A NEW PROBABILISTIC METHODOLOGY TO SUPPORT AN EMOTIVE DIALOG BETWEEN A HUMAN AND A ROBOT

José Augusto Soares Prado

PHD THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL

AND COMPUTER ENGINEERING

OF THE UNIVERSITY OF COIMBRA

FACULTY OF SCIENCE AND TECHNOLOGY

UNIVERSITY OF COIMBRA

April, 2012

I hereby declare that this thesis is the result of my own research except as cited in the references. This thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :
Student : José Augusto Soares Prado

For my beloved childreen **Gustavo Prado** and **Letícia Prado.**

Praise to the Almighty

Acknowledgment

I gratefully acknowledge support from Institute of Systems and Robotics at University of Coimbra (ISR-UC) and from Portuguese Foundation for Science and Technology (FCT).
I would like to thank my supervisor *Professor Doctor Jorge Manuel Miranda Dias* for all the support, guidance, kind comments, insightful critique, detailed and thorough corrections, and suggestions on how to improve the work during all these years.
I thank also to the members of the jury, for the time and dedication on the evaluation of this thesis.
I appreciate the support given from my family: specially from my mother Maria das Graças Neto Soares and from my father José Gouvêa Prado.
I thank also to my wife Patrícia Nóbrega, for all the support that she gave me. Also for all the nights (and days) that she had to stand alone with the baby while I was working.
Thanks to all the friends who accompanied me on this journey, mostly my colleagues from ISR: Amilcar Ferreira, Carlos Simplício, David Portugal, Diego Faria, Hadi Aliakbarpour, Hugo Faria, Jafar Hosseini, João Filipe Ferreira, José Marinho, Kamrad Khoshhal, Luis Almeida, Luis Davim, Luis Santos, Pedro Trindade, Professor Jorge Lobo, Professor Paulo Menezes, Professor Rui Rocha, Tiago Margalho, Ricardo Martins, Rita Catarino. It was risky to make this list, because it is highly probable that I forget to mention some names. If that's your case, thank you for your support.

José Augusto Soares Prado, Coimbra, 2012

Abstract

In this thesis, Bayesian approaches to analyze and synthesize human facial and vocal expressions are proposed, moreover a robot is used to influence the decision of the robot's response during a human robot interaction. In a human to human conversation, a person senses the interlocutor's face and voice, perceives his/her emotional expressions, and processes this information in order to decide which response to give. Moreover, observed emotions are taken into account, and the response may vary according to the defined for the robot. The can be understand as a personality of the robot. An emotional vector is prepared as input for this personality and carries probabilistic influences about which emotion the robot shall use in its behavior. The purpose of our structure is to endow robots with the capability of: not only recognize human emotions, but also to synthesize emotions and simulate personality. Thus, several sub problems need to be solved: feature extraction, classification, decision and synthesis. In the proposed approach, we integrate two novel Bayesian classifiers for emotion recognition from audio and video channels. Then, we use a new method for fusion with the selected *social behavior profile*. To keep the person engaged in the interaction, after each iteration of the analysis, the robot synthesizes voice with lips synchronization and also synthesizes facial expressions. The *social behavior profile* conducts the personality of the robot. The structure and the work flow of the: analysis, synthesis and decision; are addressed, and the Bayesian networks are discussed. Furthermore, we proposed assessments for measuring the engagement of the human during the interaction with the robot. Finally, we test the system with a dialog study case and by using the assessments we were able to reinforce the importance of such an emotional component for social robot applications.

Resumo

Nesta tese, abordagens Bayesianas para analisar e sintetizar expressões humanas (faciais e vocais) são propostas. Além disso, perfis de comportamento social foram integrados no sistema de forma a emular personalidade no robot. Em uma conversa de humano para humano, uma pessoa analisa o rosto do interlocutor e também a voz, percebendo assim as suas expressões emocionais; a pessoa então processa essa informação para decidir que resposta dar. As emoções observadas são então levadas em consideração, e com isso a resposta pode variar. No nosso sistema, a resposta também poderá variar, não só com base nas expressões, mas também de acordo com o perfil de comportamento social definido para o robot. O perfil de comportamento social pode ser entendido como uma personalidade do robot. Um vector emocional é preparado como entrada para esta personalidade, este vector carrega influências probabilísticas sobre as quais o robot decide qual emoção utilizará na resposta, dependendo da personalidade seleccionada. O propósito de nossa estrutura é dotar robôs com a capacidade de: não só reconhecer as emoções humanas, mas também de sintetizar emoções e simular personalidade. Contudo, vários subproblemas precisaram ser resolvidos, dentre eles: a extracção de características da imagem e do som, a classificação, a decisão e a síntese. A abordagem proposta integra dois novos classificadores Bayesianos para reconhecimento de emoções a partir de canais de áudio e de vídeo. Nós definimos e utilizamos um novo modelo de mistura Bayesiana para a fusão de ambos os canais com a personalidade. Para manter a pessoa envolvida na interacção, após cada interacção da análise, o robot sintetiza a voz humana tanto com sincronização de lábios quanto com expressões faciais. O perfil de comportamento social conduz a personalidade do robot. A estrutura e fluxo de trabalho da análise, síntese e decisão são abordados, e as redes Bayesianas são definidas e discutidas. Além disso, propusemos avaliações para medir o engajamento do ser humano durante a interacção com o robot. Finalmente, testamos o sistema em um estudo de caso de diálogo, e usando as avaliações definidas, fomos capazes de reforçar a importância de tal componente emocional para aplicações robóticas sociais.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Human regular dialogue is generally contactless. Thus, to reduce the separation between humans and machines contactless interfaces can be used. Definition of human regular dialogue is beyond spoken communication. In a face to face interaction between humans, several modalities are typically used, for example: body posture, gestures, gaze, vocalization, and facial expressions. Our focus is to improve the interaction between human and machine by exploring the non-verbal cues, namely facial and vocal expressions. In this area, researchers commonly use the term "verbal" with the strict sense meaning of: "concerned with words". Mostly the term "verbal communication" is not used as a synonym for oral communication. To understand better, take as example a grunt, a singing melody or a wordless note; these are vocal sounds that are not words, so they are nonverbal. Thus, the analysis of voice without regard to the words is called nonverbal.

There are several contributions in this study, first we do an extensive study about how humans deal with emotions and feelings. Then we present how a robotic system can address the same issues. Bayesian real time classifiers of emotions from audio and video are proposed, then the paper goes through the fusion of both modalities with a proposed Bayesian mixture model. Later synthesis is presented with an avatar capable of lips synchronization during speech. Moreover different behaviors of the robot are explored and the reaction of the users are tested.

From the *visual perception* our focus is on facial expression recognition while in *auditory perception* we focus on vocal expression recognition. The *emotion recognition* problem, in both modalities, depends on two sub-problems: the sensory processing that does the extraction of features from the input signal, and the classification of these features across a defined scope. After the classification, the robot reacts according to the human recognized emotion and also according to a robotic *social behavior profile* (SBP), see figure 1.1. Moreover, the interaction is supported by synthesizing the vocal and facial expressions, and also lips synchronization. For both analysis and synthesis, a Bayesian framework is used.

Humans express their emotional states through paralinguistic cues, e.g., gestures, gaze, facial expressions, movements or body pose. Among all, our main work focuses specifically on human facial expressions and vocal expressions. Emotion recognition systems have many potential applications. They can be used in medicine or psychology applications, surveillance or in intelligent human-machine interaction. The components proposed in this thesis can be incorporated in a companion robot, a service robot, a social robot, or even a chat robot. The robot observes and reacts according to the facial and vocal expressions of a person. This robot can be used in the context of assisted ambiance. The global motivation for this project addresses the emergent tendencies of developing new devices to the elderly community.

1.2 Theoretical Background

The emotional state that a person demonstrates as a reaction to some circumstance depends on the *social behavior profile* of the subject. According to Kau et al. [40], each Social behavior profile (SBP) is distinct by specific features of social behavior impairment. The specification of it's scope is context dependent. For example, Evers in [24] argued that both sympathy and antipathy can but do not need to be empathic, along [24] both antipathetic and sympathetic are considered as *social behavior profile*s. Moreover in [27] an attempt to do automatic analysis of learner's social behavior during computer-mediated synchronous conversations was presented. Four SBPs were analyzed in this work: moderator, valuator, seeker, interdependent. In medicine, for example on [40], we

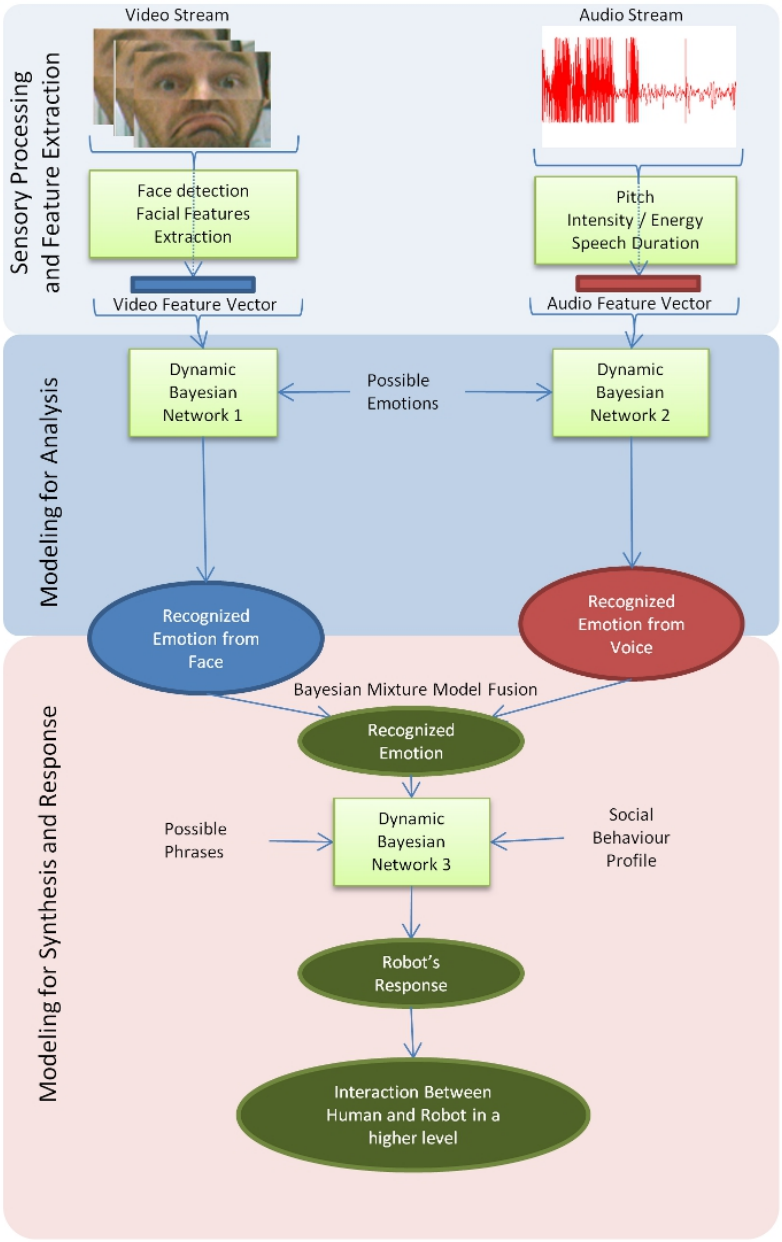


Figure 1.1: Diagram from the video and audio input streams up to robotic response. After the whole process, the result is a higher level of interaction. A higher level of interaction means that the human enjoy more to interact with an emotive robot than with a robot without emotions. Assessments for this are proposed along this thesis. Novel Bayesian classifiers for facial and vocal expressions are also proposed. A Dynamic Bayesian Network represented as DBN2 is proposed for the auditory perception; its outputs are probabilities of vocal expressions. Dynamic Bayesian network1 is proposed for visual perception, in this network certain local distortions presented over the human face are the input evidences to infer the person's facial expression. Later the information from both networks are combined with a robotic personality also called as social behavior profile *(SBP),* this happens on the third proposed DBN. The input vector for DBN3 is called *emotional vector* and the DBN3 is responsible for fusion and also decision of the robot response.

