solution in terms of cost and energy consumption due to the simplicity of the system hardware. The main disadvantage of these systems is the fluctuation of the received signal strength caused by the multipath effect so the system must be designed to compensate this issue by means of distributed algorithms and other software solutions such as self-learning mechanisms to minimize miscalculations.
- In Angle of Arrival (AoA) techniques, several fixed receiver nodes are used to determine the direction of arrival of the signal transmitted by the mobile object. These system provide a very accurate object position estimation but they also present the following difficulties: (1) a Line of Sight (LoS) is required between the mobile object and the antennas; (2) the system must be designed taking into account the multipath effect over signals; (3) peculiarities on the antenna radiation diagram and possible dark zones can complicate the measurement as well as very small changes in the orientation of the antennas; (4) specific receptors are needed for the fixed nodes that require highly directional and very complex antennas. Due to the high cost of the required receptors, these systems are not used in the market. This technique is

particularly appropriate for using with UWB communications given the immunity that this technology presents to multipath effect.

- In Time of Arrival (ToA) techniques, distances between the mobile node and at least three fixed nodes are calculated based on the time of flight of the radiofrequency signal at a known speed (according to the characteristics of the medium) and a circular trilateration process is applied to the estimated distances. These system provide a very accurate object position estimation but they also present the following difficulties: (1) possible calculation errors might arise because of the required signal processing time that depends on the manufacturer and also because of the battery level of the device; (2) absolute synchronization is needed between all nodes (mobile and fixed) what increases the cost of the system; (3) the loose of LOS can cause an error that drives to miscalculations.
- In Time Difference of Arrival (TDoA) techniques, the difference between signal arrival times to different pairs of fixed nodes is used to calculate the mobile node position. As in ToA systems, synchronization between fixed and mobile nodes is essential and its lack implies accuracy errors. The main advantage of this technique is that the loose of LoS doesn't affect as much as in ToA because the difference of times cancels potential errors caused by the reflections.

Low/medium range wireless positioning technologies:

- Radiofrequency:
  - Bluetooth [Bluetooth SIG, http://www.bluetooth.org/apps/content/] operates in the 2.4GHz Industrial, Scientific and Medical (ISM) band. Its bitrate is 1 Mbps and the range is typically 10-15 m. On the other hand, Bluetooth is a “light” standard, highly ubiquitous (embedded in most phones, personal digital assistants (PDAs), etc.) and supports several other networking services in addition to IP. Bluetooth tags are small size transceivers. As any other Bluetooth device, each tag has a unique ID. This ID can be used for locating the Bluetooth tag. The Bluetooth SIG has recently provided the Bluetooth 4.0 core specification that includes classic Bluetooth, high-speed Bluetooth and low-consumption Bluetooth.
  - WiFi [IEEE 802.11 Working Group, www.ieee802.org/11] operates in the ISM band and has become very popular in public hotspots and enterprise locations during the last few years. With a typical bitrate of 11, 54, or 108 Mbps and a range of 50-100 m, WiFi is currently the dominant local wireless networking standard. Thus, it is usually possible to use an existing infrastructure for indoor location by adding a location server.
  - UWB [WiMedia Alliance, www.wimedia.org/en/index.asp] is a technology able to provide the necessary characteristics for applications with most critical requirements in terms of accuracy and update interval. It is based on sending ultra-short pulses (typically <1 ns), with a low duty cycle (typically 1:1000) thus using a spread signal (even >500 MHz wide) in the spectral domain. It provides a very accurate spatial positioning (a few centimeters) when used in ToA and TDoA techniques [S. Gezici, H.V. Poor, “*Position estimation via ultra-wide-band signals*”, Proc. IEEE, vol. 97, no. 2, pp. 386–403, Feb. 2009] as it offers certain inmunity to multipath signals because the short duration of pulses facilitates the differentiation between the direct signal and the reflected ones.
  - RFID [ETSI RIFD, http://www.etsi.eu/website/Technologies/RFID.aspx] is a means of storing and retrieving data through electromagnetic transmission to an RF compatible integrated circuit and is now being seen as a means of enhancing data handling processes. An RFID system has several basic components, including a number of RFID readers, RFID tags, and the communication between them. The RFID reader is able to read the data emitted from RFID tags. RFID readers and tags use a defined RF and protocol to transmit and receive data. RFID tags are categorized as either passive or active. Passive RFID tags operate without a battery. They are mainly used to replace the traditional barcode technology and are much lighter, smaller in volume, and less expensive than active tags. They reflect the RF signal transmitted to them from a reader and add information by modulating the reflected signal. However, their reading ranges are very limited, typically 1–2 m, and the cost of the readers is relatively high. On the other hand, active RFID tags are small transceivers, which can actively transmit their ID (or other additional data). The advantages of active RFID are with the smaller antenna and in the much longer range (can be tens of meters). Active tags are ideally suited for the identification of high-unit-value products moving through a harsh assembly process.
  - The IEEE 802.15.4 physical (PHY) layer and medium access control (MAC) layer standard for low-rate wireless personal-area networks (LR-WPANs), and the ZigBee networking and application layer standard [ZigBee Alliance – ZigBee Specification,

http://www.zigbee.org/Specifications/ZigBee/Overview.aspx] allow for localization information to be measured between pairs of sensors. In particular, RSS can be measured in the 802.15.4 PHY standard via the link quality indication (LQI), which reports the signal strength associated with a received packet to higher layers. Most of the sensor-network-based location estimations use RSSI techniques [N. Patwari, A. O. Hero, M. Perkins, N. S. Correal, and R. J. O'Dea, "*Relative location estimation in wireless sensor networks*", IEEE Transactions on Signal Processing, vol. 51, no. 8, pp. 2137-2148, Aug. 2003] [J. N. Ash and L. C. Potter, "*Sensor network localization via received signal strength measurements with directional antennas*", in Proceedings of Annual Allerton Conference on Communications, Control and Computing, Sep. 2004, pp. 1861-1870].

- Infrared radiation is used in wireless personal area network (WPAN) since it is a short-range narrow-transmission-angle beam suitable for aiming and selective reception of signals. Most of the Infrared Data Association [IrDA, http://www.irda.org/] wireless system is based on the LoS mode. Considering the high room accuracy of the IR location [R.Want, A. Hopper, V. Falcao, and J. Gibbons, "*The active badge location system*", ACM Transactions on Information and System Security (TISSEC), pp. 91–102, Jan. 1992] and the high availability of the RF location, it makes sense to combine the two technologies into a hybrid location system that combines the advantages of each technology, this is, the accurate room location of IR and the wide range of RF; however, it could be more costly.
- Ultrasound-based systems can provide a very accurate positioning. The advantage of these systems is with the low transmission speed of acoustic waves that reduces the required spatial accuracy and temporal granularity needed to estimate distance with a high precission (a few centimeters) [E. Mangas, E. Bilas, "*FLASH: fine-grained localization in wireless sensor networks using acoustic sound transmissions and high precision clock synchronization*", in Proceedings of the 29th IEEE International Conference on Distributed Computing Systems, pp. 289–98, Montreal, 2009]. However, it is necessary to consider other factors that have a strong influence in these systems as the changes of propagation speed depending on the atmospheric conditions, the non-LoS or the reflections in walls and floor [J.H. Park, W.S. Jang, J.M. Lee, "*Localization of multiple robots in a wide workspace*", in Proceedings of the 15th International Symposium on Artificial Life and Robotics (AROB 2010), Oita, Japan, 2010].

To improve the reliability of the positioning system, different sources of location information can be combined, such as the strength indicator of the messages that the sensors send to the processor devices through a wireless communication interface, RFID tags lectures or infrared detection lectures, and measurements from different body position sensors (accelerometer…).

3.4.4 Activity and behaviour recognition

Another branch of context awareness research has concentrated on *activity analysis*. Most prominently this has been done for mobile devices and for smart environments. Activities of the mobile user can be derived by combining sensory information from the mobile phone with general knowledge about the user. Thus motion sensors give clues about the type of motion of the user (at rest, walking, running, walking steps, driving a car, etc) and the situation (indoors, outdoors, noisy or light/dark environment, in a meeting with (known) people, etc.). This research relies on sensor data processing, merging, classification and reasoning. Research in smart environments is often related to ambient assisted living and attempts to derive information about people's wellbeing (sleeping, awake, daily rhythm, falls, level of general activity) from the various sensors – often including camera's – in a smart environment. The applied technologies are similar, but tend to be computationally more demanding because of the variety of sensors and the computational power available. Knowledge about the activity of a user is a good starting point to detect behavioural patterns and help to assess the stage of the process the user is in. It therefore provides valuable input for intention detection.

Sensors to monitor user activity can provide information used to derive statistics about the user habits (e.g. sleeping and awake hours, level of general activity…) that helps to assess his circumstances thereby providing valuable input for intention detection. In this sense, the following sensor technologies can be used to estimate the user activity:

- Different types of pressure sensors (capacitive, electromagnetic, piezoelectric, optical, potentiometric…) to detect presence in bed and chairs and gates opening and closing [E.M. Tapia, S.S. Intille, K. Larson, "*Activity Recognition in the Home using Simple and Ubiquitous Sensors*", In Pervasive Computing, vol. 3001, pp. 158-175, 2004].

- Inertial sensors (such as accelerometers, gyroscopes and magnetometers) that provide information regarding body position [A.M. Sabatini, C. Martelloni, S. Scapellato, F. Cavallo, "*Assessment of walking features from foot inertial sensing*", in IEEE Transactions on Biomedical Engineering 2005, vol. 52, pp. 486–494][L. Qiang, J.A Stankovic, M.A. Hanson, A.T. Barth, J. Lach, Z. Gang "*Accurate Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information*", in the 6th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2009), pp. 138 - 143, June 2009].
- Sensors to measure different parameters such as temperature, humidity, gas concentration or noise that provide valuable information regarding the user environment [V.R. Jakkula and D.J. Cook, "*Detecting Anomalous Sensor Events in Smart Home Data for Enhancing the Living Experience*", in Artificial Intelligence and Smarter Living, vol. 07, 2011].
- Biomedical parameters (studied in previous sections for estimation of user emotional state) can also be employed as a source of user activity information (i.e. a high heartbeat can imply a user making exercise)

Several methods to estimate the user activity by analyzing data from different information sources (this is, from different types of sensors) have been developed that will be studied in section 5.2.

3.4.5 Network context

Empathic applications will interact with the network connectivity in two ways. They will of course require a certain type of network connection in terms of bandwidth, security, reliability, security, price,... But they can also benefits from triggers and events sent from the network layer.

Additional inputs from the environment:

- Dynamic HotSpot detection: from the list of the available hotspots polled at regular intervals, it is possible to deduce valuable information, like the physical location of the mobile device (usefull indoors where the GPS signal isn't available). It also allows to know if the user is moving or not and this information can be used by the Empathic application to decide, for example, to use a bigger font size. If the location has already been seen by the device, the history of the available hotspots can also be used to improve the roaming functionality by getting ready to connect to the neighborhood spots.

- Use GPS sensor to deduce possible connections: If we know (from previous experience or cartography) which hotspot are available along with their respective properties, an Empathic application with a specific need (for example bandwidth of a certain quality) may suggest the user to move to the nearest place with the desired connection.

Multipath-routing and adaptative VPN:

Traditionally, the applications must do with the current networking capacity but there's no unified way for applications to request a certain type of connectivity, there's also no unified way for the network layer to notify applications when some preset conditions are met or not.

In order to help Empathic applications to get the best features we need an API (along with the library that implements it) that will allow bi-directional communication between the network layer and the applications.

In particular an Empathic application should be able to request:

- to have a certain amount of available bandwidth,
- to be certain that the connection with another point has a certain level of security,
- to be able to securely connect together all the members of a group,
- to know the price of the network link currently used, And on the other hand and application must also be able to be notified when some event occurs:
  - if the bandwidth capacity goes under a certain threshold,
  - if the connection is likely to be cut very soon,
  - if it's possible to load-balance the bandwidth between two connection points,

When using application on mobile devices it's often desirable to use wifi access when available (often better bandwidth and free compared to 3G), but the downside is that wifi is less reliable and if the user is moving, then it's hard to have a reliable connection.

Traditionally, the network parameters are set by the administrator and the end-user applications have no way to modify them. Allowing the applications to be notified by the network layer or allowing them to require some types of connectivity will allow Empathic application to improve their behavior thanks to networking configuration and notifications.

While there are numerous VPN and connectivity solutions out there, all the API will have to be developped along with the library that implements it and communicate with the network layer of the target operating system.

When a secure connection is required, the OpenVPN software will be used as it's a very software that totally covers the needs of secure connections; however it will have to be modified/improved in order to have it send notifications when necessary and to modify it's configuration on the fly.

Technogical starting point: OpenVPN (http://openvpn.net/index.php/open-source/)

- OpenVPN's licence (GPLv2) allows the re-use and modification of it.
- It's been a stable software for a long time, both in terms of reliability and in terms of security.
- OpenVPN code is well commented and documented,
- OPenVPN is very portable and already runs on the following platforms: IOS (IPhone), Android, Windows Mobile, Windows 7/8, Linux and OS X.

4. Data fusion and interaction modelling for affective computing

4.1 Indexing and classification techniques

4.1.1 Static classification methods

The static classification methods only consider the current frame of a video to classify it according to a given emotion model.

In [Coh03], Cohen et al. use Naive Bayes classifiers, Tree-augmented Naive Bayes classifiers, After the interactive selection of the face characteristics on the first image, the system follows the face movement by using the system developed by Tao and Huang called “Piecewise Bézier Volume Deformation” (P.B.V.D.) [Tao98]. As a result, the velocities of the selected points can be computed through the video as motion units. These units are used as features for classifications into the usual emotion categories with the help of a naive Bayes classifier. The underlying assumption of feature independency is however unrealistic. This drawback is partially overcome with the Tree-Augmented Naive Bayes classification. Here, dependencies are created between the features based on the mutual information brought by a variable given the emotion class to another. It results in a tree of maximal total weight linking the correlated features (variables).

Other static classification methods use Support Vector Machine (S.V.M.), mostly of type “One-versus-All” with gaussian kernels, as in [Nic11]. Here, multiple classification methods are considered for comparisons. Let us note that authors also tackl the multi-modality issue: three different channels are considered, including face, shoulders, and sound. As a result, the classification issue is crucial, because it will depend on the way the different modalities are “joined”. Before studying the “mixture” of modalities, the article suggests results for each channel, with SVM classification.

In the same way, let us also cite the work presented in [Gun07], where H. Gunes and M. Piccardi use the variations of centroid, rotation, width, height and area for the hands, head and shoulders' regions as inputs of SVM to classify the images into 12 emotional categories.
Another model for static classification is the Gaussian Mixture model, used a lot for speaker recognition in linguistic as indicated by C. Clavel and G. Richard in [Cla10]. For each emotion class, the observation is represented as a weighted sum of Gaussian densities, whose means, covariance matrices, and respective weight are parameters that need to be determined. The estimation is made with the Hope Maximization algorithm and the classification with the Maximum a Posteriori method.

4.1.2 Dynamic classification methods

In [Bla95], Black and Yacoob use a rule-based classifier. They use affine and planar parameterized models to recognize non-rigid movements of the face. After having placed windows on relevant zones of a initial, reference image, they estimate the value of the parameters of these models for each frame throughout the video. In this manner, they determine the motion of the face element included in each window (mouth, eyebrows, eyes, face). The motions of such elements are associated with the beginning, apex or end of emotion categories according to a set of displacement rules. Such rules are also used in [Yac96].