found autistic, apathetic and aggressive as *social behavior profiles*. Several other *social behavior profiles* are listed in the literature, they may vary a lot depending on the author's interpretation and on the context of each problem. For our context, we selected five *social behavior profiles* for our robot: *neurotic, extroverted, conscientious, agreeable and humorous*.

Along the history of human kind, emotions have always been considered important; especially by their role in social behavior. Sometimes they were seen as elevating, and in some other times as being degrading. But until recently it was hard to include them in the field of science, and even harder to include them in the realm of technology. However, we can start to address these questions now because today man has a workable idea about what emotions are and that is an initial step in the attempt to discover *why* emotions are and *what* emotions do for us. Furthermore, we know that emotions play a critical role in social behavior[16, 15].

During the seventeenth century, Spinozza [90] worked on an attempt to define what human emotions are. William James continued this work, and more recently this research area was also addressed by Damasio [16, 15]. But despite all the advancements in neuroscience's understanding of what emotions are, all the work done pointed away from the possibility of mathematically representing emotions. In order to include emotions in technology, it is necessary to find a mathematical framework for emotions.

In neuroscience, as defined in [52], emotions are occurrences in our body (including facial expressions), the feelings are the neuronal representations of such emotions. We can see people´s emotions, but not their feelings. In most circumstances emotions can generate feelings, but not the other way around. What feelings often elicit is the occurrence of a simulation of an emotion induced by a feeling, an as-if emotion. However, it is possible to make use of the person's expressed emotion to guess the person's implied feeling. So what the robot will do is not to have an emotion, but rather learn the person's implied feeling and how that feeling should influence the robot's decision-making.

Spinoza et al. [90] assumed that the past was enough to determine the future, while Leibniz assumed that the universe had a certain irreducibility to it which enable it to have multiple future-possibilities given a certain past. In situations where one does not have the needed information about the past, the use of Leibniz's perspective is often a

better approach than Newton's or Spinoza's. According to Lori et al. in [51], although Damasio's model of emotions was found to be a good match to Spinoza's perspective [16], it might even be a better match to Leibniz's perspective.

Newton defined that one cause leads to one consequence, and every consequence is unequivocally and completely defined by its cause. Leibniz had a more freedom-occurring approach to causality, believing that the cause limits the possible consequences, however Leibniz did not limit it to only one consequence. Lori [52] explored the point of view of Leibniz using the approach to Leibniz's causality defended in [9] and proposed that the word "causality" be replaced by the concept of "enablement", and the word "consequence" be replaced by the concept of "alternatives". The enablement does not have the capacity to univocally define a single alternative because the amount of information, the "message", coming into a system is not capable of doing such a stringent constraining of the alternatives. The enablement can be represented as the probability distribution of a cause (stimulus) generating a certain event (posterior).

The Leibniz perspective of causality can be summed up, as a structural triplet approach [52], on which the flow of information goes into an enablement that then generates alternatives, with only a portion of those alternatives being capable of transmitting information. The enablement, the alternatives and the message/information are the three components of the Leibniz approach to causality. In the Leibniz-Damasio model of emotions/feelings, the feelings represent the fulfillment (positive feelings) or the failure (negative feelings) of a prediction. Using the Leibniz approach the negative feelings are divided into three types of failure, one for each of the components of the causality: *de-enabled*, *de-alternatived* and *de-messaged.*

A fourth group, associated to the *success* of the prediction, is also required. But as feelings are about prediction outcome, then only the success-associated emotion is a true feeling, with the negative feelings serving merely as indicators of how far one is from the positive feeling. This approach strongly coincides with the approach of positive psychology where the *Eros vs. Tanatos* duality is abandoned, in favor of a learning of how to cope with the increases and decreases in positive feelings that are a natural occurrence in life, e.g. [44]. In Damasio's perspective there were different types of positive feelings, each of them associated to a certain group of negative feelings (typically 3-4). What is

proposed in the Leibniz-Damasio model of feelings is that the different positive feelings can be assigned to different self-consciousness perspectives, with each positive feeling being associated to exactly three negative feelings [52].

Table 1.1 shows Leibniz-Damasio's levels of consciousness, with its four feelings per consciousness level [52], and also the neutral state proposed here. In this paper we will only consider the feelings associated to the Core-consciousness level (anger, fear, happy, sad, neutral). The Emotional Competent Stimulus for each one of these four emotions are:

- *Neutral* — the absence of an emotional competent stimulus is what we consider as a trigger for the *neutral* state.
- *Happy* — perception of a contribution to cooperation/communication. This emotion is associated with successful cooperation/communication between self and another individual, and so it is linked to the *successful* emotions.
- *Sad* — individual suffering/in-need. This emotion is associated with reduction/loss of the capacity to communicate with an individual, and so it is linked to the *de-messaged* emotions.
- *Fear* — weakness/failure/violation of the individual's own person or behavior. This emotion is associated with de-empowerment of the individual, and so it is linked to the *de-enabled* emotions.
- A*nger* — an interlocutor's violation of norms. This emotion is associated with loss of alternative possibilities of communion/cooperation, and so it is linked to *de-alternative* emotions.

1.3 State of the art

1.3.1 Emotive robots

According to Rich et al. [79] there has never been any doubt about the importance of emotions in human behavior, especially in human relationships. The past decade, however,

Table 1.1: DAMASIO'S CONDITION, CONSCIOUSNESS LEVEL AND HUMAN EMOTIONS.

		Consciousness Level		
		Proto-self	*Core*	Personal
Condition	De-alternative	Tension	*Anger*	Disgust
	De-enabled	Fatigue	*Fear*	Surprise
	De-message	Malaise	*Sad*	Jealousy
	Successful	Well-being	*Happy*	Pride
	Neutral	none	none	none

has seen a great deal of progress in developing computational theories of emotion that can be applied to building robots and avatars that interact emotionally with humans. According to the mainstream of such theories [29], emotions are very intertwined with other cognitive processing, both as antecedents (emotions affect cognition) and consequences (cognition affects emotions). The robot Autom, by Kidd and Breaseal [41], was designed for extended use in homes as a weight-loss advisor and coach. Autom builds on the research by Bickmore on long-term social interaction and behavior change using avatars. Recently, Schroder in [83] presented the SEMAINE API as a framework for enabling the creation of simple or complex emotion oriented systems. Their framework is rooted in the understanding that the use of standard formats is beneficial for interoperability and reuse of components. They show how system integration and reuse of components can work in practice. An implementation of an interaction system was done using a 2D displayed avatar and speech interface. More work is needed in order to make the SEMAINE API fully suitable for real case applications of emotion-aware systems [83].

Emotion recognition, in robotics context, is the capability of automatically recognizing which emotion a human is expressing among a finite scope of possibilities; this can be done using one or more modalities. In our case the scope is *{neutral, happy, sad, fear, anger}* and our modalities are image and sound.

Classifying emotions in human conversation was studied in Lee et al. in [43], where it was presented a comparison between various acoustic feature sets and recognition methods for classifying oral/spoken sentences based on the emotional state of the interlocutor. Later, Wang and Guan in [100], presented an *emotion recognition* system to classify human emotional state from audiovisual signals. The strategy was to extract

prosodic, Mel-Frequency Cepstral Coefficient (MFCC), and formant frequency features to represent the audio characteristics of the emotional speech.

A face feature extraction scheme based on HSV color model was used to detect the face from the background. The facial expressions were represented by Gabor wavelet features. This proposed emotional recognition system was tested and had an overall recognition accuracy of 82.14% of true positives. Recently, Cowie et al. in [12], described a dynamic approach to recognize emotions in video sequences. Recognition was performed by a neural network, it used a short term memory that are appropriate for modeling dynamical events in facial and prosodic expressivity.

As state-of-the-art shows, classifying emotions is a theme that shall be addressed from several different modalities that humans use in natural communication. When two or more modalities are used and fused as input to the final decision, the system is called a multimodal system. When a multimodal system is devoted to interaction, this system will perform what is called *multimodal interaction*.

Darwin studied *multimodal interaction* in humans in [17], a study about how humans express their emotional states placing a great emphasis on facial expressions. Ekman, using a more modern approach, studied emotional states and facial expressions across cultures [20] [21][22].

1.3.2 Bayesian framework and Probabilistic approach

In Bayesian algorithms, it is important to define a technique to fill-out the *Bayesian network* with information from the real world. As defined by Dahlback in [14], learning techniques are widely used for designing and testing natural language processing systems. A particular case of learning techniques was discussed by Klemmer et al. in [42], where the problem of spoken interaction was addressed. In our system, during the learning phase, the human experimenter embodies the strategy of the robot and interacts with another human. Meanwhile, the robot *looks* and *listens* to these two humans performing an interaction. From the observation, the desired variables are extracted and the likelihood tables can then be produced. At the running phase, when the decision moment arrives, the robot has a state and also a filled *Bayesian network*, thus it infers and performs the correct responses.

1.3.3 Facial expression recognition

Automatic *emotion recognition* from face images incorporates sub-problems as: finding faces, extracting features, and classification. Many of the current systems assume the presence of a face in the scene and do not automatically find faces, as for example: Paknikar in [65] and Wuhan in [101]. Pantic et al. in [66, 61] proposed to statically point a camera to the human face, so they did not really need to find the face since the image was always from a face. In human robot interaction applications, the camera is always on the robot and not on the human. Most systems assumes good illumination, a clean background and usually they do not provide any automatic tool to deal with illumination problems, in our approach we handle illumination problems for AUs detection with the dynamic threshold technique presented in chapter 2. Several improvements have been done in the area of extracting faces, a survey was presented by Yang et al. [102], moreover the OpenCV library incorporated the methodology presented by Viola and Jones in [99].

Many of the current approaches do not automatically extract the features, do not consider time sequence frames (dynamic Bayesian networks). However it is common that the face is divided in parts instead of analyzing the whole face image at once, as for example Pantic et al. in [66, 61]. About the classification techniques, we found on the literature: template-based classification (Wuhan et al. in [101]), fuzzy classification, ANN based classification (Paknikar in [65]), HMM based classification (Cohen et al. in [10]) and Bayesian classification (Sebel et al. in [84], similar but different from ours).

1.3.4 Auditory emotion recognition

Automatic *emotion recognition* of vocal expressions presented so far in the literature (for example: Mihalis et al. in [61] and Ververidis et al. in [97]), commonly use features based on: Pitch, the fundamental frequency of the acoustic signal; Energy, also called intensity or volume-level; Speech Rate, the number of words spoken in a time interval or the sentence/phrase duration when the number of the words inside each phrase is known; Pitch contour, the geometrical patterns of the pitch variations; Phonetic features, the pronunciation features.

1.3.5 Measuring engagement

It is difficult to measure how the interaction is improving. On literature, Stock and Strapparava in [91] and also in [92], claims that a interaction is better when the person consider the system to be funny. Specially those from the European project called "Hahacronym", we found descriptions of results but no detailed descriptions of assessments. However it is understandable that they performed experiment with several persons, while an external agent do a manual classification of how happy was the person with the performance of that system. Binsted et al. [6] measured the mean of "jokiness", "funniness" and also "heard before" possible classifications for each text, according to their defined assessments. The "jokiness" could be scored from 0 to 1. For "funniness" the range was defined from 1 to 5. For "heard before" the range of score was from 0 to 1. Later, Ritchie et al. [80] used the protocol of Binsted, and described the assessments more clearly. The system was shown to children and what was consider as a joke was also manually measured (by questionnaires after the dialog). In this thesis we also propose assessments to measure the engagement of the human during the interaction.

1.4 Objectives

Our objective is to create a smart robotic system with contactless interfaces (cameras and microphones) capable of a multimodal interaction with a human. Phoneme recognition is not our concern here, we are dealing with *emotion recognition* only on the auditory part, thus a story board for the human input is used, nevertheless the robot responses will vary.

Adapted from [23], figure 1.2 shows a conceptual drawing of the system, which is composed by three main parts: *memory (*where the knowledge acquired during the learning phase and also over time is stored*), analysis and synthesis.* The emotional vector is composed by (VE,FE and *SBP*) is the input for the decision process that will be explained later in section 4.4, for now it can be considered to be part of the synthesis.

Notice that there is a pair of the *analysis* part (from stimulus to Posterior), one per each channel. Later the visual and auditory channels information merges into the emotional vector. An overview of the *analysis* part is described bellow:

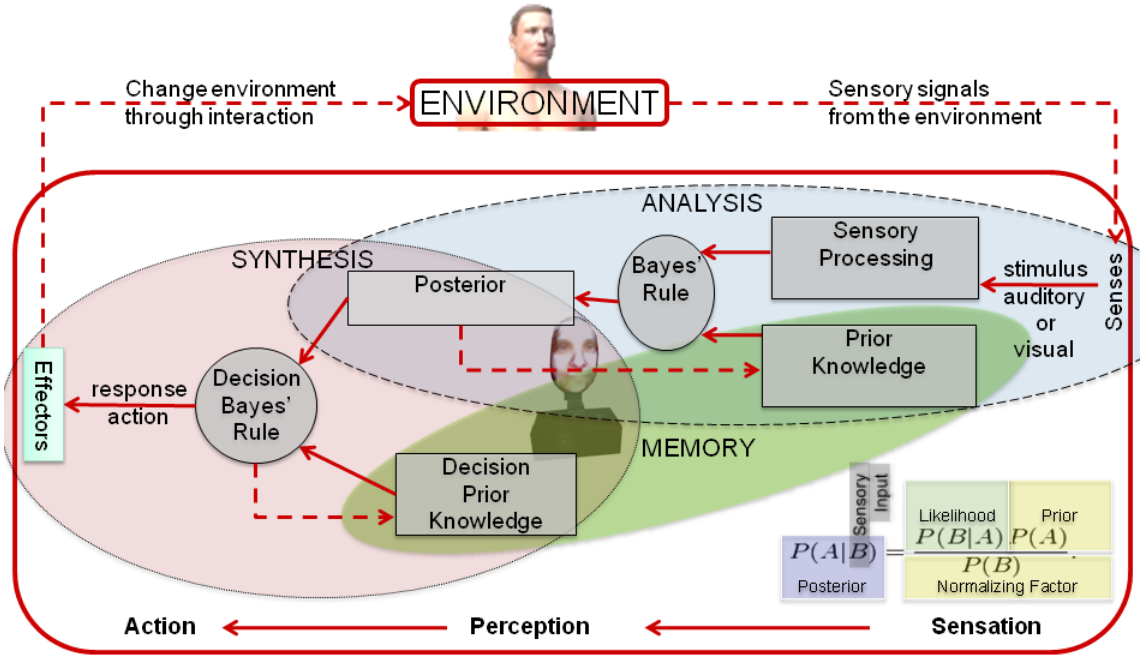


Figure 1.2: Two layers of the analysis part (represented by the dashed ellipse) shall be considered: one layer is for the visual analysis and another for the auditory analysis. The synthesis part (represented by the round dotted ellipse) is where the emotional vector is built, and also where the fusion takes place, this is a single layer. The stimulus comes from the sensors, per each modality, (image and sound), goes through the sensory processing part, where the features are extracted; the features are then input for the Bayes's rule which then infer the *posterior*. The *posterior* is the classified result. Later the two *posteriors* are merged with the robot *social behavior profile*. Finally the response (robot emotional state) is passed to the effectors.

- Stimulus: in our case, both image and sound.

- Sensory Processing: it is here that the facial Action Units (AUs) and the auditory variables are identified from the raw stimulus. This leads to the sensory input for the Bayes Rule (equation 1.1 explained on figure 1.2 on bottom right corner) to perform the correct question "what is the probability of a posterior, given the input from sensory processing?".

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

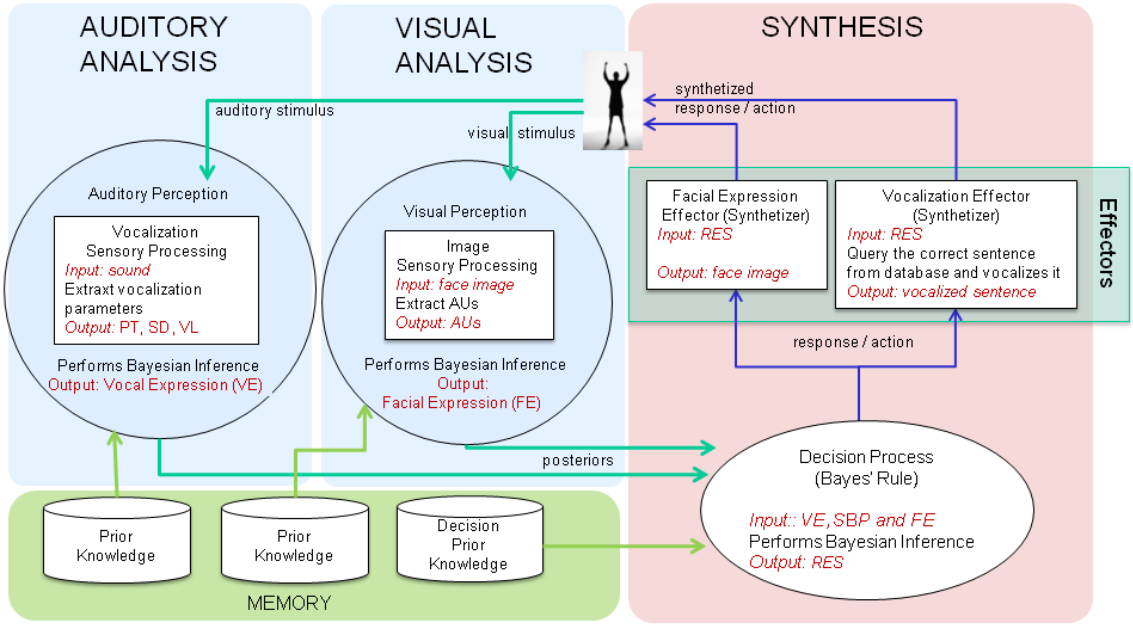


Figure 1.3: Break down through the conceptual system schema.

- Prior Knowledge: contains the prior and the likelihood. According to Bayesian Theory, prior is the probability of each event happening over a stimulus, independent of any other event. The likelihood is the probability (histogram if discrete) of an event happens given that another event already happened; it is filled out during the learning phase and it is stored in *memory*.

- Bayes's Rule: inference performed over the prior and the likelihood, in order to give the probability of posterior given the input from sensory processing.

- Posterior: inferred result.

In figure 1.2, the arrow that comes from posterior to the prior knowledge indicates that the *Bayesian network* has feedback. In other words, it starts by knowing just the prior, but through the passage of time, the robot acquires more "life experience" and the posterior will become part of the prior, thus the probability of an event that was already observed increases.

An overview of the *synthesis* part is described bellow:

- Posterior: the inferred results from the analysis are the inputs for the synthesis.
- Decision Prior Knowledge: this prior knowledge balances the fusion of the modalities.
- Decision Rule: Bayes's rule that infers a final decision over the emotional states coming from the posteriors of both modalities.
  - Response: Decision of what the effectors will actually do.

Figure 1.3 shows a break down through the conceptual system schema. The respective input and output of the two analysis processes are clearly separated, the posteriors of the analysis are VE and FE. The decision process takes the emotional vector as input. The emotional vector is composed by both classifier's results and the *social behavior profile* (*SBP*). The output of the emotional vector/decision process is *RES* that nicely stands for both Robot Emotional State and/or Robot Response. *RES* is then the input for both effectors. The effectors actually produces the robot's face and voice with lips synchronization (will be further detailed).

1.5 Experimental Platform

Our latest robotic platform was designed in a Scout platform and a head with 2 degrees of freedom and a support for a screen was added to show the expressions. Furthermore a retro-projectable mask was built for a better interactive interface (see figure 1.4 and 1.5a).

In both cases the robotic technology used as the experimental platform has an active vision system. This feature allows the robot to move its head towards the tracked face before starting to move its body, thus, the robot will avoid unnecessary movements with the body structure.