In [Ess98], I.A. Essa et A. Pentland (1998) do not provide any classification method, but underline possible features that can be used as inputs. They propose the use of two parallel models for facial actions:

- a physical based-model, consisting of the parametrized representation of groups of independent muscles responsible for facial actions,
- a geometric model representing skin tissues, resulting in a triangular mesh rule by the rigid and non-rigid dynamics equation.
-

The two models are linked by the attachment of the muscles on the vertices of the mesh surfaces (triangles).

At the beginning, the relevant facial zones (eyes, nose, lips) are initialized thanks to the eigenspace method introduced by Pentland and Moghaddam in [Pen94]. The optical flow is computed and a Kalman filter allows to take into account the previous images. The mapping function and the velocity field computed permit to determine the vertices' velocity. Essa and Pentland suggest the use of motion energy, of muscular activation and of vertices' velocity field as input features for emotion recognition.

In [Ots97], Otsuka and Ohya compute the optical flow from frame to frame and use the velocity vectors in the mouth and eyes zone to compute the low frequencies coefficients of their two dimensional Fourier Transform. These coefficients are used as inputs for Hidden Markov Models (HMM) with continuous outputs (only few parameters to estimate) represented as gaussian mixtures. It is a way to take into account the variability of facial expressions.

In [Ots96], authors use the same classification method but the inputs features built on the horizontal and vertical Haar Transform low frequencies coefficients.

In [Oli00], simple HMM are also used. Auhtors delelop the notion of “blob” with which they separate the face from the background of the image. The face blobs are represented by a Gaussian mixture with 2 or 3 components whose parameters are learned. From one image to another, 2 Kalman filters are used to control the centroid position and the dimensions of the rectangle where each blob has to be confined; thanks to that, the camera can be parametrized to put the face in the center of the image, with a constant size, without rotation. Then the mouth is extracted from the face as the face had been extracted from the whole image, but with a gaussian mixture comprising up to 5 components. The mouths area relatively to the face area, the width and length of the mouth relatively to those of the entire face, the vertices of the rectangle confining the mouth are used as input characteristics for the HMM. Particular configurations of the mouth (default position, smile, sadness, mouth opened, mouth widely opened) can be recognised and provide high level characteristics for rule-based classification.

In [Nef99], Nefian and Hayes propose a system for face recognition. Low frequencies coefficients of the two dimension Fourier Transform are extracted from rectangle segmenting each facial image, to build a feature vector. Nefian and Hayes use integrated HMM for the classification. A first level of “super-states” corresponds to a vertical segmentation of face between the forehead, the eyes, the nose, the mouth and the chin. Inside each “super-states”, there are “states” corresponding to the learning individuals. As a result, the machine learning is made with a double Viterbi algorithm. Output probabilities are represented as gaussian mixtures.

[Coh03] has already been introduced in the previous section, for the static classification methods tested. Here, we focus on the dynamic: the multi-level HMM.

There is one HMM by emotion category, and the probability of observation (features) given the class is represented by a continuous gaussian distribution. Then, the states' sequences of each categorical HMM are decoded and used as input for the “High-level” HMM which will give the final emotion.

In [Nic11], 3 channels are considered: the face, the shoulder and the audio cues.
The learning is made on the S.A.L. Database [Dou07] where persons react to avatars' solicitations. Coders annotated the videos according to the 2D Valence-Arousal emotion space to provide a ground truth measure, to define images corresponding to emotion ending and beginning of another one, and then segment the data.
After having detected the face with Haar classifiers, segmented it into Regions of Interest [eyes, mouth], extracted characteristics based on grey level and Gabor filters' processing, and selected features with GentleBoost, we have 20 points for the face. With a similar method, 5 points are chosen to describe the shoulders. Eventually 15 acoustic indices are used as audio cues.

The multi-modality is managed as follows.

- The first classification method tested uses SVM and has been introduced in the previous part. SVM are used on each modality and the classification results are compared to those of another method implementing Bidirectional Long Short-Term Memory Neural Networks (B.L.S.T.M.-N.N. which have already been used for affect prediction from audio signals as explained in [Wol08]).
- The second method predicts the emotion by Feature-level fusion and also uses BLSTM-NN. One feature vector gathers all the audio-visual cues.
- The third method uses Model-level fusion to take into account the fact that the audio and image channels are not independent.
- Finally, the last method consists of a prediction by Output-associative fusion, which does not only considers that the audio and image channels are not independent, but also that valence and arousal emotional dimensions are not independent neither.

Nicolaou et al. [Nic11] define three different measures of emotion estimations:

- The mean square error between prediction and annotation (“ground truth”) which evaluates the estimation according to variance and bias.
- The correlation between prediction and annotation, which evaluates the capacity of the estimation to be structurally close to the data annotated.
- The Sign Agreement Metric (S.A.G.R.)

Other machine learning methods relating to Neural Networks are evoked in [Sch97].

4.1.3 Video orchestration example

Having attentional immersion used for video presentation “at distance” use cases (i.e. videoconferencing) imply to develop and implement specific reasoning mechanisms. Such mechanisms enable for instance to identify which of the video events happening is the most relevant [Lav09] to display. Or, it may help to implement elements of the Cognitive Load Theory [May01] in order to support a better knowledge transfer (for instance when narration and visual information are complementary and presented simultaneously).

Video conference system can be extended to enable video orchestration supporting some of these attentional immersion aspects.

To solve the problem of camera selection a lot of solutions and systems were proposed: for basic video conference systems, one solution is to select videos/cameras to display through a remote control. Another solution is to show all the participants of the meeting through pre-defined orchestration templates. But with such system, once the number is high, a lot of details will be lost. Both systems are not convenient and an automatic and a smart camera selection are required.

Video orchestration based on “audio events”, is a solution in this direction. Nevertheless, only the speaker will be shown all the time and useful information for video orchestration are missing (for instance gesture). Research has suggested that around 70% of all the meaning is derived from nonverbal behavior/communication [Eng06].

Likewise, Al-Hames [Alh06] proved that the audio information is not sufficient and visual features are essential. One early approach was developed in [Sum04] in which high level features were used in a rule-based model taking into account both audio and visual information. Then Al-Hames [Alh061] proposed a new approach similarly based on rules but low level features were used instead, such as global motions, skin blobs and acoustic features. In [Alh07] the orchestration was improved by using the same low features but a statistical model (HMMs[Rab89]) was used instead of a classic based-rule engine. HMMs have been also used by Hörner [Hor09], and for better selection he combined low and high level features as inputs of his model.

HMMs have become one of the most popular formalism in the domain of video event modeling [Lav09],[Yam92] ,[Sta95] ,[Ngu05], [Jou11] due to their diverse benefits (e.g. temporal aspect, probabilistic aspect, efficient algorithms, etc.).

Systems based on HMMs [Alh07] [Din06] solve the problem of dynamicity and intelligent selection and present many advantages but learning mechanism is not interactive; by default it is done by an expert and a basic user is not able to train and configure his own templates and orchestration models.

Globally, the existing solutions remain still static with no intelligence in the system to select the right view/template and no programmability/leaning mechanisms dedicated to the video orchestration based on the and user interactivities.

4.2 Multi-modal model.

In [Seb04], authors point out several issues about multi-modality, that need to be tackled.

- Does multi-modality provide something better than single channels?
- What are the relevant indices in voice and image, and for what situation?
- When he is talking with somebody, which type of information does the observer focuses on? Which importance is to be given to each channel?
- Do people use voice, face and gesture in a complementary way, or with redundancy?

In [Bal04], T. Balomenos et al. combine facial features and particular hand movements and use them with HMM to classify image into 6 emotions.
Let us also cite the study in [Gun07], where wuthors use the variations of centroid, rotation, width, height and area for the hands, head and shoulders' regions as inputs of SVM to classify the images into 12 emotional categories.

In [Kan08], Kanluan et al. propose the same kind of procedure as [Nic11], but with different features, based in Discrete Cosine Transform for visual cues. Contrary to Nicolaou et al., they fuse both mono-modal emotion estimations at a decision level by a weighted linear combination.

[Nic11] proposes a multi-modal emotion recognition system which is representative of all the questions raised by the multi-modality problematic and the classification architectures it implies, as we have seen it in the previous part.

Several conclusions can be drawn from [Nic11]:

- Globally, the recognition rate is better with multi-modality than for each modality taken apart.
- Arousal is better predicted by audio cues that by the visual ones. It is the contrary for the valence.
- The multi-modality seem to be more useful to predict the valence than it seem useful to predict arousal.

In [Gun11], such kinds of conclusion are brought for voice features:

- The pitch provides at lot of information about arousal, as intensity, high frequencies energy, the timber.
- High Power seems correlated with low frequencies and long voyels, but it has to be verified.
- High valence seem to be linked to a high speech flow, long voyels, a large frequency band.

Kleinsmith et al. have shown in [Kle05] that human observers discriminate body-postures with the following emotional dimensions: scaling, arousal, valence, action tendency.

It has to be considered that sometimes, the different modalities of expression can be congruent, non-congruent, or blended. It can help to study the different emotion mixture evoked in [Och05] in the future.

In [Hud12], E. Hudlicka and H. Gunes think about an intermediary level between the expressive features and the emotional analysis and quote the appraisal theory of K. Scherer [San05]. They wonder whether it is possible to characterize the information brought by a channel and then conclude about the conscious or unconscious use of this channel in the emotion production processing. Is this “channel call” different according to the stimuli, or in other words according to the appraisal type?

4.2.1 Multimodal Emotion Recognition - Feature Fusion

Significant research has been performed on emotion recognition using unimodal information during the last decade. Today, many of the studies for emotion recognition performed on unimodal features containing only text, speech, or visual information (e.g. face, body gestures). However, recent work targets to use multimodal information for emotion recognition. According to the famous 7% , 38% , & 55% rule of Mehrabian (Mehrabian, A., Comminication without words, Psychology Today 2, pp. 53-56, 1968.), emotions are conveyed by verbal, vocal and facial features, respectively. Therefore, emotions itself are multimodal in nature.

Multimodal Emotion Recognition (MER) is a collection of interdisciplinary methods for the recognition of human emotion. MER is a new research trend to investigate roles of multiple modalities for EER and obtain better results from the power of multimodality. However, multimodality introduces the fusion problem. When and how should the information combined by other resources? In literature, there exist two different approaches namely early and late fusion. The former merges features in low-level feature space whereas the latter uses semantic output of individual unimodal classifiers as inputs to a higher-level classifier.

We compared advantages and disadvantages of early and late fusion scheme.

Early vs. Late Fusion

Multimodal analysis and integration introduces feature fusion problem. Fusion is a need for multimodal systems where a decision is required based on outputs of two or more unimodal sources. Some researcherís use the terms early and late fusion

Gunes & Piccardi (Gunes, H. & Piccardi, M., Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications, 30:1334-1345, 2007) whereas others use data, feature and decision level fusion Dasarathy (Dasarathy, B., Sensor fusion potential exploitation-innovative architectures and illustrative applications. IEEE Proceedings 85(1) pp. 24-38, 1997), or semantic fusion Wu, Oviatt, & Cohen (Wu, L., Oviatt, S. L., & Cohen, P. R., Multimodal integration – a statistical view. IEEE Transactions on Multimedia, 1(4):334-341, 1999).

Early fusion is nothing different from simple machine learning process. Since the fused feature vector size obtained from multiple modalities is usually quite large, it introduces the feature dimensionality problem. On the other hand, scalability of late fusion scheme is better than early fusion scheme because that fused number of classes is much less than fused number of features vectors. Since late fusion is much like interpretation of high-level semantics, the recognition accuracy directly depends on the performance of individual modalities.

Common belief is to use the hierarchical levels for the fusion process where output of a level is used as input for the next level in the hierarchy. This hierarchy is constructed so that semantic complexity increases proportionally with the higher levels. If the different modalities does not complement or support each other then accessing to relevant information requires well-defined assign of feature and fusion weights.

In case of MER, multimodal studies in emotion recognition are different from other domains such that, emotions do not exist for some circumstances or different emotional labels can be assigned for the same temporal section of video for different modalities. Handling of these two problems for such cases is difficult in early fusion schemes where the fusion of feature vectors from different modalities is required and absence of the feature vectors affects the classification performance. In addition, early fusion of features containing different classes of emotions is not suitable for machine learning purposes as the machine learning algorithms designed for classifying similar concepts. On the other hand, absence of features is not a problem for late fusion scheme because that outputs of individual unimodal results are used. At this point, each modality can result different emotion labels for the same video segment.

5. Behaviour modelling techniques

5.1 Behaviour, desires and intention

5.1.1 Behaviour observation to achieve intention awareness

Intention aware products assess human actions and responses through measurement and interpretation of overt behaviour in terms of cognitive models. Behavioural measurement methods can be direct or indirect. Direct methods capture human behaviour that can be seen (e.g. body and head motion, gestures, facial expressions), heard (e.g. vocalizations, speech) or felt (e.g. grip force). Indirect methods derive behaviour from human-system interaction (e.g. key strokes, mouse clicks, car driver actions). While manual data collection by human observers has been the main way of working for many years, advances in sensor and computer technology have allowed many observational methods to be automated. However, many techniques for automated behavioural observation are still limited to highly artificial circumstances and significant research challenges remain in order to capture and interpret natural behaviour of freely moving subjects in ecologically valid settings.

5.1.2 Intention

Intention is a well studied subject in psychology and has found its way into IT through early work in the 1980’s on text analysis and help facilities in Unix environments [Sch78]. The subject has been addressed in context awareness research as well, but has received surprisingly little attention in literature. Various experiments, particularly for ambient assisted living environments, have been done and have demonstrated the possibilities, and the complexities, of intention recognition [Gir08] [Pei09]. The use of intention involves the understanding of user goals and tasks (in a certain domain), the recognition of user activity and situations, and reasoning on most likely intentions based on this understanding. Situation and activity recognition is part of context awareness, as described in the previous chapters. Goals and tasks have been used for user interface modelling; see for example the work by Paterno [Pat99]. Matching tasks to observed behaviour requires logic-based, case-based or probabilistic reasoning methods. The development of generic smart space architectures utilising semantic description of components and events, like done in e.g. the DIYSE project, will also help to put intention recognition in place. Providing a rudimentary understanding of a user’s intention does not need to be as complex as described in these systems. Instead for many applications the domain is fixed and the set of intentions can be narrowed down considerably. Some good examples of emerging applications utilising human intent are presented by Ted Selker (MIT) in a seminar talk [B].
Intention awareness is a hot topic of research also in the fields of psychology and cognitive neuroscience. The main focus of attention is the discovery of the fact, that conscious intention may have less control over our behaviour than previously thought. Our intent to perform a purposeful action seems not to be only determined by an act of volition, but is also driven by unconscious neural activity.

A first clue to this effect was the discovery of the readiness potential, which shows a measurable change of electrophysical potential of the scalp approximately 500 ms before the actual occurrence of a spontaneous action[Kor65] [Lib82]. The discussion on the voluntary nature of intention goes beyond our research, but serves as good background knowledge.