The structure in the back of the head allows a simple calibration of the projection. This calibration is possible due to the adjustable distances and mirror angles. The possible adjusts are:

1. distance between projector and first mirror,

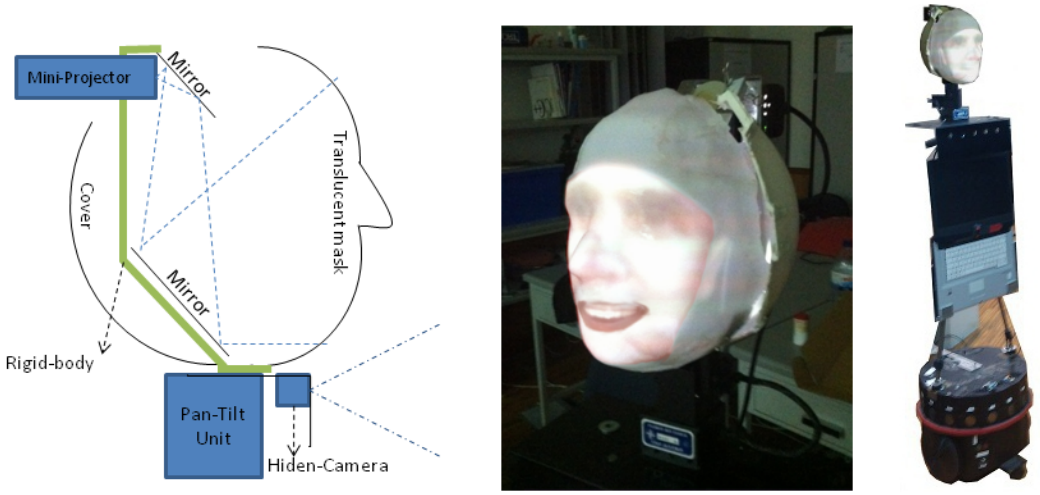


Figure 1.4: The head is composed by a retro-projected translucent mask which is attached to a rigid body. Two mirrors are attached to this same structure in order to deviate the beam of light projected. This setup was conceived to reduce the necessary distance for projection and thus close the head in a form more close to what would be a humanoid head.

2. distance between the two mirrors,
3. angle of the first mirror,
4. angle of the second mirror,
5. zoom screen is done with Linux built in desktop zoom,
6. focus of the projector.

These six parameter are adjusted in the beginning of the projection setup until the eyes and the nose fits to the correct place on the mask. This calibration is done manually and the error is visible by the displacement of the projection size, position or angle over the mask.

Furthermore, as another platform; we developed a 3D virtual world as a “Blender game”, where the same core of interactions can be used both over the real robot and/or inside the virtual world; see figure 1.5b. We generated 14 meshes of heads from 14 persons of our lab, so that we use the face of the real person on the avatar that mimics the person. Stereo vision systems are also an option that we consider. The segmentation can be used to improve the selection of which user to interact with.

(a) (b)

Figure 1.5: a) Current version of our robotic platform: With a Scout basis and a retro-projected head mounted over a pan-tilt with 2 DOF.
b) Our virtual world, it was developed as another option for interaction. When playing over it, a real person shall look at the camera and speak at a microphone; meanwhile an avatar mimics the user and another avatar simulates the robot.

1.6 Contributions

This thesis provides a number of novel contributions to the Human-Robot-Interaction field of research. The primary contributions of this research are as follows:

- A technique for dynamic background segmentation was proposed, applied for mobile robots [73].
- Bayesian models for visual and auditory expression analysis, these models were developed to be robust enough for achieving a reasonable classification, and simple enough to be computed in appropriated time for HRI [76]. Evaluation about the advantages of our proposed classifiers, and comparison with state of the art, was presented along this thesis.
- Bayesian models for synthesis of emotions on a human-robot-interactive platform

[74]. Our proposed methods of synthesis can emulate "trainable" emotive reactions in the robot.

- Emulation of personality in the robot; we did this by applying to the robotics field, the psychological concepts of five *Social Behavior Profiles* [75].
- Assessments for human engagement during a conversation were defined [75]. The advantages of emulating the five personalities were measured using the proposed assessments.
- A robotic prototype platform, SoPHIE (**S**ocial R**o**botic **P**latform for **H**uman **I**nteractive **E**xperimentation), was developed and implemented with a retro-projected translucent face mask [70].

1.7 Thesis summary

At the introduction, we studied how human emotions work, what are emotions for us and what is the importance of endowing emotions to a robot. Chapter 2 presents: the feature extraction processes, face attention, image sensory processing and auditory sensory processing. Furthermore, methods and tools, for detecting the necessary features from image and sound, are proposed and described. Still in chapter 2, sub-problems like noise in the image background and noise in the auditory background are addressed; a technique for dynamic background segmentation is presented.

Chapter 3 presents models for emotion analysis, once the extraction of features are covered by chapter 2, chapter 3 explains how to use these features in two different proposed Bayesian networks for classification; one for image, another for sound. Moreover, the classifiers are evaluated and compared with the state of the art by using known assessments. Discussion about the advantages of our methods are presented. Furthermore, the learning procedures for the dynamic Bayesian networks of analysis are described.

In chapter 4 models for emotion synthesis are presented. In order to endow the robot with the capability to express emotions, reverse Bayesian models are proposed, analogues to the classification ones. While the classification models comes from features to emotion, the reverse Bayesian models proposed in chapter 4 are able to come from the

emotion (that the robot shall express) to features. The learning for the synthesis (reverse models) are not presented in this chapter because synthesis reuses the learning of classification. An emotional vector is proposed to emulate personality to the robot, a Bayesian fusion method is presented and the learning for the emotional vector is described.

Chapter 5 presents how the emotional components are used into a dialogue between human and machine, how the dialogue takes place and the synchronization of visual and auditory channel. Several assertions are made in this chapter in order to turn the dialogue plausible for experimentation. After the constraints for the dialogue are defined, still in chapter 5 we define assessments to measure human engagement during the interaction. Experiments with the full system are performed and results concerning the human engagement are presented.

The conclusions and discussion of future work are presented in chapter 6. Appendix A presents a method used for comparing histograms. This method was used to check if the performed emotions from the reverse model matches the expected. Appendix B talks about how to implement Bayesian programs without the help of any library, it is a golden egg gift from this thesis for those who are starting with Bayesian programming. An example of the application of the Bayesian framework for a visuovestibular-based gaze control is presented on appendix 2.5.

1.8 Publications

The following publications resulted of the work leading to this thesis, this list of publications can be also found in *http://paloma.isr.uc.pt/mrl/people/peopleinformation.php*:

2012

8: "José Prado, Jorge Dias: *Emotive Dialog between Human Robot: A robot with personality* — submitted to ***International Journal of Human-Computer Interaction***, status: under review".

7: "José Prado, Jorge Dias: *Bayesian Emotional Behavior Analysis and Synthesis on*

Robot Artifacts — submitted to ***IEEE Transactions on Affective Computing***, status: under review".

6: "José Prado, Carlos Simplício, Nicolás Francisco Lori, Jorge Dias: *Visuo-Auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction* — ***International Journal of Social Robotics***, ISSN: 1875-4791, Jan 2012; *http://dx.doi.org/10.1007/s12369-011-0134-7"*.

2011

5: "José Prado, Lakmar Seneviratne, Jorge Dias: *Synthesis of Emotions on a Human-Robot-Interactive Platform* — in Proceedings of IASTED, Robo 2011, The 16th IASTED International Conference on Robotics, Pittsburgh, USA, November 2011".

4: "José Prado, Carlos Simplicio, Jorge Dias: *Robot Emotional State through Bayesian Visuo-Auditory Perception* — In proceedings of the 2nd Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS11, Costa da Caparica, Lisbon, Portugal, February 2011".

2010

3: "José Prado, Jorge Lobo, Jorge Dias: *SoPHIE Social Robotic Platform for Human Interactive Experimentation* — COGSYS 2010, 4th International Conference on Cognitive Systems, ETH Zurich, Switzerland".

2009

2: "José Prado, Luis Santos, Jorge Dias: *A Technique for Dynamic Background Segmentation using a Robotic Stereo Vision Head* — in Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication Toyama, Japan, September 27 to October 2, 2009; *http://dx.doi.org/10.1109/ROMAN.2009.5326303"*.

1: "José Prado, Luis Santos, Jorge Dias: *Horopter based Dynamic Background Segmentation applied to an Interactive Mobile Robot* — in Proceedings of the 14th International Conference on Advanced Robotics, ICAR 2009, Munich, Germany, June 2009".

Moreover, several cooperative work was done inside the Institute of System and Robotics, the following papers: [36], [71], [18], [1], [25], [82], [26], [88], [86] were resultant from this cooperation.

Chapter 2

Pre-Processing and Feature Extraction

2.1 Face Attention

In order to perform automatic facial expressions recognition, detecting the face among the image is necessary, additionally it is desirable to keep the robot targeting this face. We call this process face attention. The face attention technique proposed here takes advantage of the symmetry extant in human faces. The objective is to endow the robot with the ability to keep always targeting the human face frontally. In order to accomplish this task, the robot performs an arc movement, the arc is centered in the human and ending at the robotic head.

The face attention process can be decomposed in two parts: face pose identification and robotic system controller. The face pose identification system is the part of the system responsible for detecting the human 2D head orientation (we are not detecting tilt orientation), while the robotic system controller is responsible for moving the robot in order to keep the face inside the image captured from the camera. An overview of the face attention system is presented in figure 2.1, notice that the purpose of these components are to prepare the robot for the facial expression recognition. In summary, the face pose identification system processes an image to provide an angle (2D head orientation); then the robotic systems controller receives as input this angle and move the robot accordingly. Later with the robotic head positioned frontally to the human, the facial expression recognition system can start.

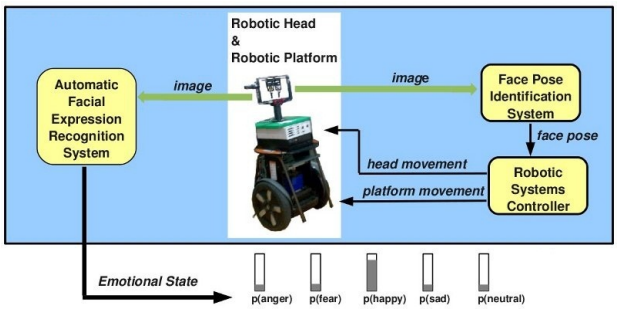


Figure 2.1: Designation of the modules necessary for Face Attention.