5.1.3 Beliefs, desires, intentions, plans and goals and their recognition

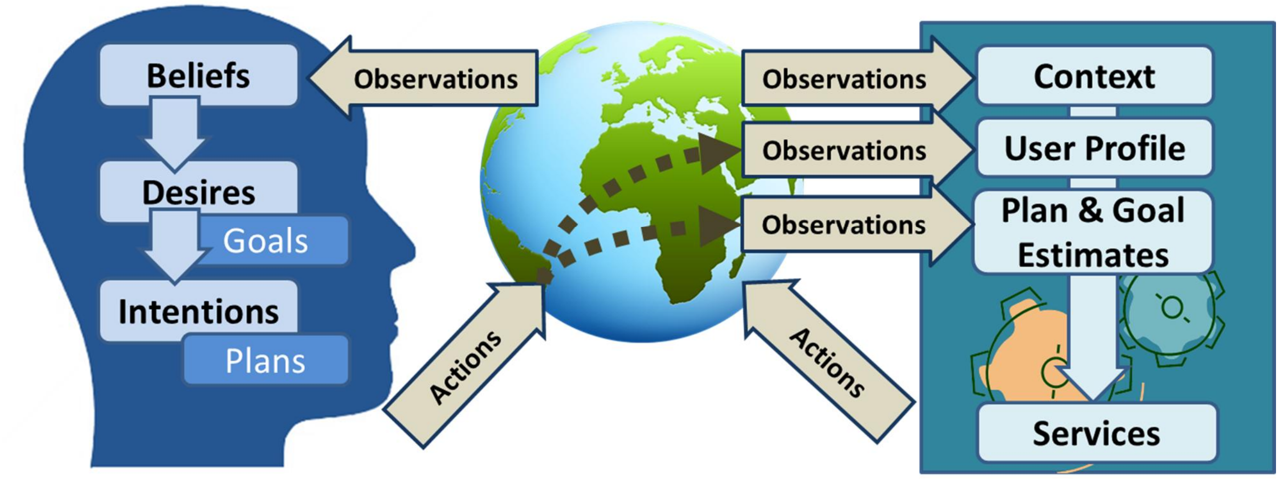


Figure 39: A model for intention aware services basing plan and goal estimates on user action observations

An important model used in software engineering explicitly utilising intention is the Belief-Desire-Intention model developed by Bratman [Bra99] [Geo99]. The model has been very influential for the communication between agents. It is based on a psychological model of the behaviour of people. Beliefs are the knowledge obtained by an agent about its environment, where the term belief emphasises the fact that this knowledge is not necessary a true picture of the actual situation, but there is an uncertainty factor. Desires depict the set of possibly conflicting objectives of the agent. It differs from the goals, in that a goal is chosen from the desires to be pursued. An agent is not supposed to have conflicting goals, i.e. when a goal is formulated, a choice has been made between the desires. An intention describes in the BDI model a state of commitment to a desire (or goal). A desire becomes an intention when steps are made to pursue this desire. In practice this means that a plan is made to achieve the goal. The plan again consists of a sequence of consecutive actions.

The BDI model was successfully applied in a number of agent-based applications, most notably the fault diagnostics system of the Space Shuttle.

When viewing the user of a system as an agent, the BDI model would apply here as well. Becoming aware of the intention of the human agent would thus require an understanding of the beliefs and desires as well. In practice, only the actions of a human can be observed, so that unless the human agent provides explicit background information, the system needs to make assumptions on the BDI from these actions alone. When assuming that actions are a sequence that together form a plan, the first step would be plan recognition. However, plans are not always that easy to derive from the actions, as not all actions are necessarily related to one single plan, but the human agent may simultaneously pursue several plans, backtrack on his/her actions, or have other unexpected behaviour. Given the uncertainty factors involved in plan recognition, establishing a full BDI model may be a rather tedious task in a natural environment. Kiefer [Kie12] concludes in his dissertation, that for the purpose of intention recognition for location based services, the BDI approach is indeed not feasible due to its complexity, referring thereby to Schlieder [Sch05].

An overview of plan recognition techniques is given in [Carberry 2001]. Besides her application area in natural language processing, she identifies translation, CAD interfaces, help systems, collaborative problem solving and image sequence description as areas with attention to the topic. She describes a typical approach, which requires the composition of a plan library with recipes that specify the (sequence of) actions that can be performed in a system and the goals or subgoals that they achieve. Plan recognition is then achieved by observing the actions and forming hypotheses about the goals and plans followed. She identified three main challenges to the field: 1) coping with the noisy data and providing a robust system, 2) effective discrimination between the various hypothesis, and 3) scalable recognition algorithms.

Figure 39 shows a model of how the intention aware services might work in practice. The user's beliefs are based on the user's observation of the environment. Desires are formed, and goals are formulated after a selection process. Eventually the user will commit to these goals, formulate plans and execute his intentions. This results in actions, which can be observed by the system (services) in the environment. The intention aware service will collect these observations and form a model of the context, user, and the estimated plans and goals. This can subsequently be used to adapt the services and perform supportive actions. Note, that in an agent BDI model, the system actually also includes beliefs (the context and user model), and has desires and intentions of its own (implemented in the services, not shown in the figure).

5.2 Analysing/predicting user intentions.

Predicting user intention takes context awareness one step further by interpreting context sources to deduce user intentions. Machine learning algorithms widely applied to human behaviour detection. Among others, the following approaches have been the most commonly studied:

- Support Vector Machines (SVM) [N. Cristianini and J. Shawe-Taylor, "*An introduction to support vector machines and other kernel-based learning methods*", 1st ed. Cambridge University Press, March 2000] are supervised learning models based on the following methodology: input data vectors are non-linearly mapped to a very high-dimension feature space. In this feature space, a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine.
- Bayesian networks (BN) [F. Ruggeri, F. Faltin and R. Kenett, "*Bayesian Networks*", Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons, 2007] are a knowledge representation and reasoning mechanism. BN represent events and causal relationships between them as conditional probabilities involving random variables. Given the values of a subset of these variables (evidence variables) BN can compute the probabilities of another subset of variables (query variables). BN can be created automatically (learnt) by using statistical data (examples). Examples of special forms of BN that have been used for estimating user behavior are Naive Bayes Networks and Dynamic Bayesian Netowrks,
- Hidden Markov Models (HMMs) [Phil Blunsom, "*Hidden Markov Models*", 2004] is a statistical tool for modelling sequences of probabilistic states where the state sequence that the model passes through is unknown and only some probabilistic function of it is known. HMM and their extensions, such as the Hierarchical HMM, use sequential classification algorithms and because of their capability of incorporating statistical information about the activity dynamics, they offer the opportunity to interpret and answer queries at varying levels of abstraction.

Some solutions for estimating user intention based on the machine learning methods enumerated previously are the following:

- In [Oliver C. Schrempf and Uwe D. Hanebeck, "*A Generic Model for Estimating User-Intentions in Human-Robot Cooperation*", in Proceedings of the 2nd International Conference on Informatics in Control, Automation and Robotics, ICINCO 05] user intention is estimated as an algorithmic reasoning process based on a Dynamic Bayesian Network that infers hidden intentions from observed actions.
- In [Peter Krauthausen and Uwe D. Hanebeck, "*A Model-Predictive Switching Approach to Efficient Intention Recognition*", in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), pp. 4908-4913, 18-22 Oct. 2010] user intention is estimated based on three structured Dynamic Bayesian Networks for large environments with many features.
- In [Mitesh Patel, Jaime Valls Miro and Gamini Dissanayake, "*Probabilistic Activity Models to Support Activities of Daily Living for Wheelchair Users*", in Proceedings of IROS 2012 Workshop on Navigation and Manipulation Assistance for Robotic Wheelchairs] the intentions of a human wheelchair operator are calculated based on a Hierarchical Hidden Markov Model.

- In [B.M. Bowen, M.B. Salem, S. Hershkop, A.D. Keromytis and S.J. Stolfo, "*Designing Host and Network Sensors to Mitigate the Insider Threat*", in IEEE Security & Privacy Magazine, pp. 22-29, Nov-Dec 2009] a system for detection of malicious insider behaviors is designed based on a one-class Support Vector Machine.
- In [W. Li, "*SMART: An SVM-based Misbehavior Detection and Trust Management Framework for Mobile Ad hoc Networks*", 2010] the SVM method is used to distinguish misbehaving nodes from well-behaved nodes without requiring any pre-determined threshold to determine misbehaving nodes.

Other solutions that have been studied for estimating user intention based on different types of model are the following:

- In [Glenn Wasson, Pradip Sheth, Cunjun Huang and Re Ledoux Majd Alwan, "*A Physics-based Model for Predicting User Intent in Shared-Control Pedestrian Mobility Aids*", in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 04)] navigational intent is estimated based on the dynamic model of a walker system, with the key features of this model given by the fact that the walker system's wheels can slip, slide, and roll as the user and the walker controller cooperate or oppose each other.
- In [Pascal Bihler, Vasile-Marian Scuturici and Lionel Brunie, "*Expressing and Interpreting User Intention in Pervasive Service Environments*", in Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2005), vol. 4, pp. 102-106, November 27 - December 1, 2005] user intention is calculated based on a mathematical model and transformed into actions in a pervasive services environment.

5.3 Affective models

5.3.1 EmotionML (Emotion Markup Language)

5.3.1.1 Introduction[4].

The Emotion Markup Language (EmotionML) is a W3C candidate recommendation (10 May, 2012) aiming to strike a balance between practical applicability and scientific well-foundedness. The language is conceived as a "plug-in" language suitable for use in three different areas: manual annotation of data, automatic recognition of emotion-related states from user behavior, and generation of emotion-related system behavior.

So, roughly, use cases for EmotionML can be grouped into three broad types:

1. Manual annotation of material involving emotionality, such as annotation of videos, of speech recordings, of faces, of texts, etc;
2. Automatic recognition of emotions from sensors, including physiological sensors, speech recordings, facial expressions, etc., as well as from multi-modal combinations of sensors;
3. Generation of emotion-related system responses, which may involve reasoning about the emotional implications of events, emotional prosody in synthetic speech, facial expressions and gestures of embodied agents or robots, the choice of music and colors of lighting in a room, etc.

Concrete examples of existing technology that could apply EmotionML include:

- Opinion mining / sentiment analysis in Web 2.0, to automatically track customer's attitude regarding a product across blogs;
- Affective monitoring, such as ambient assisted living applications for the elderly, fear detection for surveillance purposes, or using wearable sensors to test customer satisfaction;
- Character design and control for games and virtual worlds;
- Social robots, such as guide robots engaging with visitors;
- Expressive speech synthesis, generating synthetic speech with different emotions, such as happy or sad, friendly or apologetic; expressive synthetic speech would for example make more

[4] This section has been taken directly from W3C EmotionML specification (10 May 2012). http://www.w3.org/TR/2012/CR-emotionml-20120510/

information available to blind and partially sighted people, and enrich their experience of the content;

- Emotion recognition (e.g., for spotting angry customers in speech dialog systems);
- Support for people with disabilities, such as educational programs for people with autism. EmotionML can be used to make the emotional intent of content explicit. This would enable people with learning disabilities (such as Asperger's Syndrome) to realize the emotional context of the content;
- EmotionML can be used for media transcripts and captions. Where emotions are marked up to help deaf or hearing impaired people who cannot hear the soundtrack, more information is made available to enrich their experience of the content.

In April 2011, an XML Schema and a MIME-type for EmotionML were defined. The following example shows how automatically annotated data from three affective sensor devices might be stored or communicated:

```
<emotionml xmlns="http://www.w3.org/2009/10/emotionml"
    category-set="http://www.w3.org/TR/emotion-voc/xml#everyday-categories">
 ...
<emotion start="1006526160000" expressed-through="face">
  <!--the first modality detects excitement.
      It is a camera observing the face. A URI to the database
      is provided to access the video stream.-->
  <category name="excited"/>
  <reference uri="http://www.example.com/facedb#t=26,98"/>
</emotion>

<emotion start="1006526160000" expressed-through="facial-skin-color">
  <!--the second modality detects anger. It is an IR camera
      observing the face. A URI to the database
      is provided to access the video stream.-->
  <category name="angry"/>
  <reference uri="http://www.example.com/skindb#t=23,108"/>
</emotion>

<emotion start="1006526160000" expressed-through="physiology">
  <!--the third modality detects excitement again. It is a
      wearable device monitoring physiological changes in the
      body. A URI to the database
      is provided to access the data stream.-->
  <category name="excited"/>
  <reference uri="http://www.example.com/physiodb#t=19,101"/>
</emotion>

<emotion start="1006526520000" expressed-through="physiology">
  <category name="angry"/>
  <reference uri="http://www.example.com/physiodb2#t=2,6"/>
</emotion>
 ...
</emotionml>
```

Example 1. “Automatically annotated data from three affective sensor devices”.

5.3.1.2 Sentiment Analysis and Emotion Detection from Textual Information.

Within the EMPATHIC PRODUCTS project, LORIA (UMR 7503) / University of Lorraine will use EmotionML for sentiment analysis and emotion recognition. The general idea is to perform sentiment analysis and emotion recognition from the syntactic and semantic analysis of sentences from chat-based tools or social networks such as Twitter™. The following example shows how annotated lexical data might be associated to Ekman’s basic emotions [Ekm99]:

```
<sentence id="1414">
        <content>Trolley Square shooting leaves 6 dead, including gunman</content>
        <tagged-content>
```

```
            <tagged-word pos="NN" lemma="trolley" morph="trolley"/>
            <tagged-word pos="NN" lemma="square" morph="Square"/>
            <tagged-word pos="NN" lemma="shooting" morph="shooting"/>
            <tagged-word pos="VBZ" lemma="leave" morph="leaves"/>
            <tagged-word pos="CD" lemma="6" morph="6"/>
            <tagged-word pos="JJ" lemma="dead" morph="dead"/>
            <tagged-word pos="," lemma="," morph=","/>
            <tagged-word pos="VBG" lemma="include" morph="including"/>
            <tagged-word pos="NN" lemma="gunman" morph="gunman"/>
        </tagged-content>
        <emotion>
            <info>
                <origin source="semeval-ekman"/>
            </info>
            <category confidence="1.0" value="0.87" name="sadness"/>
            <category confidence="1.0" value="0.18" name="surprise"/>
            <category confidence="1.0" value="0.0" name="joy"/>
            <category confidence="1.0" value="0.12" name="anger"/>
            <category confidence="1.0" value="0.0" name="disgust"/>
            <category confidence="1.0" value="0.57" name="fear"/>
        </emotion>
        <emotion>
            <info>
                <origin source="semeval"/>
            </info>
            <category confidence="1.0" value="0.0" name="valence"/>
        </emotion>
</sentence>
```

Example 2. “Associating basic emotions to annotated lexical data”.

Remarks and comments about Example 2:

- <sentence> represents the sentence that has been analyzed;
- <content> represents the sentence as a string;
- <tagged-content> contains the elements in the sentence. The meaning of attributes is the following:
    - pos: part-of-speech morphological tags. Morphological tags represent different kinds of words such as: verbs, nouns, adjectives, etc.
    - lemma: the canonical form of a word;
    - morph: a word.
- <emotion> an EmotionML element allowing to associate emotions to sentences. This element is fully based on EmotionML.

In the framework of sentiment analysis, we will also use Emotion-ML. The following example shows how annotated lexical data might be associated to positive, negative, or neutral sentiments:

```
<sentence id="266290644207681536">
        <content>@bryony_q Omg really totes loving this.. :) xxx</content>
        <emotion>
            <info>
                <origin source="twitter"/>
            </info>
            <category confidence="1.0" value="1.0" name="valence"/>
        </emotion>
</sentence>
<sentence id="266460172858310656">
        <content>Major migrane going on now :( q</content>
        <emotion>
            <info>
                <origin source="twitter"/>
            </info>
            <category confidence="1.0" value="0.0" name="valence"/>
        </emotion>
</sentence>
```

Example 3. “Associating positive, negative, or neutral sentiments to annotated lexical data”.

Remarks and comments about Example 3:

- <sentence> represents the sentence that has been analyzed;
- <content> represents the sentence as a string;
- <emotion> an EmotionML element allowing to associate sentiments to sentences;
- Within the <category> element, the “value” attribute means: 0.0 is a fully negative sentiment, 1.0 is a fully positive sentiment, and 0.5 is a neutral sentiment.

5.3.2 Psycho-cognitive models

As Gregory Beller indicated in [Bel08], emotions include:

- Practical emotions : facilitating our adaptation or our reactions to a important event
- Preferences : judgements relating to the appraisal or the comparison between objects (“I like”, “I don't like”...)
- Attitudes : judgements, beliefs, relatively stable
- Moods : diffuse affective states, whose cause can hardly be known (melancholic, apathetic, depressed...)
- Personal traits (nervous, jealous...)
- Relational position : spontaneous or strategic affective style (polite, cold, friendly...)
- Aesthetical emotions : uninterested pleasures as defined by the philosopher Kant (afraid, fascinated, moved...)

In affective computing of course, it has seemed at first sight very difficult to go so deep into such details. The emotions are most represented with model validated by psychology literature.

5.3.2.1 Categorical representations and reflexion

According to Paul Ekman, emotions are the mean we use to survive in the nature, and consist in a set of universal categories. In an article published in 1999, he explains that there is an agreement on 6 emotions: joy, sadness, anger, surprise, fear, disgust. He talks about emotional needs, innate, necessary, in the continuity of Charles Darwin [Dar72].

Since 1984, the “palette-theory” of Scherer allows express secondary emotions as a composition of primary emotions.

In [Cra94], Vincent Crapanzano thinks about the possibility of an anthropology of emotions and quotes several researchers. According to Daniel Rosenberg, it exists a terminology delimitation of the anterior states due to culture and notably language. European languages have a nominalization grammar, or at least a grammar focusing on categories with limits, and Papataxiarchis explains that such a grammar restricts the emotional descriptions to relationship between words and referents. Crapanzano puts the emphasis on the idea of “emotional discourse”: the emotions are produced by a context as they create a context. A particular view on our emotions can reveal an independency regarding someone else, a continuity or on the contrary a dependency, a discontinuity, the particular.

In [Dev03], Laurence Devillers develops the concept of “lexical vectors” to describe the emotions, when several words are necessary to characterize an affective state. She also tackles the problem of multiple emotions.

5.3.2.2 Dimensional representations

The dimensional representations consist in describing emotions as elements in a vector space.
Below, we give some examples of spaces: most of them have 2 or 3 dimensions:

The tool “FeelTrace” of Schöeder (2003) is used for the annotation of videos and allows the coder to move a marker towards 2 directions in the plan corresponding to 2 dimensions :

- positive/negative
- aroused/calm

The emotional state can be described all through the video as a vector of 2 values.

The Wundt's space comprises 3 dimensions:

- positive/negative
- aroused/calm
- tension/release

The affective lexical space of Fitrianie (2006) resembles the one of Schröder :

- positive/negative
- active/passive

The Genève wheel of Scherer (2005) comprises 4 dimensions :

- valence
- control
- arousal
- conductive/obstructive

The PAD (Pleasure, Arousal, Dominance) space (2010) is a 3D space sometimes completed with 2 dimensions :

- valence
- arousal
- dominance/submission
- (intensity)
- (anticipation)

In [Dev12], Devillers et al. propose an new annotation protocol which combines the categorical and dimensional approches.

5.3.2.3 Appraisal representations

According to K. Scherer, the perception and the cognitive appraisal of an event determine the type and intensity of the emotion felt by a person.

Darwin considered the emotions as prototypical reactions whose functions were linked to survival. In 1884, William James put the emphasis on the link between the reflex of the emotional reaction and the perception of this reaction which constitutes the proper emotion, and then introduced the concept of cognitive appraisal process.

In 1980, Averill opposes the social-constructivist theories to the automatism of the emotional reactions highlighted by the unconscious neural circuits between perception and action.
The following models are mainly used in emotion generation.

5.3.2.3.1 The O.C.C. model

In 1988, the theory of Ortony, Clore and Collins [Ort88] divides the emotions into 3 categories:

- the ones relating to the interpretation of events
- the ones relating to the actions of individuals or oneself
- the ones relating to the evaluation of objects

We will stress the fact that for such representation, the place of the subject is of crucial importance. For example, a victory can imply pride for the winner and deception for the looser. The temporal aspect has to be indexed too. In effect, a victory can be celebrated (present) or hoped (future).

The O.C.C. Model defines 3 classes of emotions. Evaluation criteria is associated with each class and permits to determine the type and intensity of the emotion created according to the aims, principles and preferences of the agent. The authors of this model specify the mental states corresponding to 22 typical emotions

5.3.2.3.2 The Component Process Model

In [San05] is exposed an evaluative theory of emotions.

The cognitive appraisal has primitive functions:

- certainty
- agreement
- responsibility
- effort
- control

K. Scherer studied the notion of "control" and developed the "push-pull" theory. "Push" refers to indices we can't control (acoustic indices due to physiologic and somatic changes). "Pull" refers to indices we can control (socio-cultural codes).

The emotion is presented as a processus consisting in the appraisal of different stimuli which relate to:

- the relevance of an event
  - Novelty check
  - Intrinsic pleasantness check
  - Goal/need relevance check
- the social implication of the event
  - Causal attribution check
  - Outcome probability check
  - Discrepancy-from-expectation check
  - Goal/need conduciveness check
  - Urgency check
- the coping potential determination
  - Control
  - Power
- the normative evaluation
  - Internal standards check
  - External standards check

The appraisal processing belongs to a general information processing system which notably manages the action tendencies.

We can notice that in these stimuli, the subjectivity of the feelings doesn't seem to be taken into account. In fact, Scherer integrates it in the representation of the responses driven by the appraisal: the more the events are considered to be relevant, the more they generate consciousness, and the subjective feelings are precisely conscious. When an event occurs, the brain integrates all the changes of all the components in representations; one part of these representations becomes conscious and is used for regulation (personal representation, social norms): subjectivity takes place here, and on part of these conscious representations can be verbally expressed.

We will conclude this paragraph on appraisal representation by quoting the work of Antonio Damasio whose book on Spinoza is introduced in [Dam03]. The philosopher said several centuries ago that emotions relate to the body and precede the feelings which relate to the mind and are determined by the experience, are produced by reason and imagination. On the contrary, René Descartes thought that the body and the soul were two different substances. In his book, Damasio draws a parallel between the idea of a unique substance (as Spinoza thought) and the material conception of mind who has known for several years of growth of interest with neurosciences.

6. Affective feedback technologies

It is still difficult to create a purely emotional interaction between a Human and a Computer or an Agent. Currently, we are still talking about ”Rational Agents” or ”Conversational Agents”, and most of them are built on the OCC model [Ort88]. In the following list, we introduce the main improvments on the emotion computational models:

- In the P.E.T.E.E.I. (P.E.T. with Evolving Emotional Intelligence) model [Nas98], the emotion is generated according to the desirability of an event, and its probability to be realized.
- The model developed by Prendinger and al. [Pre02] comprises a module which evaluates the emotional meaning of an event in terms of aim, principles and preferences. Here, the uncertainty is not considered.
- In the E.M.A. (Emotion and Adaptation) model [Gra04], the emotion is generated according to the desirability of an event, and its probability to be realized (as for the P.E.T.E.E.I.), but it also takes into account the type of agent responsible for the event and the coping degree of the agent regarding the situation, thanks to a causal representation of the events (past, present and future) and the resulting agent’s states.
- The model developed by Rosis and al [Ros03] uses a dynamic network of beliefs: the emotions are generated when the agent undergoes a change of belief concerning the realization or the threat of one of its goals.
- The “Beliefs, Desires and Intentions” approach (B.D.I.) provides a framework for the representation of mental states. The technology A.R.T.I.M.I.S [Sad97] developed by France Télécom uses the Theory of Rational Interaction (T.R.I.) which a variation of the ”B.D.I.” approach. It defines 3 primitive mental attitudes of a rational agent regarding a particular situation (event, action or object, according to the O.C.C model): Beliefs, Uncertainties, and Choices. The configuration of primitive mental attitudes constitute the mental state of the agent. In other words, evaluation criteria constitute the mental state of an emotion.
- In [Och05] is M. Ochs and al. propose generation rules for two types of emotion mixture: superposition of emotions and blended of one emotion by another. They show examples of facial expression resulting of these complex emotions.
- Finally in [Gri06], A. Grizard and al. work with the appraisal model of Scherer [San05] to animate a robot developed by Philips (the platform ”iCat”) and a graphic avatar created with the commercial tool ”Haptek”. the facial expressions based on the famous Action Units of Ekman and Friesen [Ekm02])

In [Hud12], E. Hudlicka and H. Gunes think about an intermediary level between the expressive features and the emotional analysis. This level would relate to cognitive appraisal and the mapping should be studied:

- the relation between appraisal variables and expressiveness
- the relation between emotion and appraisal variables

The use of appraisal variables could notably allow to better taking into account the context (notably the variables related to stimuli of relevance and social implications provoked by an event [San05]).
Hudlicka and H. Gunes also wonder whether it is possible to characterize the information brought by a channel (voice, face, hands, head...) and then conclude about the conscious or unconscious use of this channel in the emotion production processing. Is this “channel call” different according to the stimuli, or in other words according to the appraisal type?

More generally, is it only possible to analyse to spontaneous state of an agent, or can't we consider larger temporal sequences and characterize emotional state in the longer term mood: long term)

- Mood: long term
- Personality: very long term.

6.1 Avatar

Avatars have, among others, the following means to transmit affective feedback:

- *Facial Expressions (can be stylized or realistic).*
- *Corporal expressions (actions like dancing or being still for too long).*
- *Color settings (blue for sadness or red for excitement or wanting to intervene).*
- *Face Detection Video conference (overlapped on the head paralelipiped).*

7. Conclusions

This document is a state of the on affective technologies used in the Empathic Products project. This document is build around technologies available in the project.

8. References

[A]	"Immersion Corporation, Cyberglove II Datasheet." [Online]. Available : http ://www.immersion.com/3d/docs/cybergloveII dec07v4-lr.pdf
[Age10]	Agerri, R., Garcia-Serrano, A. 2010. Q-WordNet : Extracting Polarity from WordNet Senses. In *Proceedings of the Language Resources and Evaluation Conference* (LREC 2010). 2010.
[Agg97]	J. K. Aggarwall and Q. Cai, "Human motion analysis: a review," in IEEE Proceedings of Nonrigid and Articulated Motion Workshop, 1997.
[Alb12]	A. Albiol, J. Oliver and M. Mossi, "Who is who at different cameras: people re-identification using depth cameras," Computer Vision, IET, vol. 6, no. 5, pp. 378-387, 2012.
[Alh06]	M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang, "Multimodal Integration for Meeting Group Action Segmentation and Recognition," in *Machine Learning for Multimodal Interaction*, vol. 3869, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 52–63.
[Alh07]	M. Al-Hames, B. Hornler, R. Muller, J. Schenk, and G. Rigoll, "Automatic Multi-Modal Meeting Camera Selection for Video-Conferences and Meeting Browsers," in *Multimedia and Expo, 2007 IEEE International Conference on*, 2007, pp. 2074 –2077.
[Alh11]	Alhothali, Areej, Modeling User Affect Using Interaction Events, A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree of Master of Mathematics in Computer Science, Waterloo, Ontario, Canada, 2011, http://hdl.handle.net/10012/6027
[Alm05]	Alm, C., Roth, D., Sproat, R. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of Human Language Technologies Conference / Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP 2005), Vancouver, Canada.
[An10]	Angela Caunce, Chris Taylor, T. C. "Improved 3D Model Search for Facial Feature Location and Pose Estimation in 2D images" *BMVC,* 2010
[And07]	Andreevskaia, A., Bergler, S. 2007. CLaC and CLaC-NB : Knowledge-based and Corpus-based Approaches to Sentiment Tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007), pp. 117-120, 2007.
[And08]	Andreevskaia, A., Bergler, S. 2008. When Specialists and Generalists Work Together : Overcoming Domain Dependence in Sentiment Tagging. In *Proceedings of ACL-2008 : HLT*, pp 290-298. Columbus, USA. 2008.
[AOKI2009]	Aoki, H., Hansen, J.P., Itoh, K. (2009). Learning gaze typing: what are the obstacles and, what progress to expect? *Universal Access in the Information Society*, 8(4), 297–310.
[Ash07]	A.B. Ashraf, S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B.J. Theobald, "The Painful Face: Pain Expression Recognition Using Active Appearance Models," Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07), pp. 9-14, 2007
[Aue05]	Aue, A., Gamon, M. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP-05). 2005.

[B]	http://the-distance-learning.com/context-aware-computing-understanding-human-intention
[Ba04]	Baker, S. Matthews, I. Xiao, J. Gross, R. Kanade, T. & Ishikawa "Real-Time Non-Rigid Driver Head Tracking for Driver Mental State Estimation" *in 11th World Congress on Intelligent Transportation Systems,* 2004
[Ba08]	Jan Bandouch, Florian Engstler and Michael Beetz "Evaluation of Hierarchical Sampling Strategies in 3D Human Pose Estimation" *BMVC,* 2008
[Ba11]	Ba, S. O. & Odobez, J.-M. "Multiperson Visual Focus of Attention from Head Pose and Meeting Contextual Cues" *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2011, 33, 101-116
[Ba94]	Beymer, D. J. "Face Recognition Under Varying Pose" *CVPR,* 1994, 756-761
[Bac10]	Bacciannella, S., Esuli, A., Sebastiani, F. 2010. SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Language Resources and Evaluation Conference* (LREC 2010). 2010.
[Bak10]	Baker, R., D'Mello, S., Rodrigo, M.M., Graesser, A.C. 2010. Better to be frustrated than bored : the incidence, persistence, and impact of leaerner's cognitive-affective states during interactions with three different computer-based learning environments. In *International Journal of Human-Computer Studies,* Volume 68, Issue 4, April, 2010, pp 223-241.
[Bal04]	T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion Analysis in Man-Machine Interaction Systems," Proc. Workshop Machine Learning for Multimodal Interaction, pp. 318-328, 2004
[Bal11]	Balahur, A., Hermida, J.M, Montoyo, A., Muñoz, R. 2011. EmotiNet : A Knowledge Base for Emotion Detection in text Built on the Appraisal Theories. In *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems* (NLDB 2011), pp. 27-39.
[Bal11]	Leslie Ball, David Bradley, Andrea Szymkowiak, Simon Brownsell, Linking recorded data with emotive and adaptive computing in an eHealth environment, 2011 IEEE First International Conference on Healthcare Informatics Imaging and Systems Biology IEEE, 2011.
[Ban08]	Banea, C., Mihalcea, R., Wiebe, J., Hassan, S. 2008. Multilingual Subjectivity Analysis Using Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2008). 2008.
[Ban10]	Banea, C., Mihalcea, R., Wiebe, J. 2010. Multilingual Subjectivity: Are More Languages Better? In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING-2010). 2010.
[Ban11]	Banea, C., Mihalcea, R., Wiebe, J. 2011. Multilingual Sentiment and Subjectivity Analysis. In *Multilingual Natural Language Processing*, editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.
[Ban12]	S. Banerjee, D. L. Woodard, Biometric Authentication and Identification Using Keystroke Dynamics: A Survey, Journal of Pattern Recognition Research 7 (2012): pg. 116-139, 2012.
[Bas08]	Shishir Bashyal, Ganesh K. Venayagamoorthy, Recognition of facial expressions using Gabor wavelets and learning vector quantization, Engineering Applications of Artificial Intelligence, Volume 21, Issue 7, October 2008, Pages 1056-1064, ISSN 0952-1976, 10.1016/j.engappai.2007.11.010.
[Be08]	Benfold B. & Reid I. "Colour Invariant Head Pose Classification in Low Resolution Video" *BMVC,* 2008
[Be10]	BenAbdelkader C. "Robust Head Pose Estimation Using Supervised Manifold Learning" *ECCV,* 2010
[Bel08]	PhD. Thesis, Université Paris VI, "Analyse et Modèle Génératif de l'Expressivité -Application à la Parole et à l'Interprétation Musicale" , Grégory Beller, Thèse soutenue le mercredi 24 Juin 2009
[Ber07]	Daniel Bernhardt and Peter Robinson. 2007. Detecting Affect from Non-stylised Body Motions. In Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII '07), Ana C. Paiva, Rui Prada, and Rosalind W. Picard (Eds.). Springer-Verlag, Berlin, Heidelberg, 59-70.
[Ber07]	D. Bernhardt and P. Robinson, "Detecting Affect from Non-Stylised Body Motions," Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction, pp. 59-70, 2007
[Ber09]	D. Bernhardt, P. Robinson, "Detecting Emotions from Connected Action Sequences, Visual Informatics: Bridging Research and Practice," LNCS: 5857, pp. 1-11, 2009.
[Bi11]	G. Bijlstra and R. Dotsch (2011). FaceReader 4 emotion classification performance on images from the Radboud Faces Database. Unpublished manuscript retrieved from http://www.gijsbijlstra.nl/ and http://ron.dotsch.org/.