2.1.1 Robotic Systems Controller

The robotic platform does three types of movements (longitudinal or transversal translations and rotations) following the provided commands. Longitudinal translations are performed to approach or move away the robot from the human. However the method used for face tracking works independently of the size of the image, it is easy to understand that as many pixels the image has, as better the recognition of facial expressions can be. The objective of using this movement is to keep the robot close to the human. Transversal translations are performed to keep always the face present in the image, the face does not need to be centralized, but need to be there. Rotations correspond to an arc of circle centered in the human. These movements are performed to follow the rotation movements done by the human being, getting always the image from a frontal point of view. The robotic head moves in synchronization with the basis if necessary.

2.1.2 Face Pose Identification System

In a perfect bilateral symmetric image, the difference between a pixel value and the respective symmetric counterpart is zero. By nature, human faces do not present a perfect bilateral symmetry and it is reflected in the acquired images. Moreover, the "imperfections" in the images are worsened by the noise associated to the acquisition process or due to lights distribution. Despite all these noises, gray level differences can still be used to detect the bilateral symmetry axis of an human face.

The method proposed here to identify the axis of symmetry is based on rather simple principles but is very effective. A vertical axis is defined to divide the face image

region in two parts with equal number of pixels and a Normalized Gray-level Difference Histogram (NGDH) is built.

To calculate the NGDH, let's state some variables only for this section:

- h is height of the image,
- w is the width of the image,
- M is a 2D matrix with dimensions w per h, and it contains the image,
- i is an integer variable that can vary from 0 to $w/2$,
- j is an integer variable that can vary from 0 to h,
- D is a 2D matrix with dimensions $w/2$ per h.

Then, D will be used to collect the differences between the left half of the image when compared to the mirrored right half of the image, in gray scale unsigned 8-bit image. The following equation shows how the D matrix is computed:

$$D_{i,j} = M_{i,j} - M_{w-i,j}$$

An histogram is a function that counts the number of observations that fall into each of the disjoint categories (this categories are known as bins). Since we used 8-bit unsigned image, the gray level varies among 0 and 255. Thus, the possible difference values ($D_{i,j}$) varies from -255 to $+255$, thus the number of beans in our NGDH is defined as 511.

When the face is frontal, this vertical axis bisects it and the information collected in the NGDH is strongly concentrated near the mean. Since we have positive and negative values, the mean is usually around 0 zero. If the image is 100% symmetric, the NGDH is a *Dirac* in zero (all the $w/2 * h$ measured pixels has 0 difference).

While the robot is static, the camera is in a fixed position, thus, when the human face rotates the defined axis is not anymore the perfect symmetry axis and the information is scattered along the histogram. Instead of using directly the mean, which could lead to undesired peaks, we used a narrow region around it: we called the *pseudomean*.

If μ is the mean of the histogram (NGDH), so the *pseudomean* is computed by:

$$pseudomean = \frac{\sum_{k=\mu-10}^{k=\mu+10} NGDH_k}{20}$$

Our method requires a frontal face just in the initialization, but it assumes in all phases an upright face. The algorithm begins performing the face detection using Haar-like features [98]. In this way a region of interest is found and a vertical axis is established in the middle of this region. To find the real face orientation, the region of interest is successively rotated about that axis using a 3D transformation. For our purpose, this angle does not need to be determined perfectly. To keep the system running fast we limited the search for five rotation angles. Thus, angles taken from interval $[-30^o; +30^o]$ with 15^o steps are used to generate five synthetic images. For every synthesized image, the NGDH is built and the *pseudomean* is computed. The five *pseudomeans* are compared. The detected face orientation, corresponding to the greatest *pseudomean*, is then sent to the robotic systems controller module.

2.1.3 Results

Figure 2.2 presents some results obtained by the proposed technique for face pose identification. Figure 2.2 (a) shows a scene in our laboratory with the human face segmented. Still in figure 2.2, from (c) to (g) are presented the synthetic images generated by the application of the 3D transformation, and from (j) to (n), the respective NGDHs. The *pseudomean* increases as the synthesized image becomes more similar to a frontal face, as it is observable at the histograms. In (h) is illustrated graphically the output as a line segment drawn superimposed to the segmented face. The angle of this segment, referenced to the image vertical axis, corresponds to the detected face angle. In this case a vertical segment means a frontal face.

When the human is not facing the robot frontally, one of the synthesized images presents an approximation to a synthetic frontal face, this approximation is enough for the correct angle to be detected.

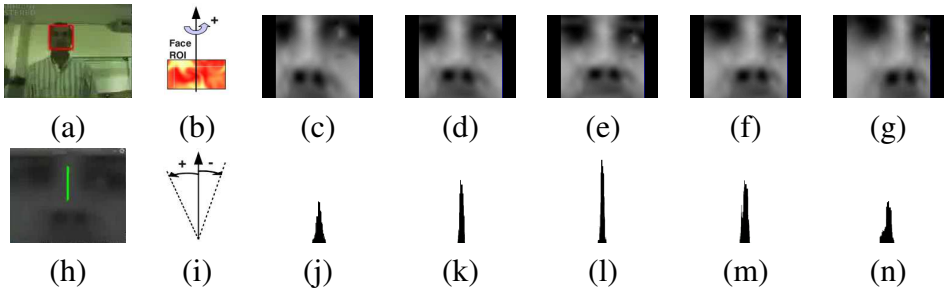


Figure 2.2: Principles of our method for face pose identification. (a) shows the image of a scene with a frontal face segmented. (b) defines the direction of the signal used on the rotations. In (c),(d),(e),(f) and (g) synthetic images from the application of a 3D rotation, angles are respectively -30^o,-15^o,0,$+15^o$,$+30^o$. The respective NGDHs are presented in (j), (k), (l), (m) and (n). At (h), a vertical line means a 0^o face pose (frontal face). In (i) is defined the vertical axis reference for the output image.

2.2 Image Sensory Processing / Feature Extraction

The first step in image emotion recognition, before classification, is the extraction of features which in our case are the Action Units. Action Units (AUs) are small distortions over the face that together characterize a facial expression, as defined by Ekman in [20].

Human facial expressions are created by movements of facial features (e.g., eyebrows, eyes, nose, mouth) and arise as a result of muscular activity. This activity can be performed voluntarily; for example when performing a grimace. However, human beings systematically perform involuntary facial expressions. These are a form of nonverbal communication used in social contact as means to express emotions. In fact, humans are beings with a strong social characteristic, and facial expressions are a primary mean of conveying social information. In figure 2.3, examples of facial expressions are presented, typically associated to some emotional states. The association between emotional states with facial expressions is substantial. For certain emotions it is inevitable that the person performs the characteristics of the correspondent facial expressions; even when one wants to hide the real emotional state. Nevertheless, this close relationship between emotional states and facial expressions may work “in the opposite direction”. Thus, in our opinion, the expressiveness of the robot may influence on the emotional state of the human interlocutor, leading him/her to different levels of engagement/commitment along

the interaction with the robot.

Darwin studied how humans express their emotional states [17]. It is an extensive study focusing on the various forms used by humans to express themselves: facial expressions, gestures, vocalization, etc. More recently, Paul Ekman devoted specific attention to the subject of emotional states and facial expressions [20] [21][22]. In [22] it is mentioned that an alternative approach to measuring facial expressions of emotion is through systematically examining video records to identify the muscular movements that constitute the emotional expressions. One advantage of this approach is that it is totally unobtrusive.

Figure 2.3: Facial expressions examples with different users among the scope: *{neutral, happy, sad, fear, anger}*.

Facial Action Coding System (FACS) [20] defines a total of 52 Action Units (AUs) where 8 of them are related with the head pose. The remainder 44 concern small distortions over the face which characterize the facial expressions. Each of these AUs is anatomically related to the activity of a specific set of muscles which produces changes in the facial appearance. Therefore, a facial expression can be seen as a set of specific AUs, which causes "distortions" in facial features (i.e., mouth, eyes, eyebrows or noise). By identifying these distortions, facial expressions can be recognized. In our work, only a small sub-set of the AUs introduced by Ekman was used (examples are presented in figure 2.4).

The initialization of visual feature extraction takes place when a facial expression is performed by the user in front of the camera. If there is no face among the captured image, the process does not fail, because it does not even starts. The Viola-Jones haarlike features [98] embedded on the OpenCV library [33] is used to detect a human face from each frame. Then, the robot stores each frame of the video for on-the-fly analysis. Later

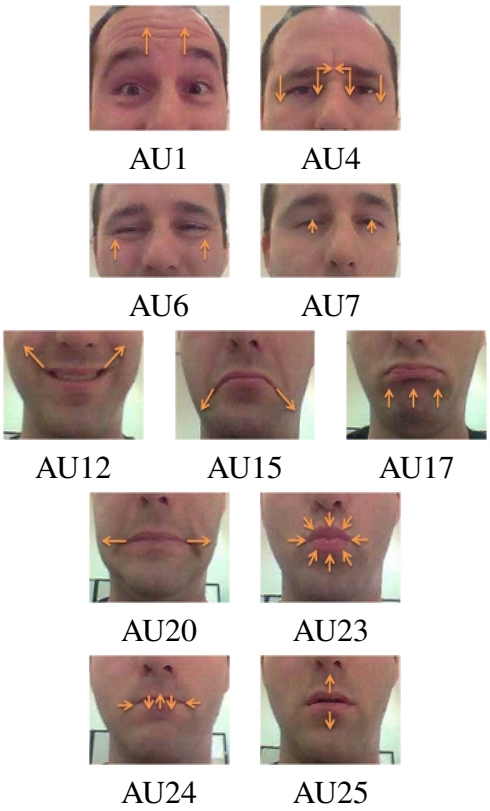


Figure 2.4: The covered Action Units, description can be found at table 2.1.

the face is divided into upper face and lower face. From this point on, the AUs are detected over the upper face image and lower face image.

To find these AUs among the input image, the upper face and lower face are treated independently, a feature vector is then extracted from each image while the user performs that specific expression. After this, we have information enough to detect whether the Action Unit is present or not in the upper face. The same was done for the lower face, independently of the result for the upper face. The method that is detailed on the next section succeeded to detect 10 of the 11 presented AUs. At our current implementation, we are not detecting AU17, since the remaining AUs demonstrated to be enough for the classification purpose. Some results can be seen in figure 2.6.

Table 2.1: DESCRIPTION OF ACTION UNITS.

Action Unit	Description
AU1	raised inner portion of the eyebrows
AU4	lowered eyebrows
AU6	raised cheeks
AU7	raised lower eyelids
AU12	lip corners up
AU15	lip corners down
AU17	chin boss up
AU20	mouth horizontally stretched
AU23	mouth horizontally tightened
AU24	mouth closed
AU25	mouth opened

2.2.1 Detecting the Action Units

Each AU is defined by a scope of the position of some feature in the image. Thus we find out that to fill every possible case of the action units, some points in the image are important to be detected. In our perspective, the Action Units are a scope of the distances defined between some of these points. This scope is not thresholded, but defined later by the Bayesian learning, for now, let's clarify what points are necessary to be find in the image.