[Bi95]	C.M. Bishop. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995.
[Bin11]	Bingbing Ni; Gang Wang; Moulin, P.; , "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on , vol., no., pp.1147-1153, 6-13 Nov. 2011
[Bir66]	R.L. Birdwhistell. Some relations between American kinesics and spoken American English. In A.G. Smith (Ed.), Communication and culture. New York: Holt, Rinehart and Winston, 1966.
[Bla95]	M.J. Black, Y. Yacoob, *Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion*, in: Proc. Internat. Conf. Computer Vision (ICCV95), 1995, pp. 374–381
[Bra98]	G.R. Bradski. “Real time face and object tracking as a component of a perceptual user interface”. In Proceedings of Fourth IEEE Workshop on Applications of Computer Vision (WACV) ‘98, pages 214-219, 1998.
[Bra99]	Bradley, M.M., Lang, P.J. 1999. *Affective Norms for English Words (ANEW): Stimuli, instruction manual, and affective ratings* (Tech. Report C-1). Gainesville: University of Florida, Center for Research in Psychophysiology.
[Bra99]	Bratman, M. 1999, *Intention, plans, and practical reason* Center for the Study of Language and Information, Stanford, Calif.
[Buc80]	G. Buchsbaum, “A spatial processor model for object colour perception,” Journal of Franklin Institute, vol. 310, pp. 1–26, 1980.
[Bul77]	P.E. Bull, and R. Brown, The role of postural change in dyadic conversation. British Journal of Social and Clinical Psychology, vol. 16, 1977, pp. 29-33.
[Cag06]	M. Baris Caglar and Niels Lobo, “Open Hand Detection in a Cluttered Single Image using Finger Primitives”, Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, 2006.
[Cal10]	Calvo, R.A., D'Mello, S. 2010. Affect Detection : An Interdisciplinary Review of Models, Methods, and Their Applications. In *IEEE Transactions on Affective Computing*, Volume 1 Issue 1, January 2010, pp 18-37.
[Cam04]	A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe,“Multimodal analysis of expressive gesture in music and dance Gesture-based Communication in HCI, pp. 20-39, 2004.performances,”
[Ch11]	I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato and A. Sugimoto “Appearance-Based Head Pose Estimation with Scene-Specific Adaptation” *ICCV,* 2011
[Cha07]	Chaumartin, F-R. 2007. UPAR7 : A Knowledge-Based System for Headline Sentiment Tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007), pp. 422-425, 2007.
[Che08]	Qing Chen, “Real-Time Vision-Based Hand Tracking and Gesture Recognition”, Phd thesis in Electrical and Computer Engineering, Ottawa-Carleton Institute for Electrical and Computer Engineering, Canada, 2008.
[Che09]	Yeongjae Cheon, Daijin Kim, Natural facial expression recognition using differential-AAM and manifold learning, Pattern Recognition, Volume 42, Issue 7, July 2009, Pages 1340-1350, ISSN 0031-3203, 10.1016/j.patcog.2008.10.010.
[Cho08]	Choi, Y., Cardie, C. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2008), pp 793-801. Honolulu, 2008.
[Cla10]	C. Clavel, G. Richard, *Reconnaissance acoustique des émotions*, Chapter 5 in Systèmes d’Interaction Emotionnelle, Editor: C. Pelachaud, Hermès, 2010 (in French).
[Co00]	T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, 2000.
[Co01]	T. Cootes, G. Edwards, and C. Taylor, “Active Appearance Models,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, June 2001
[COCAIN2012]	Available from: http://wiki.cogain.info/index.php/Eye_Trackers
[Coh03]	Facial expression recognition from video sequences: temporal and static modeling, Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen and Thomas S. Huanga. Computer Vision and Image Understanding 91 (2003) 160–187
[Col81]	Colby, K. 1981. Modeling a paranoid mind. *The Behavioral and Brain Sciences*, 4(4):515-560, December 1981.

[Con05]	Simon CONSEIL, Salah BOURENNANE, Lionel MARTIN, “Suivi Tridimensionnel en Stéréovision”, GRETSI, Groupe d’Etudes du Traitement du Signal et des Images, 2005.
[Coo06]	Thomas Coogan, George Awad, Junwei Han and Alistair Sutherland, “Real Time Hand Gesture Recognition Including Hand Segmentation and Tracking”, In: ISVC 2006 – 2nd International Symposium on Visual Computing, 2006.
[Coo95]	T. F. Cootes , C. J. Taylor , D. H. Cooper , J. Graham, Active shape models—their training and application, Computer Vision and Image Understanding, v.61 n.1, p.38-59, Jan. 1995 .
[Cou12]	Coutrix, C., Mandran, N., Identifying Emotions Expressed by Mobile Users through 2D Surface and 3D Motion Gestures, In Proceedings of the 14th ACM International Conference on Ubiquitous Computing (Ubicomp'12), September 5-8, 2012, Pittsburgh, Pennsylvania, United States.
[Cow01]	Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. 2001. Emotion recognition in human-computer interaction. In *Signal Processing Magazine, IEEE*, Vol 18. Issue 1. pp 32-80. Jan 2001.
[Cow03]	Cowie, R., Cornelius, R.R., Describing the emotional states that are expressed in speech (2003) Speech Communication, 40 (1-2), pp. 5-32.
[Cra94]	Vincent Crapanzano, « Réflexions sur une anthropologie des émotions », *Terrain*, 22 | 1994,mis en ligne le 15 juin 2007, 09 octobre 2012. URL : http://terrain.revues.org/3089 ; DOI : 10.4000/ terrain.3089
[Cri07]	Cristiana Bolchini and Carlo A. Curino and Elisa Quintarelli and Fabio A. Schreiber and Letizia Tanca (2007). "A data-oriented survey of context models" (PDF). *SIGMOD Rec.* (ACM) **36** (4): 19–26. doi:10.1145/1361348.1361353
[Cut98]	Ross Cutler and Matthew Turk, “View-based Interpretation of Real-time Optical Flow for Gesture Recognition”, Proceeding on Third IEE International Conference on Automatic Face and Gesture Recognition, 1998.
[Da11]	Dahmane, M. & Meunier, J. “Object Representation Based on Gabor Wave Vector Binning : An Application to Human Head Pose Detection” *ICCV,* 2011
[Dam03]	Antonio Damasio : « Spinoza *avait raison : joie et tristesse, le cerveau des émotions* », Odile Jacob, Paris, 2003, 346 p. (ISBN 2738112641).
[Dan08]	Danisman, T., Alpkocak, A. 2008. Feeler : Emotion Classification of Text Using Vector Space Model. In *AISB 2008 Convention, Communication, Interaction and Social Intelligence*, Vol. 2 (April 2008).
[Dan12]	Taner Danisman, Ioan Marius Bilasco, Jean Martinet, Chabane Djeraba, Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron, Signal Processing, Available online 20 August 2012, ISSN 0165-1684, 10.1016/j.sigpro.2012.08.007.
[Dar72]	Charles Darwin: “The Expression of the Emotions in Man and Animals”, London, John Murray, 1872
[Das01]	Das, S., Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of APFA-2001*. 2001.
[Dav10]	Davidov, D., Tsur, O., Rappaport, A. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceeding of the 23rd International Conference on Computational Linguistics* (COLING-2010). 2010.
[Den02]	Jiangwen Deng and H.T. Tsui, “A PCA/MDA Scheme for Hand Posture Recognition”, Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002.
[Den08]	Denecke, K. 2008. Using SentiWordNet for Multilingual Sentiment Analysis. In *Proceedings of the IEEE 24th International Conference on Data Engineering Workshop* (ICDEW 2008).
[Des09]	Sachin Deshpande, Chang Yuan, Scott Daly, and Ibrahim Sezan, "A Large Ultra High Resolution Tiled Display System: Architecture, Technologies, Applications, and Tools," in the Proc. of 16th International Display Workshops, Miyazaki, Japan, 12/2009
[Deubel2012]	Available publication list from: http://www.psy.lmu.de/exp/people/prof/deubel/publications/index.html
[Dev03]	“Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches”
[Dev12]	L. Devillers, R. Cowie, J-C. Martin, E. Douglas-Cowie, S. Abrilian, M. McRorie, LIMSI-CNRS; France, Queen’s University Belfast; UK{devil, martin,abrilian}@limsi.fr, {r.cowie, e.douglas-Cowie, m.mcrorie}@qub.ac.uk

[Dey01]	Dey, Anind K. (2001). "Understanding and Using Context". *Personal Ubiquitous Computing* **5** (1): 4–7. doi:10.1007/s007790170019.
[Din06]	Y. Ding and G. Fan, “Camera View-Based American Football Video Analysis,” in *Multimedia, 2006. ISM’06. Eighth IEEE International Symposium on*, 2006, pp. 317 –322.
[Do11]	Dong Huang, Markus Storer, Fernando De la Torre and Horst Bischof “Supervised Local Subspace Learning for Continuous Head Pose Estimation” *CVPR,* 2011
[Done2005]	Donegan, M., Oosthuizen, L., Bates, R., Daunys, G., Hansen, J. P., Joos, M., Majaranta, P., & Signorile, I.(2005). User requirements report with observations of difficulties users are experiencing. Communication by Gaze Interaction, IST-2003-511598, available at http://wiki.cogain.org/images/e/ef/COGAIN-D3.1.pdf
[Dor11]	F. Dornaika, E. Lazkano, B. Sierra, Improving dynamic facial expression recognition with feature subset selection, Pattern Recognition Letters, Volume 32, Issue 5, 1 April 2011, Pages 740-748, ISSN 0167-8655, 10.1016/j.patrec.2010.12.010.
[Dou07]	E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, and K. Karpouzis, “The humaine database: addressing the needs of the affective computing community,” in *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, 2007, pp. 488–500.
[Drewes2010]	Heiko Drewes (2010), Eye Gaze Tracking for Human Computer Interaction, Dissertation an der LFE Medien-Informatik der Ludwig-Maximilians-Universität, München
[DUCHO2002]	DUCHOWSKI ANDREW T. (2002), A breadth-first survey of eye-tracking applications, *Behavior Research Methods, Instruments, & Computers, 34 (4), 455-470.*
[Ducho2003]	Duchowski, A. T. (2003). *Eye tracking methodology: Theory and practice*. London: Springer-Verlag Ltd.
[Ek70]	P. Ekman. Universal facial expressions of emotion. California Mental Health Research Digest, 8: 151-158, 1970.
[Ekm02]	Ekman P., Friesen W. V., Hager J. C., *Facial Action Coding System Investigator’s Guide*, A Human Face, 2002
[Ekm72]	Ekman, P. 1972. Universals And Cultural Differences In Facial Expressions Of Emotions. In J. Cole (ed.), *Nebraska Symposium on Motivation*, 1971. Lincoln, Neb.: University of Nebraska Press, 1972. pp 207- 283.
[Ekm78]	Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978).
[Ekm99]	Ekman, P. 1999**.** Basic emotions. In T. Dalgleish and T. Power (Eds.) The handbook of cognition and emotion, pages 45-60. New York. Ed. John Wiley & Sons.
[Ell92]	Elliott, C. 1992. The Affective Reasoner: A Process Model of Emotions in a Multi-agent System. PhD thesis, Northwestern University, May 1992. The Institute for the Learning Sciences, Technical Report No. 32.
[Eng06]	I. N. Engleberg and D. R. Wynn, *Working in Groups: Communication Principles and Strategies*. 2006.
[Epp11]	Epp, C., Lippold, M. Mandryk, R.L. Identifying emotional states using keystroke dynamics. In Proceedings of the Proceedings of the 2011 annual conference on Human factors in computing systems, ACM Press (2011), 715-724.
[Eri75]	F. Erikson, One function of proxemics shifts in face-to-face interaction. In A. Kendon, R. M. Harris and M.R. Key (Eds.), Organization of behaviour in face-to-face interaction. The Hague: Mouton, 1975.
[Ess98]	I.A. Essa, A.P. Pentland, *Coding, analysis, interpretation, and recognition of facial expressions*, IEEE Trans. Pattern Anal. Machine Intell. 19 (7) (1997) 757–763
[Esu06]	Esuli, A., Sebastiani, F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation* (LREC 2006), Genova, IT, 2006, pp. 417-422.
[Eva07]	Evans, D.K., Ku, L-W., Seki, Y., Chen, H-H., Kando, K. 2007. Opinion Analysis across Languages : An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task. In *Proceedings of the Workshop on Cross-Language Information Processing*, vol. 4578 *of Lecture Notes in Computer Science*, pp. 456-463. 2007.
[EyeDraw2012]	Available from: http://www.cs.uoregon.edu/Research/cm-hci/EyeDraw/
[facelab2012]	Available from: http://www.seeingmachines.com/product/facelab

[Fas04]	B. Fasel, F. Monay, and D. Gatica-Perez, “Latent Semantic Analysis of Facial Action Codes for Automatic Facial Expression Recognition,” Proc. Sixth ACM Int’l Workshop Multimedia Information Retrieval (MIR ’04), pp. 181-188, 2004.
[Fra09]	Frantova, E., Bergler, S. 2009. Automatic Emotion Annotation of Dream Diaries. In *Proceedings of the Analyzing Social Media to Represent Collective Knowledge Workshop at K-CAP 2009, The Fifth International Conference on Knowledge Capture*. 2009.
[Fre96]	Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the 13th International Conferene on Machine Learning, Morgan Kaufmann, 1996, pp. 148–146.
[Fun98]	B. Funt, K. Bernard, L. Martin, “Is Machine Colour Constancy Good Enough”, Proceedings of the 5th European Conference on Computer Vision (ECCV’98), Freiburg, Germany, pp. 445-459 , 1998
[Gar12]	M. Garber-Barron, M. Si, “Using body movement and posture for emotion detection in non-acted scenarios”, 2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, pp. 1-8, 2012.
[Ge02]	X. Ge and J. Tian, “An automatic active contour model for multiple objects,” IEEE, pp. 881-884, 2002.
[Gej04]	Peter Gejgus, Jaroslav Placek, Martin Sperka, “Skin color segmentation method based on mixture of Gaussians and its application in Learning System for Finger Alphabet”, International Conference on Computer Systems and Technologies, 2004.
[GEM11]	http://www.affective-sciences.org/gemep., 2011.
[Geo99]	Georgeff, M., Pell, B., Pollack, M., Tambe, M. & Wooldridge, Chapter 1, M. 1999; 2003, "Intelligent Agents V: Agents Theories, Architectures, and Languages; The Belief-Desire-Intention Model of Agency ", vol. 1555, pp. 1-10.
[Gil09]	Gilroy, S.W.; Cavazza, M.; Niiranen, M.; Andre, E.; Vogt, T.; Urbain, J.; Benayoun, M.; Seichter, H.; Billinghurst, M.; , "PAD-based multimodal affective fusion," Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on , vol., no., pp.1-8, 10-12 Sept. 2009
[Gir08]	Giroux S., Bauchet J., Pigot, H., Lusser-Desrochers, D., Lachappelle, Y. (2008). Pervasive behavior tracking for cognitive assistance. The 3rd International Conference on Pervasive Technologies Related to Assistive Environments, Petra’08, Greece, July 15-19, 2008.
[Glo10]	Multi-Scale Entropy Analysis of Dominance in Social Creative Activities, Donald Glowinski, Paolo Coletta, Gualtiero Volpe, Antonio Camurri, Carlo Chiorri, Andrea Schenone, 2010
[Glo11]	D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, K.Scherer, “Towards a Minimal Representation of Affective Gestures,”IEEE Trans. on Affective Computing, vol.2, no. 2, pp. 106-118, 2011.
[Glo11]	*Toward a Minimal Representation of Affective Gestures*, Donald Glowinski, Member, IEEE, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 2, NO. 2, APRIL-JUNE 2011
[Gold2003]	Goldberg, H. J., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner’s guide. In J. Hyönä, R. Radach, & H. Deubel (Eds.), The mind's eye:Cognitive and applied aspects of eye movement research (pp. 493-516). Amsterdam:Elsevier.
[Gon10]	M. Gonzalez, C. Collet, R. Dubot, “Head Tracking and Hand Segmentation during Hand over Face Occlusion in Sign Language”, International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV, 2010.
[Gra04]	Gratch, J. and Marsella, S. A *domain-independant Framework for modeling emotion.Journal of Cognitive Systems Research, 5 (4), 269-306.2004.*
[Gri06]	*Adaptation d’une théorie psychologique pour la generation d’expressions facials synthétiques pour des agents d’interface.* Amandine Grizard, Marco Paleari et Christine Lisetti, 2006, Groupe d’informatique affective sociale, Département communications multimedias, Institut Eurocom
[Gun07]	H. Gunes and M. Piccardi, “Bi-Modal Emotion Recognition from Expressive Face and Body Gestures,” J. Network and Computer Applications, vol. 30, no. 4, pp. 1334-1345, 2007
[Gun11]	”Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey”, Hatice Gunes, Bj¨orn Schuller, Maja Pantic and Roddy Cowie, European Community’s 7th Framework Programme [FP7/2007-2013]
[Gut96]	S. Gutta, J. Huang, I.F. Imam, and H. Wechsler.“Face and hand gesture recognition using hybrid classifiers”, Technical report, 1996.
[Ha11]	Hao Ji, Risheng Liu, Fei Su, Zhixun Su and Yan Tian “Robust Head Pose Estimation via Convex Regularized Sparse regression” ICIP 2011

[Hab05]	Hakan Haberdar and Songül Albayrak, “Real Time Isolated Turkish Sign Language Recognition from Video Using Hidden Markov Models withGlobal Feature”, Proceedings of Computer and Information Sciences, Image and Speech Processing, pp. 677-687, 2005.
[Hal061]	M. Al-Hames, B. Hörnler, C. Scheuermann, and G. Rigoll, “Using Audio, Visual, and Lexical Features in a Multi-modal Virtual Meeting Director,” in *Machine Learning for Multimodal Interaction*, vol. 4299, S. Renals, S. Bengio, and J. G. Fiscus, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 63–74.
[Ham02]	Yasushi HAMADA, Nobutaka SHIMADA, Yoshiaki SHIRAI, “Hand Shape Estimation Using Sequence of Multi-Ocular Images Based on Transition Network”, In Vi 2002, 2002.
[Han08]	D. W. Hansen, M. S. Hansen, M. Kirschmeyer, R. Larsen and D. Silvestre, “Cluster tracking with Time-of-Flight cameras,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08, 2008.
[Han99]	M. Handouyahia, D. Ziou and S. Wang, “Sign Language Recognition Using Moment-Based Size Functions”, in Proc. International Conference of Vision Interface, pp. 210-216, 1999.
[Hansen2009]	Dan Witzner Hansen, IEEE Member and Qiang Ji, IEEE Senior Member(2009), In the Eye of the Beholder: A Survey of Models for Eyes and Gaze, Available from http://people.ict.usc.edu/~gratch/CSCI534/WitznerJi_EyeTrackSurvey%282009%29.pdf
[Hat97]	Hatzivassiloglou, V., McKeown, K.R. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (EACL 1997), pp 174-181.
[Hea96]	A. J. Heap and D. C. Hogg, “Towards 3-D hand tracking using a deformable model”, In 2nd International Face and Gesture Recognition Conference, pp. 140–145, Killington, USA,Oct. 1996.
[Hel02]	Helena M. Mentis, and Geri K. Gay, “Using touchpad pressure to detect negative affect”, In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02). IEEE, 406, 2002.
[Her11]	Hernandez, D.; Castrillon, M.; Lorenzo, J.; , "People counting with re-identification using depth cameras," Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference on , vol., no., pp.1-6, 3-4 Nov. 2011
[Hor09]	B. Hörnler, D. Arsic, B. Schuller, and G. Rigoll, “Boosting multi-modal camera selection with semantic features,” in *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, Piscataway, NJ, USA, 2009, pp. 1298–1301.
[Hornof2004]	Hornof, A., Cavender, A., and Hoselton, R. EyeDraw (2004), A System for Drawing Pictures with the Eyes. In *Extended Abstracts on Human Factors in Computing Systems*. CHI '04. ACM Press (2004), 1251 – 1254.
[Hu04]	Hu, M., Liu, B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining* (KDD 2004), pp 168-177. 2004.
[Hu98]	Huang, J.; Shao, X. & Wechsler, H. “Face Pose Discrimination Using Support Vector Machines (SVM)” *Proceedings of the 14th International Conference on Pattern Recognition-Volume 1 - Volume 1, IEEE Computer Society,* 1998
[Hud12]	Guest editorial Preface, *Benefits and Limitations of Continuous Representations of Emotions in Affective Computing: Introduction to the Special Issue*, Eva Hudlicka, Psychometrix Associates, USA Hatice Gunes*,* Queen Mary, University of London, UK
[Hy08]	M.J. den Uyl and H. van Kuilenburg. The FaceReader: Online Facial Expression Recognition. Proceedings of Measuring Behavior 2005, Wageningen, The Netherlands, August 30 - September 2, 2008, pp. 589-590.
[Hyrsky2006]	Aulikki Hyrskykari (2006), Eys in Attentive Interfaces: Experiences from iDict, a Gaze-Aware Reading Aid, Academic Dissertation
[Is11]	I. Chamveha, Y. Sugano, D. S. T. S. T. O. Y. S. & Sugimoto, A. “Appearance-Based Head Pose Estimation with Scene-Specific Adaptation” *ICCV,* 2011
[Isa98]	M. Isard and A. Blake. “CONDENSATION—Conditional Density Propagation for Visual Tracking”, International Journal of Computer Vision, 1998.
[Ja99]	Jamie Sherrah, S. G. & Ong, E.-J. “Understanding Pose Discrimination in Similarity Space” *BMVC,* 1999
[Jac2003]	Jacob, R.J.K & Karn, K.S (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary). In J. Hyona, R. Radach & H.. Deubel (Eds.) The Mind’s Eyes: Cognitive and Applied Aspects of Eye Movements. Elsevier Science.

[Jai00]	Anil K. Jain, Fellow, Robert P.W. Duin, and Jianchang Mao, “Statistical Pattern Recognition: A Review”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, 2000.
[Jan08]	D. Janssen, W. I. Schollhorn, and J. Lubienetzki. Recognition of emotions in gait patterns by means of artificial neural nets. Journal of Nonverbal Behavior, vol. 32, 2008, pp. 79-92.
[Jep98]	Michael J. Black and Allan D. Jepson, “Recognizing Temporal Trajectories using the Condensation Algorithm”, IEEE, Automatic Face and Gesture Recognition, April 1998.
[Jou11]	E. Jouneau and C. Carincotte, “Particle-based tracking model for automatic anomaly detection,” in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 513 –516.
[Jun12]	H. Jungong, E. J. Pauwels, P. M. d. Zeeuw and P. H. N. d. With, “Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment,” IEEE Transactions on Consumer Electronics, pp. 255-263, May 2012.
[Jus06]	Agnès Just, Yann Rodriguez and Sébastien Marcel, “Hand Posture Classification and Recognition using the Modified Census Transform”, Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, 2006.
[Kak09]	Kaklauskas, A, Krutinis, M, Seniut, M., Biometric Mouse Intelligent System for Student’s Emotional and Examination Process Analysis, Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on Digital Object Identifier: 10.1109/ICALT.2009.130, Publication Year: 2009 , Page(s): 189 – 193.
[Kan08]	16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP : *Audio-Visual Emotion Recognition Using an Emotion Space Concept*, Ittipan Kanluan, Michael Grimm, Kristian Kroschel
[Kap04]	A. Kapoor, R. W. Picard, and Y. IvanovProbabilistic combination of multiple modalities to detect interest. Proc. 17th Int. Conf. Pattern Recog., vol. 3, 2004, pp. 969-972.
[Kap05]	A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P.F. Driessen, “Gesture-Based Affective Computing on Motion Capture Data,” Proc. First Int’l Conf. Affective Computing and Intelligent Interaction, pp. 1-7, 2005
[Kap05]	A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P.F. Driessen,“Gesture-based affective computing on motion capture data,” In Proc First Int Conf on Aff Comp and Intelligent Interaction, pp. 1-7, 2005.
[Kap07]	A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. Int. Journal. Human-Computer Studies, vol. 65, no. 8, pp. 724-736.
[Kas06]	W lodzimierz Kasprzak and Piotr Skrzynski, “Hand image interpretation based on double active contour tracking”, Journal of Courses and Lectures-International Centre for Mechanical Sciences, pp. 439-446, 2006
[Kas88]	Michael Kass, Andrew Witkin and Demetri Terzopoulos, “Active Contour Models”, International Journal of Computer Vision, pp. 321-331, 1988.
[Kat07]	Katz, P., Singleton, M., Wicentowski, R. 2007. SWAT-MP : The Semeval-2007 Systems for Task 5 and Task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007), pp. 308-313, 2007.
[Ken95]	J. Kennedy and R. Eberhart, “Particle swarm optimization.” Proc. IEEE International Conf. on Neural Networks (Perth, Australia), IEEE Service Center, Piscataway, NJ, 1995 (in press)
[Kha10]	P. Khanna and M. Sasikumar. Recognising Emotions from Keyboard Stroke Pattern. International Journal of Computer Applications, 11(9), December 2010.
[Kie12]	Kiefer, P. 2012, "Mobile Intention Recognition ",
[Kim04]	Kim, S., Hovy, E. 2004. Determining the sentiment of opinions. In *Proceedings of International Conference on Computational Linguistics* (COLING-2004). 2004.
[Kim06]	Kim, S., Hovy, E. 2006. Identifying and Analyzing Judgment Opinions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (HLT-NAACL '06), pp. 200-207.
[Kim10]	Kim, S.M. Valitutti, A., Calvo, R.A. Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. In *Proceedings of the NAACL HTL 20120 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 62-70, Los Angeles California. June 2010.

[Kle05]	A. Kleinsmith, P. R. De Silva, & N. Bianchi-Berthouze, “Recognizing emotion from postures: Cross–cultural differences in user modeling,” in *Proc. the Conf. on User Modeling*, 2005, pp. 50–59
[Kle05]	A. Kleinsmith and N. Bianchi-Berthouze, “Grounding affective dimensions into posture features,” LNCS: Proc. 1st Int. Conference on Affective Computing and Intelligent Interaction, pp. 263-270, 2005.
[Kle11]	A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic Recognition of Non-Acted Affective Postures. IEEE Transactions on Systems, Man, and Cybernetics, Part B 41(4), 2011, 1027-1038.
[Kle12]	Kleinsmith, A.; Bianchi-Berthouze, N.; , "Affective Body Expression Perception and Recognition: A Survey," Affective Computing, IEEE Transactions on , vol.PP, no.99, pp.1, 2012
[Kol04]	Mathias Kolsch and Matthew Turk, “Fast 2d hand tracking with flocks of features and multi-cue integration,” in Proceedings of CVPR Workshop on Real-Time Vision for HCI, 2004.
[Kor01]	Kort, B., Reilly, R. Picard, R.W. 2001. An Affective Model of Interplay Between Emotions and Learning : Reengineering Pedagogy – Building a *Learning Companion*. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*. Madison.
[Kor65]	Kornhuber, H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkurbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Archive, 284, 1-17.
[Kov03]	J. Kovac, P. Peer and F. Solina, “2D versus 3D colour space face detection”, 4th EURASIP Conference on Video/Image Processing and Multimedia Communications, Croatia, pp. 449-454, 2003.
[Koz07]	Kozareva, Z., Navarro, B., Vazquez, S., Montoyo, A. 2007. UA-ZBSA: A Headline Emotion Classification Through Web Information. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007), 2007.
[Ku05]	H. van Kuilenburg, M. Wiering and M.J. den Uyl. A Model Based Method for Automatic Facial Expression Recognition. Proceedings of the 16th European Conference on Machine Learning, Porto, Portugal, 2005, pp. 194-205, Springer-Verlag GmbH.
[Ku08]	H. van Kuilenburg, M.J. den Uyl, M.L. Israël and P. Ivan. Advances in face and gesture analysis. Proceedings of Measuring Behavior 2008,26-29, 2008, pp. 371-372. Maastricht, The Netherlands, August
[Kun98]	Kang-Hyun Jo_, Yoshinori Kuno and Yoshiaki Shirai, “Manipulative Hand Gesture Recognition Using Task Knowledge for Human Computer Interaction”, Proceedings of the 3rd. International Conference on Face & Gesture Recognition, 1998.
[La04]	S. Langton, H. Honeyman, and E. Tessler, “The Influence of Head Contour and Nose Angle on the Perception of Eye-Gaze Direction,” Perception and Psychophysics, vol. 66, no. 5, pp. 752-771, 2004.
[La10]	9. O. Langner, R. Dotsch, G. Bijlstra, D.H.J. Wigboldus S.T. Hawk, and A. van Knippenberg (2010). Presentation and validation of the Radboud Faces Database. Cognition and Emotion, 24(8), 1377-1388.
[Lai06]	Wen-Hsiang Lai and Chang-Tsun Li, “Skin Colour-based Face Detection in Colour Images”, Proceedings of the IEEE International Conference on Video and Signal Based Surveillance, 2006.
[Lan02]	Lanitis, A., Taylor, C., Cootes, T.F.: Automatic interpretation and coding of face images using flexible models. IEEE Transactions on Pattern Analysis and Machine Intelligence,743–756 (2002).
[Lav09]	G. Lavee, E. Rivlin, and M. Rudzsky, “Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 5, pp. 489 –504, Sep. 2009.
[LC2012]	LC Technogies. (2012) The Eyegaze Communication System, The EyeFollower, 600 Series Eye Tracker, VA, http://www.eyegaze.com/
[Lee12]	Lee, H., Choi, Y.S., Lee, S. Park, I.P. Towards Unobtrusive Emotion Recognition for Affective Social Communication. In 9th Annual IEEE Consumer Communications and Networking Conference, IEEE Press (2012).
[Lee95]	J. Lee, T. Kunii, “Model Based Analysis of Hand Posture”, IEEE Computer Graphics and Applications.Vol.15, No. 5 pp. 77-86, 1995.
[Li01]	Li, S.; Fu, Q.; Gu, L.; Scholkopf, B.; Cheng, Y. & Zhang, H. “Kernel machine based learning for multi-view face detection and pose estimation” *Proceedings. Eighth IEEE International Conference on,* 2001, 2, 674 -679 vol.2