For the lower face, consider that: The leftmost point of the mouth is ($mouthleft_{x,y}$), the rightmost point of the mouth is ($mouthright_{x,y}$), the highest point of the mouth is ($mouthhigh_{x,y}$), the lowest point of the mouth is ($mouthlow_{x,y}$), the lowest point of the chin is ($chin_{x,y}$). For the upper face, consider that: The center points of the eyes are ($eye1_{x,y}$ and $eye2_{x,y}$), the central point of each eyebrow is detected as ($eyebrow1_{x,y}$,$eyebrow2_{x,y}$), the highest pixel of the eye region are $starteye1_{x,y}$ and $starteye2_{x,y}$; the lowest pixel of the eye region is $endeye1_{x,y}$ and $endeye2_{x,y}$. When we raise our lower eyelids for anger expression we usually involuntary set the eyebrows together, thus we assume that when the eyebrows are together the AU7 is present. Table 2.2 shows the relation of the Action Units with the interest points to track in the image/video sequence.

Table 2.2: RELATION BETWEEN ACTION UNITS, THE RELATED FEATURES AND THE POINTS TO DETECT.

Face Part	Action Unit	Related Feature	Relation to the Detected Points
Lower	AU20 and AU23	Mouth Form (horizontal)	$\lvert mouthright_x - mouthleft_x \rvert$
Lower	AU24 and AU25	Mouth Aperture (vertical)	$\lvert mouthlow_y - mouthhigh_y \rvert$
Lower	AU12 and AU15	Lips Corners (height)	$(mouthleft_y + mouthright_y)/2$
Lower	AU17	Chin Boss	$chin_y$
Upper	AU1 and AU4	EyeBrows (vertical distance between eyebrows and eyes)	$(\lvert eye1_y - eyebrow1_y \rvert + \lvert eye2_y - eyebrow2_y \rvert)/2$
Upper	AU7	Lower Eyelids (horizontal distance between eyebrows)	$\lvert eyebrow2_x - eyebrow1_x \rvert$
Upper	AU6	Cheeks (raising the cheeks, closing the eyes)	$(\lvert endeye1_y - starteye1_y \rvert + \lvert endeye2_y - starteye2_y \rvert)/2$

Selecting Region of Interest

After the steps of initialization, the first step in the AUs detection is selecting the lower face ROI and the upper face ROI. From the original input image matrix, we first select and crop the face (based on the result of the haarlike features detection), then the nose region is cropped out to reduce the image size. Then we sub-select a region of interest related to the lower face (*lowerFaceROI*) and a region related to the upper face (*upperFaceROI*), these sub-selection are static, relative to the face detected position, considering that every person has the mouth and the eyes in similar region of the face. Then *lowerFaceROI* contains only the image of the mouth region and *upperFaceROI* has only the region of the eyes and eyebrows, as shown in figure 2.5. Our input image is currently 320x200, so of course after so many digital zooms, the ROI image has a very low resolution. We are keeping this input resolution to maintain the real time processing, however it is possible to easily increase the size of the input image having though a better resolution ROI.

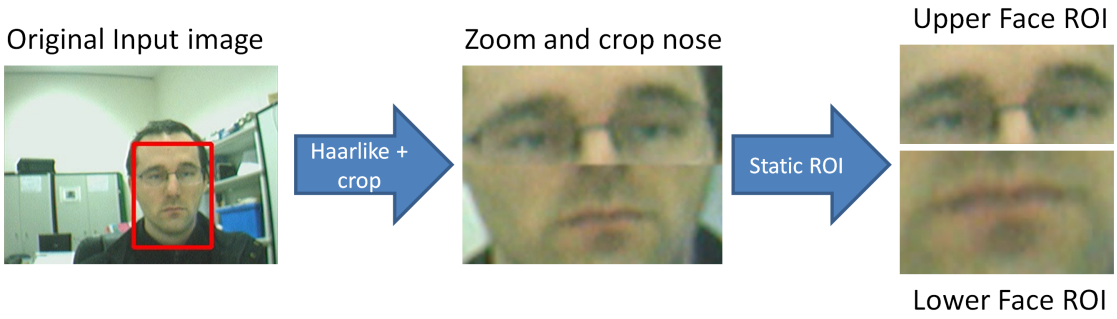


Figure 2.5: From original input image to face ROIs.

Dynamic Threshold

This is the second step, it is performed in order to the detection to be independent of illumination conditions. It is a simple step, but primordial for the rest of the detection. First we traverse the image ROI and find the darkest pixel. From this darkest pixel we then selected three different thresholds, these are dynamic because they are summed with the darkest pixel intensity. After testing in several different datasets, from different persons, we empirically find our three dynamic thresholds: $DT1$ for mouth, $DT2$ for eyes and $DT3$ for eyebrows. When applying these, we have the first level of detection, illustrated

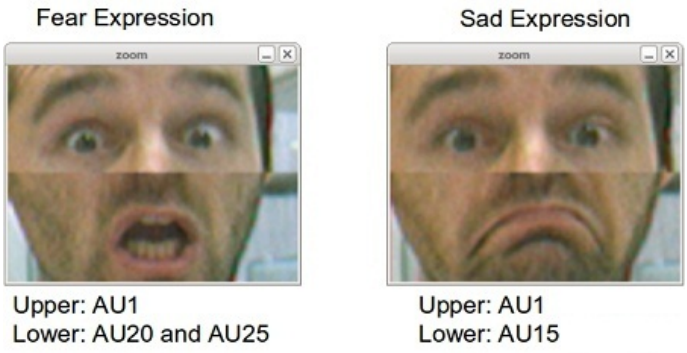


Figure 2.6: Sample results of our method for image face feature extraction. Left image presents a fear expression; right image presents a sad expression. The face is separated in upper and lower face. The nose part is discarded. The detected AUs are presented under each of the respective images.

in figure 2.7.

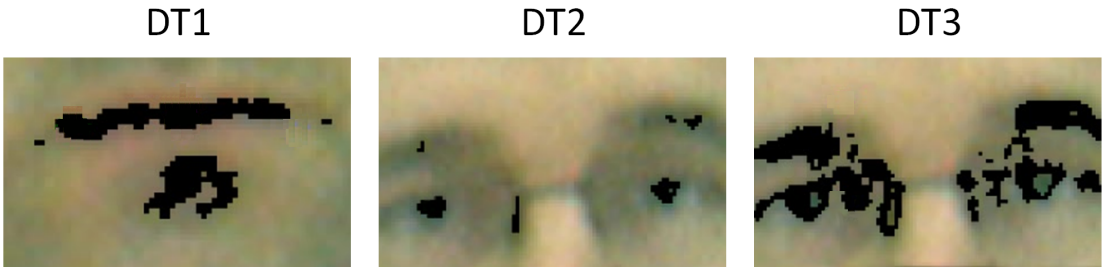


Figure 2.7: Dynamic Threshold.

Region Eroding

The third step in our detection is region eroding. To reduce the noise and to disconnect possible noises from the interested regions, we then apply a region eroding algorithm, as defined in [28], onto the pre-detected regions. This algorithm is based in a $3x3$ mask and the result of the erosion is presented in figure 2.8. Blue indicates the eroded pixels while black indicate the remaining pixels.

Region Segmentation

The fourth step is segmenting the regions. We implemented a variant of k-means algo-

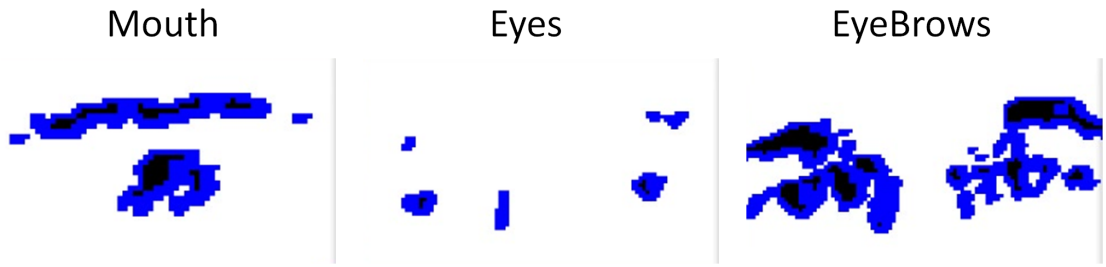


Figure 2.8: Region Eroding.

rithm in order to segment the regions of each image. Our variation set together regions that are very close and finally paint each region with a different color index for further analysis. A sample result of our variation of k-means is presented in figure 2.9.

Figure 2.9: Region Segmentation.

Finding Points

We took advantage of the symmetry of our facial expressions, thus, the side of the face with a better detection in the previous step is mirrored to the other side (left to the right or right to the left). The mirror decision is fully automatic, the non interest regions are also discarded automatically. For the mouth and for the eyebrows, the first region from top-down is selected, for the eyes the first image from down-top is selected. The result of mirroring and region discarding are shown in figure 2.10. Later a pseudo-face is composed (figure 2.11) and the points are easily found.

For some frames, the haarlike features might fail, consequently the pseudo-face is not built correctly, if this happens the frame is automatically discarded.

Figure 2.10: Region mirroring and discarding.

2.2.2 Assessments for the Image Sensory Processing

It is known that the classifier's results depend directly on the effectiveness of the detectors. Usually researchers test the system in optimal conditions and do not do stress tests with different environments. We decided to also define assessments for the used detectors. Once again this was done for both the auditory feature extraction and for the visual (face images) feature extraction.

For the images we measured the *percentage of correct face feature extractions* across 100 frames. Ten iterations were done in 3 different environments keeping the same facial expression, the same performed emotion (neutral) and the same user. The environments differs by the amount of noise in the background. Good, medium and bad were used to indicate the quality of the input image, as regards the amount of noise on the image background.