[Li08]	Bing Li, Student Member, IEEE, and Scott T. Acton, Senior Member, IEEE, ", Automatic Active Model Initialization via Poisson Inverse Gradient," IEEE Trans. On image processing, Vo. 17, No. 8, 2008.
[Li10]	Li, S., Lee, S.Y.M., Chen, Y., Huang, C., Zhou, G. 2010. Sentiment Classification and Polarity Shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING-2010). 2010.
[Lib82]	Libet, B., Wright, E. W., & Gleason, C. A. (1982). Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. Electroencephalography and clinical neurophysiology, 54(3), 322-35.
[Liu03]	Liu, H., Lieberman, H., Selker, T. 2003. A Model of Textual Affect Sensing using Real-World Knowledge. In *Proceedings of the 2003 International Conference on Intelligent User Interfaces* (IUI 2003). January 12-15. Miami, USA.
[Liu05]	C. Li, J. Liu, and M. D. Fox, "Segmentation of external force field for automatic initialization and splitting of snakes," Pattern Recognit., vol. 38, pp. 1947–1960, 2005.
[Liu07]	Liu, Y., Huang, X., An, A., Yu, X. 2007. ARSA : A Sentiment-Aware Model for Predicting Sales Performance Using Blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR 2007), pp. 607-614.
[Liu12]	Liu, B. 2012. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, May 2012.
[Loc02]	Raymond Lockton and Andrew W. Fitzgibbon, "Real-time gesture recognition using deterministic boosting", Preceedings of British Machine Vision Conference, pp. 817-826, 2002.
[Lu06]	Lu, C-Y., Hong, J-S., Cruz-Lara, S. 2006. Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques. In *Proceedings of the Third Taiwanese-French Conference on Information Technology* (TFIT 2006).
[Luy12]	Luyckx, K., Vaassen, F., Peersman, C., Daelemans, W. 2012. Fine-Grained Emotion Detection in Suicide Notes : A Thresholding Approach to Multi-Label Classification. In *Biomedical Informatics Insights (BII)*, Vol. 5, Issue 1, pp. 61-69. 2012.
[Majaranta2002]	Majaranta, P., and Räihä, K. (2002), Twenty Years of Eye Typing: Systems and Design Issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*. ETRA '02. ACM Press (2002), 15 – 22. (http://www.cs.uta.fi/ucit/ETRA2002-Majaranta-Raiha.pdf)
[Mam01]	James P. Mammen, Subhasis Chaudhuri, Tushar Agrawal. "Simultaneous Tracking of Both Hands by Estimation of Erroneous Observations". In Proceedings of BMVC'2001, 2001.
[Mar00]	Jérôme Martin and Jean-Baptiste Durand, "Automatic handwriting gestures recognition using Hidden Markov Models", Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000.
[Mar06]	*Multimodal Complex Emotions: Gesture Expressivity and Blended Facial Expressions*, JEAN-CLAUDE MARTIN, RADOSLAW NIEWIADOMSKI, LAURENCE DEVILLERS STEPHANIE BUISINE, CATHERINE PELACHAUD, International Journal of Humanoid Robotics Vol. 3, No. 3 (2006) 269–291, World Scientific Publishing Company
[May01]	R. E. Mayer, "Multimedia learning." Cambridge University Press., 2001.
[Med05]	Medler, D.A., Arnoldussen, A., Binder, J.R, Seidenberg, M.S. 2005. The Wisconsin Perceptual Attribute Ratings Database. 2005.
[Mer07]	A. Mehrabian, *Nonverbal Comm.* Aldine, 2007
[Met11]	Metaxas, P.T., Mustafaraj, E., Gayo-Avello, D. 2011. How (Not) To Predict Elections. In *Proceedings of Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE third international conference on Social Computing (socialcom)*.
[Mih07]	Mihalcea, R., Banea, C., Wiebe, J. 2007. Learning Multilingual Subjective Language via Cross-Lingual Projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, June 2007.
[Mil95]	Miller, G.A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41.
[Mish05]	Mishne, G. 2005. Experiments with Mood Classification in Blog Posts. In Proceedings of the first Workshop on Stylistic Analysis Of Text For Information Access. 2005.
[Moh10]	Mohammad, S.M., Turney, P., 2010. Emotions Evoked by Common Words and Phrases : Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HTL 20120 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 26-34, Los Angeles California. June 2010.

[Moi07]	Moilanen, K., Pulman, S. 2007. Sentiment Composition. In *Proceedings of Recent Advances in Natural Language Processing* (RANLP 2007). September 27-29, Borovets, Bulgaria. pp. 378-382.
[Mon93]	J. M. Montepare, and L. A. Zebrowitz, A cross-cultural comparison of impressions created by age-related variations in gait. Journal of Nonverbal Behavior, vol. 17, 1993, pp. 55-68.
[Mot03]	S. Mota and R.W. Picard, “Automated Posture Analysis for Detecting Learner’s Interest Level,” Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop, vol. 5, p. 49, 2003
[Mu07]	Murphy-Chutorian E.; Doshi A. & Trivedi M. “Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation” *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE,* 2007, 709-714
[Mu09]	Murphy-Chutorian, E. & Trivedi, M. M. “Head Pose Estimation in Computer Vision: A Survey” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* 2009, 31, 607-626
[Mu12]	Murad Al Haj, J. G. & Davis, L. S. “On Partial Least Squares in Head Pose Estimation: How to simultaneously deal with misalignment” *CVPR,* 2012
[Mul10]	Jörg Müller, Florian Alt, Daniel Michelis, and Albrecht Schmidt. 2010. Requirements and design space for interactive public displays. In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA, 1285-129
[Mur09]	E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell., 31 (4) (2009), pp. 607–626
[Nak10]	Nakagawa, T., Inui, K., Kurohashi, S. 2010. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. In *Proceedings of Human Language Technologies : the 2010 Annual Conference of the North American Chapter of the ACL* (HLT 2010), pp. 786-794. Los Angeles. 2010.
[Nam96]	Yanghee Nam and KwangYunWohn, “Recognition of Space-Time Hand-Gestures using Hidden Markov Model”, In Proceedings of ACM Symposium on Virtual Reality Software and Technology, 1996.
[Nas98]	El-Nasr, M. S., Ioerger, T. R. and Yen, J. Learning and Emotional Intelligence in Agents. In: Proceedings of AAAI (American Association for Intelligence) Fall Symposium on Emotional Intelligence, Floride. 1998Artificial
[Nav09]	Navigli, R. 2009. Word Sense Disambiguation : A Survey. In *ACM Computing Surveys (CSUR)*, Vol. 41 Issue 2, February 2009.
[Nef99]	A. Nefian, M. Hayes, *Face recognition using an embedded HMM, in: IEEE Conf. on Audio and Video-based Biometric Person Authentication*, 1999, pp. 19–24
[Ngu05]	N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, “Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 2, pp. 955 – 960 vol. 2.
[Nic11]	IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. , NO. , MONTH 2011 “Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence–Arousal Space”, Mihalis A. Nicolaou *Student Member, IEEE*, Hatice Gunes, *Member, IEEE* and Maja Pantic, *Senior Member, IEEE*
[Och05]	*Intelligent Expressions of Emotions*. Magalie Ochs, Radoslaw Niewiadomski, Catherine Pelachaud, and David Sadek. The First International Conference on Affective Computing & Intelligent Interaction (ACII 2005), p. 707-714, 2005, Pékin, Chine
[Oct11]	J.R. Octavia, K. Coninx, and K. Luyten, “Squeeze me and I'll change: An exploration of frustration-triggered adaptation for multimodal interaction”. In 2011 IEEE Symposium on 3D User Interfaces (3DUI), pages 79-86, march 2011.
[Oli00]	N. Oliver, A. Pentland, F. Berard, LAFTER: *a real-time face and lips tracker with facial expression recognition*, Pattern Recognition 33 (2000) 1369–1382
[Ong04]	Eng-Jon Ong and Richard Bowden, “A Boosted Classifier Tree for Hand Shape Detection”, Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition, 2004.
[Orman2011]	Zeynep Orman1, Abdulkadir Battal2 and Erdem Kemer3, a study on face, eye detection and gaze estimation, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.2, No.3, August 2011.
[Ort88]	Ortony, A., Clore, G. L. and A., C. *The cognitive structure of emotions*, Cambridge University Press, 1988.

[Ots97]	T. Otsuka, J. Ohya, *Recognizing multiple persons facial expressions using HMM based on automatic extraction of significant frames from image sequences*, in: Proc. Internat. Conf. on Image Processing (ICIP97), 1997, pp. 546–549
[Pa05]	Pan, Y.; Zhu, H. & Ji, R. “3-D Head Pose Estimation for Monocular Image” *Springer,* 2005, 293-301
[Pak10]	Pak, A., Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC 2010).Valletta, Malta. 2010.
[Pak12]	Pak, A., Bernhard, D., Paroubek, P., Grouin, C. 2012. A Combined Approach to Emotion Detection in Suicide Notes. In *Biomedical Informatics Insights (BII)*, Vol. 5, Issue 1, pp. 105-114. 2012.
[Pan02]	Pang, B., Lee, L., Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86, Philadelphia, PA.
[Pan04]	Pang, B., Lee, L. 2004. A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings *of the Annual Meeting of the Association of Computational Linguistics*, pages 271–278.
[Pan04]	M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. IEEE Transactions on Systems, Man and Cybernetics, 34(3):1449–1461, 2004.
[Pan08]	Pang, B., Lee, L. 2008. Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval* 2(1-2), pp. 1–135, 2008.
[Pan2004]	Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L.A., Feusner, M. K., and Newman, J. K, 2004. The Determinants of Web Page Viewing Behavior: An Eye-Tracking Study. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*. ETRA '04. ACM Press (2004), 147 – 154. (http://panb.people.cofc.edu/pan/ETRA04.pdf)
[Pat01]	H.M. Paterson, F.E. Pollick, and A.J. Sanford, “The role of velocity in affect discrimination,” Proc. 23rd Annual Conference of the Cognitive Science Society, pp. 756-761, Lawrence Erlbaum Associates, 2001.
[Pat99]	Paterno, Fabio, “Model-Based Design and Evaluation of Interactive Applications”, Springer Verlag, Lodon, UK, 1999, 192 p.
[Pav97]	Vladimir I. Pavlovic, Rajeev Sharma, and Thomas S. Huang, “Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review”, IEEE Transactions on pattern analysis and machine intelligence, pp. 677-695, Vol. 19, No. 7, JULY 1997.
[Pei09]	Pereira, L.M., Anh, H.T. (2009). Elder care via intention recognition and evolution prospection, in: S. Abreu, D. Seipel (eds.), Procs. 18th International Conference on Applications of Declarative Programming and Knowledge Management (INAP'09), Évora, Portugal, November 2009.
[Pen94]	A. Pentland, B. Moghaddam, and T. Starner. *View-based and modular eigenspaces for face recognition*. In *Computer Vision and Pattern Recognition Conference*, pages 84–91. IEEE Computer Society, 1994.
[Pes12]	Pestian, J.P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Bretonnel Cohen, K., Hurdle, J., Brew, C. 2012. Sentiment Analysis of Suicide Notes: A Shared Task. In *Biomedical Informatics Insights (BII)*, Vol. 5, Issue 1, pp. 3-16. 2012.
[Pic97]	Picard, Rosalind W., and Jennifer Healey. "Affective wearables." Personal and Ubiquitous Computing 1, no. 4 (1997): 231-240.
[Pol04]	Polanyi, L., Zaenen, A. 2004. Contextual valence shifters. In J. Shanahan, Y. Qu, and J. Wiebe (eds.), *Computing Attitude and Affect in Text: Theory and Applications*, pp. 1–9. The Information Retrieval Series, Vol. 20, Springer, Dordrecht, The Netherlands.
[Poole2005]	Alex Poole and Linden J. Ball (2005), Psychology Department (, Lancaster University, UK, Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects.
[Pre02]	Prendinger, H., Descamps, S. and Ishizuka, M. “Scriptinag Affective Communicationwith Life-like Characters in Web-based Interaction Systems”. Applied Artificial Intelligence Journal, 16(7-8), 519-553.2002.
[Qua90]	D.L. Quam, “Gesture Recognition With a DataGlove”, Proc. IEEE National Aerospace and Electronics Conf., pp. 755-660, 1990.
[Ra08]	Ranganathan, A. & Yang, M.-H. “Online sparse matrix gaussian process regression and vision applications” *ECCV,* 2008
[Rab89]	L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 –286, Feb. 1989.
[Ram03]	Aditya Ramamoorthya, Namrata Vaswania, Santanu Chaudhurya and Subhashis Banerjeeb, “Recognition of dynamic hand gestures”, Journal of Pattern Recognition Society, pp. 2069-2081, 2003.

[Ras09]	Omer Rashid, Ayoub Al-Hamadi, Axel Panning and Bernd Michaelis, “Posture Recognition using Combined Statistical and Geometrical Feature Vectors based on SVM”, International Journal of Information and Mathematical Sciences, pp. 590-597, 2009.
[RAUDO2012]	Vidas RAUDONIS, Agne PAULAUSKAIT E-TARASEVICIENE, Laura KIŽAUSKIEN (2011), The Gaze Tracking System with Natural Head Motion Compensation, INFORMATICA, 2012, Vol. 23, No. 1, 105–124.
[Rea05]	Read, J. 2005. Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics*. Ann Arbor, US. 2005.
[Rea09]	Read, J., Caroll, J. 2009. Weakly Supervised Techniques for Domain-Independent Sentiment Classification. In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion* (2009), pp. 45-52. 2009.
[Ric00]	J.S. Richman and J.R. Moorman. ”Physiological time-series analysis using approximate entropy and sample entropy”. *American Journal of Physiology- Heart and Circulatory Physiology*, 278(6):H2039, 2000.
[Ril03]	Riloff, E., Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of* the *2003 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2003), pp. 105–112.
[Ros03]	Rosis, F. d., Pelachaud, C., Poggi, I., Carofiglio, V. and Carolis, B. D. “From Greta's mind to her face: modelling the dynamics of affective Studies, 59(1-2), 81-118.2003.states in a conversational embodied agent”. *International Journal of Human-Computer*
[Sad97]	Sadek, D., Bretier, P. and Panaget, F. ARTIMIS: Natural Dialogue Meets Rational Agency. In: *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI'97), Nagoya, Japon, 1030-1035. 1997*
[Sah07]	Sahlgren, M., Karlgren, J., Eriksson, G. 2007. SICS: Valence Annotation Based on Seeds in Word Space. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007).
[San05]	2005 Special Issue: “A systems approach to appraisal mechanisms in emotion” David Sander, Didier Grandjean, Klaus R. Scherer Geneva Emotion Research Group, Department of Psychology, University of Geneva, 40, Bd. du Pont d’Arve, CH-1205, Geneva, Switzerland. Received 23 March 2005; accepted 24 March 2005
[Sav11]	N. Savva, N. Bianchi-Berthouze, “Automatic recognition of affective body movement in a video game scenario,” International on Intelligent Technologies for interactive entertainment, 2011 Conference
[Sav12]	Savva, N.; Scarinzi, A.; Bianchi-Berthouze, N.; , "Continuous Recognition of Player's Affective Body Expression as Dynamic Quality of Aesthetic Experience," Computational Intelligence and AI in Games, IEEE Transactions on , vol.4, no.3, pp.199-212, Sept. 2012
[Sch05]	Schlieder, Christoph, 2005, GeoSpatial Semantics, Representing the meaing of spatial behaviour by spatially grounded intentional systems
[Sch73]	A. Scheflen, Communicational structure: Analysis of a psychotherapy transaction. Bloomington: University of Indiana Press.
[Sch78]	Schmidt, C., Sridharan, N., Goodson, J. (1978). The plan recognition problem: an intersection of psychology and artificial intelligence. Artificial Intelligence, Vol. 11, 1978, 45-83.
[Sch94]	Scherer, K., Wallbott, H. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66, pp. 310-328.
[Sch94]	B. Schilit, N. Adams, and R. Want. (1994). "Context-aware computing applications" (PDF). *IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'94), Santa Cruz, CA, US*. pp. 89–101.
[Sch97]	M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Trans. on Signal Processing*, vol. 45, pp. 2673–2681, November 1997
[Seb04]	June 18, 2004 16:2 WSPC for Review Volume emotion, CHAPTER 1: MULTIMODAL EMOTION RECOGNITION, Nicu Sebe, Ira Cohen, and Thomas S. Huang
[Sha04]	Caifeng Shan, Yucheng Wei, Tieniu Tan and Frédéric Ojardias, “Real Time Hand Tracking by Combining Particle Filtering and Mean Shift”, Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.
[Sha07]	Noor Shaker and M. Abou Zliekha, “Real-time Finger Tracking for Interaction”, Proceedings of the 5th International Symposium on image and Signal Processing and Analysis, pp. 141-145, 2007.

[Sha07]	Shaikh, M., Prendinger, H., Mitsuru, I. 2007. Assessing Sentiment of Text by Semantic Dependency and Contextual Valence Analysis. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction* (ACII '07), pp. 191 – 202.
[Sil06]	De Silva, P.R.; Osano, M.; Marasinghe, A.; Madurapperuma, A.P.; , "Towards recognizing emotion with affective dimensions through body gestures," Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on , vol., no., pp.269-274, 2-6 April 2006
[Sin02]	Singh, P. 2002. The Public Acquisition of Commonsense Knowledge. In *Proceedings of AAAI Spring Symposium*. Palo Alto, CA. 2002.
[smivision2012]	Available from: http://www.smivision.com/en/gaze-and-eye-tracking-systems
[Soc11]	Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2011), pp 151-161. Edinburgh, 2011.
[Sta95]	Thad Starner and Alex Pentland, “Real-Time American Sign Language Recognit ion from Video Using Hidden Markov Models”, Proceedings of the International Symposium on Computer Vision, pp. 265-270, 1995.
[Sta95]	T. Starner and A. Pentland, “Real-time American Sign Language recognition from video using hidden Markov models,” in *Computer Vision, 1995. Proceedings., International Symposium on*, 1995, pp. 265 –270.
[Ste01]	B. Stenger, P. R. S. Mendonc¸a and R. Cipolla, “Model-Based Hand Tracking Using an Unscented Kalman Filter”, In Proc. British Machine Vision Conference, 2001.
[Sto 07]	J. Van den Stock, R. Righart, and B. de Gelder, “Body expressions influence recognition of emotions in the face and voice,” Emotion,vol. 7, no. 3, pp. 487-494, 2007.
[Sto66]	Stone, P. J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press. 1966.
[Str04]	Strapparava, C., Valitutti, A. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of the Language Resources and Evaluation Conference* (LREC 2004), Lisbon, May 2004, pp. 1083-1086.
[Str07]	Strapparava, C., Mihalcea, R. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations* (SemEval 2007), Prague, Czech Republic, June 2007.
[Str08]	Strapparava, C., Mihalcea, R. 2008. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM symposium on Applied computing* (SAC 2008), pp. 1556-1560.
[Sub95]	Subutai Ahmad, “A Usable Real-Time 3D Hand Tracker”, IEEE, 28th Asilomar Conference on Signals, Systems and Computers, pp1257-1261, 1995.
[Sum04]	S. Sumec, “Multi camera automatic video editing,” presented at the in Proceedings of the ICCVG., 2004.
[Sun12]	Jaeyong Sung, Colin Ponce, Bart Selman, Ashutosh Unstructured Human Activity Detection from RGBD Images, Saxena. In *International Conference on Robotics and Automation (ICRA)*, 2012
[Syk03]	Jonathan Sykes, and Simon Brown, “Affective gaming: measuring emotion through the gamepad”, In Extended abstracts CHI '03, pages 732-733, New York, USA, 2003. ACM.
[Tab11]	Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, Vol. 37, No. 2, pp 267-307. 2011.
[Tao98]	H. Tao, T.S. Huang, Connected vibrations: a modal analysis approach to non-rigid motion tracking, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR98), 1998, pp. 735–740
[Tobii2012]	Tobii Techonoly. (2012) Tobii eye trackers, information available at http://www.tobii.com/
[Tri96]	Triesch, J., Malsburg, C., “Robust Classification of Hand Postures Against Complex Back-ground”, Intl Conf. On Automatic Face and Gesture Recognition, 1996.
[Tur02]	Turney, P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL 2002), pp. 417–424, Philadelphia, PA.

[Tur04]	Mathias K¨olsch and Matthew Turk, “Analysis of Rotational Robustness of Hand Detection with a Viola-Jones Detector”, Proceedings of the 17th International Conference on Pattern Recognition, 2004.
[Ued03]	E. Ueda and al. , “A Hand Pose Estimation for Vision Based Human Interfaces”, IEEE Transactions on Industrial Electronics, Vol. 50, No. 4, pp. 676–684, 2003.
[Val04] [Val11]	M. Valstar, M. Pantic, and I. Patras, “Motion History for Facial Action Detection from Face Video,” Proc. IEEE Int’l Conf. Systems, Man, and Cybernetics (SMC ’04), vol. 1, pp. 635-640, 2004. Valstar, M.F.; Bihan Jiang; Mehu, M.; Pantic, M.; Scherer, K.; , "The first facial expression recognition and analysis challenge," Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on , vol., no., pp.921-926, 21-25 March 2011
[Van08]	W.M. van den Hoogen, W.A. IJsselsteijn, and Y.A.W. de Kort, “Exploring behavioral expressions of player experience in digital games”. In A. Nijholt and R. Poppe, editors, Workshop proceedings on Facial and Bodily Expression for Control and Adaptation of Games, pages 11-19, 2008.
[Verte2002]	Vertegaal, R., Dickie, C., Sohn, C.& Flickner, M. (2002). Designing attentive cell phone using eyecontact sensors. In CHI ’02 Extended Abstract on Human Factors in Computing Systems. ACM Press, 647-647
[Vi04]	P. Viola and M. Jones, 2004. Robust Real-time Face Detection. International Journal of Computer Vision 57(2), 137–154, 2004.
[Villa2008]	Arantxa Villanueva, Juan J. Cerrolaza and Rafael Cabeza (2008). Geometry Issues of Gaze Estimation, Advances in Human Computer Interaction, Shane Pinder (Ed.), ISBN: 978-953-7619-15-2, InTech, Available from: http://www.intechopen.com/books/advances_in_human_computer_interaction/geometry_issues_of_gaze_es
[Wa07]	Wang, J.-G. & Sung, E. “EM enhancement of 3D head pose estimated by point at infinity” *Image Vision Comput., Butterworth-Heinemann,* 2007, 25, 1864-1874
[Wal86]	D. Janssen, W. I. Schollhorn, and J. Lubienetzki. Recognition of emotions in gait patterns by means of artificial neural nets. Journal of Nonverbal Behavior, vol. 32, 2008, pp. 79-92.
[Whi06]	J. Whitehill and C.W. Omlin, “Haar Features for FACS AU Recognition,” Proc. IEEE Int’l Conf. Automatic Face and Gesture Recognition (AFGR ’06), pp. 217-222, 2006.
[Wie10]	Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A. 2010. A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp 60-68. Uppsala, July 2010.
[Wil05]	Wilson, T., Wiebe, J., Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technologies Conference / Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP 2005), Vancouver, Canada.
[Wil09]	Wilson, T., Wiebe, J., Hoffmann, P. 2009. Recognizing Contextual Polarity : An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, Vol. 35, No. 3, pp 399-433. 2009.
[Wo24] [Wol08]	W.H. Wollaston, “On the Apparent Direction of Eyes in a Portrait,” Philosophical Trans. Royal Soc. of London, vol. 114, pp. 247-256, 1824. M. Wollmer, and F. Eyben, and S. Reiter, and B. Schuller, and C. Cox, and E. Douglas-Cowie, and R. Cowie , “Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies,” in *Proc. of 9th Interspeech Conf.*, 2008, pp. 597–600
[Xi10]	Xiangyang Liu, Hongtao Lu and Wenbin L. “Multi-Manifold Modeling for Head Pose Estimation” *ICIP,* 2010
[Xu98]	C. Xu and J. L. Prince, “Snakes, shapes, and gradient vector flow,” IEEE Trans. Image Process., vol. 7, no. 3, pp. 359–369, 1998.
[Yac96]	Y. Yacoob, L.S. Davis, *Recognizing human facial expressions from long image sequences using optical flow*, IEEE Trans. Pattern Anal. Machine Intell. 18 (6) (1996) 636–642
[Yac96]	Yacoob, Y.; Davis, L.S.; , "Recognizing human facial expressions from long image sequences using optical flow," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.18, no.6, pp.636-642, Jun 1996
[Yam92]	J. Yamato, J. Ohya, and K. Ishii, “Recognizing Human Action in Time Sequential Images Using Hidden Markov Model,” 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 379-385, 1992.

[Yam92]	J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden Markov model,” in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR ’92., 1992 IEEE Computer Society Conference on*, 1992, pp. 379 –385.
[Yan99] [Yil06]	Ming-Hsuan Yang and Narendra Ahuja, “Recognizing Hand Gesture Using Motion Trajectories”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition A. Yilmaz, O. Javed and M. Shah, “Object tracking: A survey,” ACM Comput. Surv. 38, 2006.
[Yin01] [Yu03]	Y. Wu, L. J. Y., and T. S. Huang. “Capturing natural hand Articulation”, In Proc. 8th Int. Conf. on Computer Vision, volume II, pp. 426–432, Vancouver, Canada, July 2001. Yu, H., Hatzivassiloglou, V. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2003)
[Zh07] [Zha07]	Zhang Z.; Hu Y.; Liu M. & Huang T. “Head Pose Estimation in Seminar Room Using Multi View Face Detectors” *Multimodal Technologies for Perception of Humans, Springer Berlin Heidelberg,* 2007, 4122, 299-304 Guoying Zhao; Pietikainen, M.; , "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.29, no.6, pp.915-928, June 2007, doi: 10.1109/TPAMI.2007.1110.
[Zhu 2007]	Zhiwei Zhu and Qiang Ji (2007), Novel Eye Gaze Tracking Techniques Under Natural Head Movement, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 54, NO. 12, DECEMBER 2007
[Zhu00]	X. Zhu, J. Yang, and A. Waibel. “Segmenting Hands of Arbitrary Color”. In Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition, pp.446-453, 2000.
[Zil09]	Ying Zilu; Zhang Guoyi; , "Facial Expression Recognition Based on NMF and SVM," Information Technology and Applications, 2009. IFITA '09. International Forum on , vol.3, no., pp.612-615, 15-17 May 2009, doi: 10.1109/IFITA.2009.279.
[Zocco2010]	Davide Zoccolan, Brett J. Graham1 and David D. Cox, 2010, A self-calibrating, camera-based eye tracker for the recording of rodent eye movements, Frontiers in Neuroscience, November 2010, Vol 4, Article 193

9. Annex 1: Available technologies in the “Empathic Products” project

Available Affective Technologies		
Partner	Technology	Description
FADO	Avatar Facial Expressions (feedback)	Can be stylized (drawn mouth and eyebrows) or more realistic (3d or image).
FADO	Avatar Corporal Expressions (feedback)	Actions like dancing or being still for long periods of time.
FADO	Avatar Color Settings (feedback)	Parts or the whole avatar turns blue for sadness or red for excitement and asking to intervene.
FADO	Avatar Video with Face Detection (feedback)	Overlapping face centered video on the avatar head (paralelipiped).
Tecnalia	physiological-based emotion monitoring system	smarphone app to monitor, record and visualization emotional valence
Inabensa	User activity detection and analysis	A network of low-consumption sensors to capture information regarding user location (based on a RSSI technique) and user activity (such as presence in bed and chairs) for further processing in an application server
Inabensa	User intention prediction	Server application intelligence to estimate user intention based on a model that combines information regarding its activity and context (hour of the day…)
ALBLF	Jazz Analyzer	Speech analyzer / Prosody analysis
ALBLF	Gesture Analyzer	Video analyzer enabling hand gesture posture recognition (matlab)
ALBLF	Attention measurement	Video analyzer enabling to measure basic visual attention characteristic (webcan/python/opencv)
ALBLF	Video Orchestration	Video engine enabling to orchestrate video
ALBLF	Video conference Client/ webRTC/HTML5	Client enabling to render/visualize the ALBLF vision conferencing system
ALBLF	Metadata aggregator	Metadata server storing results of analyzers and supporting push, pull, subscription mechanisms
VV	FaceReader 5.0	Software automatically analyzing facial expressions of emotions
Lille 1	Facial Expression recognition	Frame-based detection of facial expressions (happy, surprise, angry, sad, normal)
Lille 1	Facial Drowsyness detector	Frame-based detection of eye blinks and PERCLOS based drowsiness evaluation
Lille 1	Head orientation detector	Frame-based detection of discrete (7 classes) yaw orientation of the head.
VTT	People Tracker and people counter	Analyses depth sensor data by locating people from the sensor data and following the movement of the people
VTT	Activity and intention of spectator/user of DS	Can detected and analyse behavior of the people front of the digital signage, Under Development, not ready
VTT	Keystroke dynamics	Can analyse people emotions/excitment from

		keystroke, under evelopment, not ready
VTT	Gaze detection	To detect gaze from short range available, determing gaze direction from longer distance under development
Noldus	The Observer XT – release 11	a software tool to analyze who does what, where, when and how/to whom as a human observes it (available)
Noldus	uLog – release 3	event logger tool for automatic logging of user behavior (version 2 is already available, 3 not yet)
Noldus	Media Recorder – release 2.0	an easy-to-use tool for a stable way of recording up to four video streams simultaneously (available)
LORIA	Textual Sentiment Analysis	Finding automatically the polarity (positive, negative or neutral) of a natural language sentence or a document
LORIA	Textual Emotion Detection	Finding automatically a basic emotion (joy, sadness, fear, anger, trust, disgust, neutral) of a natural language sentence or a document
SKSW	iisu 3.5.1 3D gesture recognition SDK	hand, body pose, skeleton 3D analysis from a Depthsense Time-of-flight camera
SKSW	ethics questionnaire	questionnaire to review privacy and ethics of a product for user tests
IMT	Gesture Analyzer	motion-based gesture recognition
IMT	affective avatars	MPEG-4 3D animation techniques
CTP	Hotspot dynamic detection	Determines physical location indoors and best possible connection(s) (wrt/ reliability/cost/security)
CTP	Multipath Routing & Adaptative VPN	Allows to use multiple simultaneous connections to get the best reliability and performances
CTP	GPS deduced relative location/speed	Use the GPS functionality to deduce the device speed/location (Train, Car, Walking, …)
CTP	jsondb	Quick portable JSON database for inter-Empathic application communitations