1. Good environment: person alone with a clean image background. Studio conditions, a curtain was setup behind the person;

2. Medium environment: person alone with a random background. Standard noise, no curtain was setup, the wall and objects behind the person appears. Thus, sometimes the features of the objects can be confused with face haarlike features;

3. Bad environment: person not alone, other faces were in the image. High background noise, no curtain was setup, the wall and objects behind the person appears, other faces behind the target person appears. Thus, several times the features of

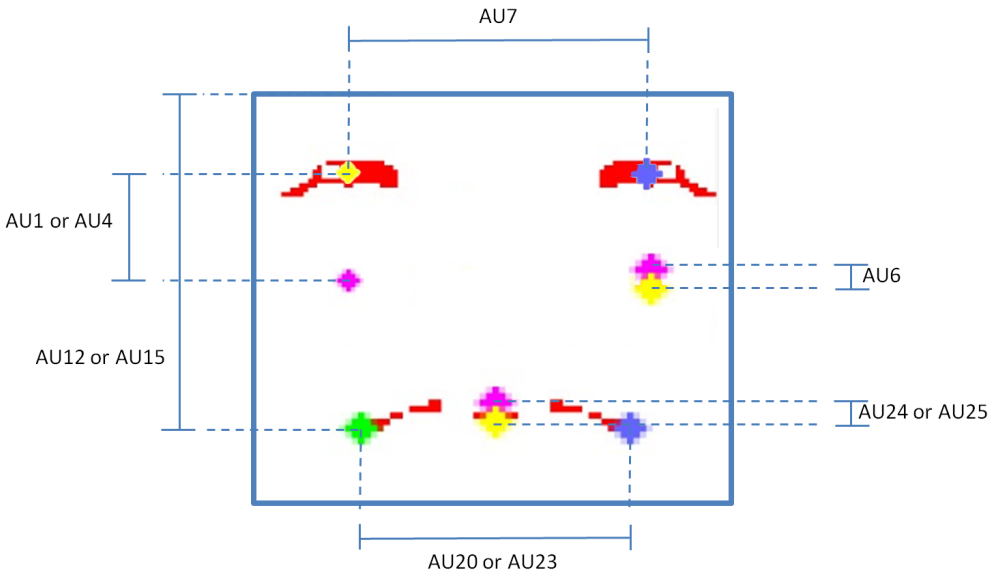


Figure 2.11: Pseudo-face and interest points.

the objects and the persons behind causes confusion on the haarlike features face detection.

Results for Visual Sensory Processing and Discussion

According to what was defined for our assessments, the hit rate was manually annotated according to the expected Action Unit. After annotation, a percentage of the correct detections was calculated and are shown in table 2.3. For good and medium environments, the Action Unit detector presented results with accuracy higher than 78%. To overcome the low hit-rate problem identified on the bad environment, a dynamic background segmentation technique, for mobile robots, will be presented.

Table 2.3: THE PERCENTAGE OF CORRECT DETECTED ACTION UNITS ARE SHOWN, RESPECTIVELY, OVER THE THREE DIFFERENT ENVIRONMENTS AS DEFINED ON ASSESSMENTS. ONE HUNDRED FRAMES WERE ANNOTATED FOR EACH ENVIRONMENT.

	Good Environment	Medium Environment	Bad Environment
Correct Detections	98% hit rate	78% hit rate	53% hit rate

Figure 2.12: Four degrees of freedom robotic head mounted on Segway robotic platform body.

2.3 Dynamic Background Segmentation and Zone of Interaction

Several approaches like face detection and face recognition [69, 85, 94, 77] often have to deal with issues associated to bad illumination and strong featured background. These problems also imply lack of performance because human detection algorithms will frequently analyze the whole image searching for features. Hence we propose a stereo vision dynamic background segmentation (DBS) to reduce the searching space to an *zone of interaction*[1]. We explain how the horopter calculation proceeds and further we give an example of how face and hand recognition frequently used on gesture recognition algorithms could have better results with our approach. This research focus was also explored by us on [73] and [72]. For the the horopter segmentation to be possible a stereo vision system is required, an example of one of the platforms used during this research is shown at figure 2.12.

[1] *zone of interaction* is the region inside the horopter 3D space (see theoretical horopter definition on section 2.3.3)

2.3.1 Related work on background subtraction

By using a reference image, a video coding approach with Motion JPEG2000 has previously been developed in the context of road surveillance [95]. Moreover it was shown how the image reference was built during initialization phase. The classical background subtraction technique was used to perform the segmentation of mobile objects. Instead of updating the remote reference with a specific period, [95] presented a technique to update the remote background image by pieces. The updating of the remote reference is triggered when some specific conditions are met, depending on the amount of moving areas.

In [56] an integrated system for smart encoding in video surveillance was presented. Their system aims at defining an optimized code-stream organization directly based on the semantic content of the video surveillance analysis module. The proposed system produces a fully compliant motion stream that contains regions of interest (typically mobile objects) data in a separate layer than regions of less interest (e.g. static background). First the system performs a real-time unsupervised segmentation of mobiles in each frame of the video. The smart encoding module uses these regions of interest maps in order to construct a code-stream that allows an optimized rendering of the video surveillance stream in low bandwidth wireless applications, allocating more quality to mobiles than for the background. The integrated system of [56] improves the coding representation of the video content without data overhead. It can also be used in applications requiring selective scrambling of regions of interest as well as for any other application dealing with regions of interest.

On [81] horopter is calculated and vergence control is explained on a stereo-vision-system applied to *tracking by using optical flow*. The robotic stereo head presented on [81] is not mounted on a mobile robot. It is also noticeable by the result images of [81] that the resolution of the disparity map is 36x36 pixels. Our approach focus on applying the system to an *interactive mobile robot*, in this case, the calculation of disparity need to be very fast, otherwise robot body rotation and translation will easily generate errors into the disparity map.

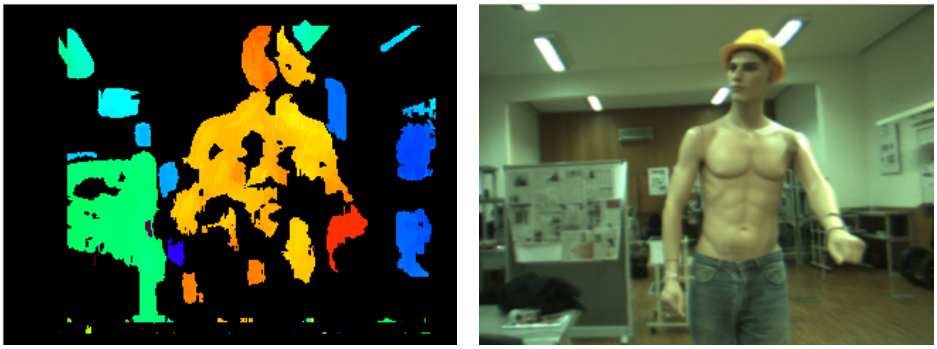
(a) (b)

Figure 2.13: Horopter segmentation schema: a)Noisy scenario, another subject trying to interfere during the interaction; b) From the robot point of view, ignoring the interference.

2.3.2 Motivation

Within the context of human-robot interaction based on computer vision, there is a prerequisite, the need for the robot to recognize the person with whom it will interact. Usually it is done using a video sensing. Since the system is implemented in a mobile platform, to separate the person from the background demands more complex processing, due to dynamic characteristics of the background. This means that an approach based in static background, as in [95] and [56], is not plausible for this case. The challenge was thus to have a robust real time solution for dynamic background segmentation on mobile robotics.

Our approach is then based on the *Geometric Horopter* as will be shown in section 2.3.3. The robot considers objects as "visible" only if they are inside the *zone of interaction* region (projected on 2D space of camera image plane). In figure 2.13a it is possible to observe the horopter represented by a semi-circle dashed line at the floor; the subject on the right of the image is purposely in a pose that would interfere on the analysis of several algorithms [69, 85, 94, 77, 63, 62]. Once applying our strategy of DBSH (Dynamic Background Segmentation based on Horopter) the robot will only see the person that is inside the horopter, according to figure 2.13b.



(a) (b)

Figure 2.14: Results of depth-map calculation: a)Depth map (*'hot'* colors represent nearest areas, *'cold'* colors represent further ones; b) Dominant eye raw image.

2.3.3 Horopter and Zone of interaction

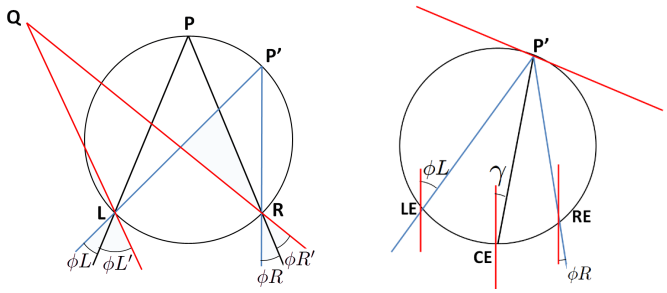
Our approach is based on the *Geometric Horopter*. This technique used stereo vision to produce a *depth map*. It is presented in figure 2.14a the *depth* map resulting from the application of this algorithm over the input image shown in figure 2.14b, while the right side shows the image from one of the stereo cameras.

The application of the *horopter* leads us to the definition of the *zone of interaction*. Only objects inside the circled area of the *zone of interaction* are possible of being detected, and consequently to interact with the robot.

The ViethMuller Circle The concept of *interaction zone* has been defined as dependent of a circle. That circle is called the Vieth-Muller Circle, as defined in [93], the following properties can be observed:

- In a pure version eye movement, the fixation point stays on the same ViethMuller Circle. A pure version eye movement is when either the two eyes, or the two cameras, rotate exactly the same angle. Figure 2.15a illustrates this fact showing how P moves to P' along the Vieth-Muller Circle whileϕL is equal to ϕR.

- It the fixation point remains static, the disparity for various points can be calculated.



(a) (b)

Figure 2.15: a) Calculating the Disparity; b) Disparity Properties on Vieth-Muller Circle.

The disparity map and consequently the horopter changes every time that the angles ϕLC and ϕR vary.

Property 1 *If a point Q lies on ViethMuller Circle, its disparity is zero.*
As Q moves outside (e.g. point P moves to position Q in figure 2.15a), ϕL decreases whilst ϕR will naturally increase. However ff point Q moves inside the circle, the opposite relation between ϕL and ϕR occurs.

Property 2 *Disparity is nonzero outside the circumference line of the Vieth-Muller Circle (with opposite signals, depending on whether side of the circle it lies in, outside or inside).*
For human vision system, when the disparity has high enough values, the object is seen in double (one from left eye and the other from right eye). This phenomenon is called *Diplopia*. The maximum disparity prior to the diplopia even is defined as *Panum's Fusional Limit*.

Calculating Disparity The ϕL and ϕR are made by line of sight with the straight ahead direction. The *GazeAngle* γ (see figure 2.15b) and *VergenceAngle* μ (see figure 2.16) are defined as

$$\gamma = \frac{1}{2}(\phi L + \phi R)$$

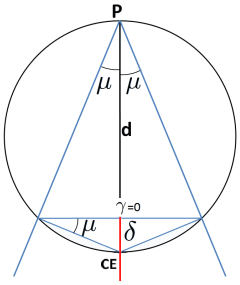


Figure 2.16: Simples justification scheme for value $\gamma = 0$. See γ angle also on figure2.15b.

$$\mu = \frac{1}{2}(\phi L - \phi R)$$

CE represents the cyclopean eye and $(d+\delta)$ is the distance from CE to the target object (see figure 2.16).

The Horizontal Disparity is

$$h = \frac{I\cos\gamma}{d}\left(\frac{\delta}{\delta+d} + \frac{d\tan\gamma}{\delta+d}x + x^2\right)$$

and Vertical Disparity

$$v = \frac{I\cos\gamma}{d}\left(\frac{d\tan\gamma}{\delta+d}y + xy\right)$$

where (x,y) are cyclopean image coordinates and I is the interocular distance.

Property 3 $d = I\cos\gamma/\sin 2\mu$

Practical Demonstration A simple practical demonstration can be presented for the value of $\gamma = 0$, as it can bee seen in figure 2.16.

$$I/2 = d \times \sin u \times \cos u \Rightarrow d = I\cos r/\sin 2u$$

Having disparity calculated, the resulting depth image (figure 2.14a) is correlated with the CE image. Pixels that present negative values for disparity, will be assigned zero value (black color pixels). The result is a segmented image where the pixels calculated to

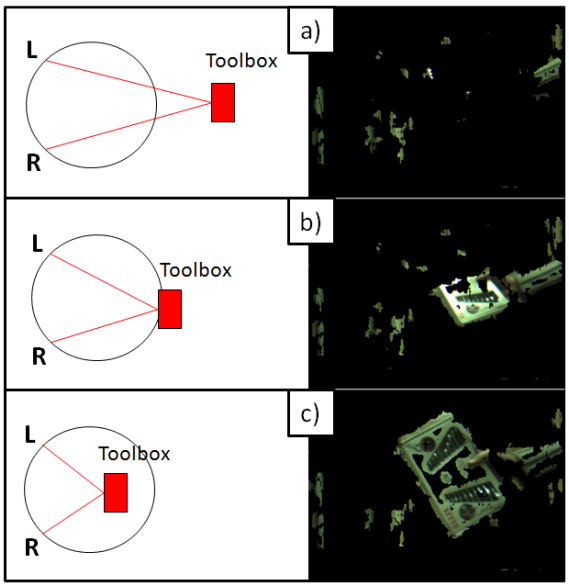


Figure 2.17: a) The toolbox is yet outside the Vieth-Muller Circle; b) Toolbox starting to enter the horopter zone; c) The object is fully inside the Vieth-Muller circle, and thus, visible.

be inside the *Vieth-Muller* circle define the visible objects within the circle (the interaction zone). The segmented image (right column of figure 2.17) results in a region of interest and this region will define the true input pixels for the *face/hand* detector. Consequently the robot will interact only with subjects inside *Vieth-Muller* circle, *i.e.* inside its current horopter.

Aperture Problem

As mentioned by Lobo et al. [47]: "the correspondence of an isointensity contour region cannot be estimated, this is the so called aperture problem". Thus, the erroneous areas extant on the results presented on figure 2.17 are due to homogeneous areas in the original image. Homogeneous areas and also very similar neighbor features of the image disturbs the depth map and consequently disturbs the final segmented image. Although we did not overcome the aperture problem, the result is still better for hand and face detection than if you have no segmentation.

Hardware Platform

Our robotic head (figure 2.12) is a common platform consisted by many sensors. In

this particular case we used two monocular cameras (which compose our stereo vision system) and the four degrees of freedom of our robotic head figure2.12 (head pan, head tilt, vergence in eye left, vergence in eye right). The cameras are two AVT Guppy Fire-Wire Cameras. Thus, we calculate stereo imaging for real-time depth-map using the triangulation principle. Camera images are transferred to a PC using the IEEE 1394 (Fire-Wire) bus. The PC is a laptop computer that can be attached in a tray inside the robot body which is an adapted Segway RMP (Robotic Mobility Platform). The Segway RMP adaptions made by us consists basically on four suspended legs to fall avoidance, a strong box for sensor batteries and a tray for the robotic head hardware controllers and laptop attachment.

2.3.4 Face detection

Featured base face detection

A multi-stage classification procedure has been proposed by Viola and Jones in [99], that reduces the processing time substantially while achieving almost the same accuracy as compared to a much slower and more complex single stage classifier. Later Lienhart and Maydt [46] extends their rapid object detection framework in two important ways: Firstly, their basic and over-complete set of haar-like feature is extended by an efficient set of 45° rotated features, which adds additional domain-knowledge to the learning framework and which is otherwise hard to learn. These novel features can be computed rapidly at all scales in constant time. Secondly, [46] derive a new post-optimization procedure for a given boosted classifier that improves its performance significantly.

More recently Bau-Cheng Shen and Chu-Song Chen proposed a new method to retrieve similar face images from large face databases. The proposed method extracts a set of Haar-like features, and integrates these features with supervised manifold learning. Haar-like features are intensity-based features. The values of various Haar-like features comprise the *rectangle feature vector* (RFV) (detailed by Chen et al. in [85]), to describe faces. Compared with several popular unsupervised dimension reduction methods, RFV is more effective in retrieving similar faces. To further improve the performance, Chen et al. [85] combine RFV and a supervised manifold learning method and obtain satisfactory retrieval results.

Skin color hand detection

According to Tarek [54], skin color can provide a useful and robust cue for human-related image analysis, such as face detection, hand detection and tracking, people retrieval in databases and Internet, etc. The major problem of such kinds of skin color detection algorithms is that it is time consuming and hence cannot be applied to a real time system. To overcome this problem, a fast technique for skin detection was introduced, which can be applied in a real time system. In this technique, instead of testing each image pixel to label it as skin or non-skin (as in classic techniques), it was suggested to skip a set of pixels to improve performance. The reason of their skipping process is the high probability that neighbors of the skin color pixels are also skin pixels.

Segmentation

For our main objective, which is the emotional interaction components, finding hands would not be necessary and are not part of the whole system. However in this particular case, for testing the method of finding the user (person to interact), we combined face and hand detection algorithms with the horopter dynamic segmentation. We firstly do the dynamic background segmentation, hence it is only necessary to slide on the remaining pixels; this significantly increases the detection performance. Thus we have very fast (10 fps) results on the segmentation plus detection. Here we used the gesture recognition algorithm proposed by Rett et al. in [78], which assumes always the same default initial position for face and hands, later on the process it tracks the real position; this approach implies on performance lost during user localization. Thus, in order to save start up time, our choice was to firstly detect the face and the hands position with the algorithms previously mentioned (haarlike features and skin color hand detection) and give this as input to the gesture recognition algorithm.

The red oval on *figure 2.18 c)* is an approximation of the search region. It is observable on the *right c)* image that there are areas with skin color on the wall and floor, so if the full image was passed to the hand algorithm hand false positives would certainly occurs. Furthermore similar errors could happen for the face algorithm if the background was strongly and randomly featured.

2.3.5 Dynamic Background Segmentation Results

When using color tracking schemes, the tracker sometimes loses the target by means of generating false positives for body part identification; this is due to multi-colored backgrounds. Thus, by applying the geometric horopter technique to the system used in [78] was able to reduce the search area within the image. The perfect scenario occurs when a perfect bounding box around the human silhouette is generated, as it was theoretically represented on figure 2.13b. There is no shadow effect on this kind of segmentation, since shadows are 2D and will be considered according to the distance of the plane they are projected, the errors noticeable on figure 2.18 takes place due to the small number of correlated points we set up to guaranty a real time application runnable in a ordinary computer. Moreover, homogeneous regions like the illuminated white wall might also generate some erroneous correlated pixels (2.3.3). The algorithm slowed its tracking computational time, from deploying 15 frames/second to 10 frames/second, which is not considered critical, as 10 frames is still a good rate. This happened because the old version used one camera only, and after the application of this method, most processing time is dedicated to the computation of the depth image. However tracking results increased dramatically, by reducing the tracking false positives in 87%. To strengthen our tracking rate, geometric constraints were also applied. The results of movement classification are out of the scope of this work and hence, will not be discussed.

Dynamic background segmentation is a good strategy to reduce the false positives of several algorithms that are based rather on pixel color or features. By reducing the scope of the searching image to an *zone of interaction* area, the applications of the dynamic background segmentation we proposed here are wide open on the field of Social Robots. In all the cases (haar like features face detection, gesture recognition, facial expression recognition), our dynamic background segmentation approach shown to improve the performance **and** the results.

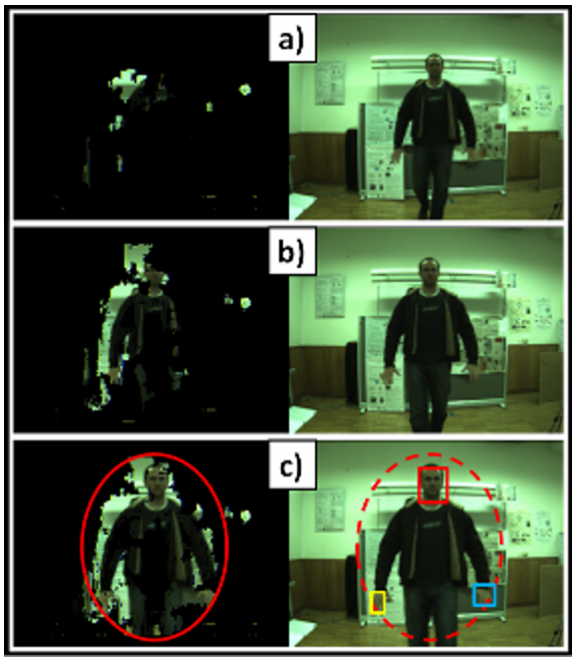


Figure 2.18: a) and b) Subject entering in horopter, consequently entering in the field of view of the robot. c) subject is inside the horopter and thus have his face and hands localized.

2.4 Auditory Sensory Processing / Feature Extraction

2.4.1 Overview

The initial step in auditory emotion recognition is the feature extraction. In our interaction, a human speaks a phrase on the microphone, each phrase is recorded on a wav file (we used wave files of the format wav: the most common wav format contains uncompressed audio in the linear pulse code modulation format). Starting from the wav file, it is possible to detect some parameters that arise as a result of the wave signal characteristics. These characteristics are involuntarily performed by the human when a certain emotion affects the voice.

The audio feature extraction takes place when the human spoke a phrase on the microphone. The distance used to capture the sounds was approximately 10cm (head set that can be wireless). The uttered phrases, at this test, were the following:

1. Human: How are you?

Table 2.4: DESCRIPTION OF VARIABLES EXTRACTED FROM SOUND.

Variable	Description
SDur	Stands for Sentence Duration: Since we know the sampling frequency ($sfreq$) of the acquired sound, we also know the beginning and the end of each sentence, and consequently the number of samples ($nsam$); and then it is simple to determine the duration in seconds by $SDur = nsam/sfreq$.
SPit	Stands for Sentence Pitch: Pitch represents the perceived fundamental frequency of a sound. The pitch extraction was done by autocorrelation method [89].
SEne	Stands for Sentence Energy: This variable is actually the *energy* or *intensity* of the signal, which for a theorically continuous-time signal $x(t)$ is given by $SEne = \int \|x(t)^2\|dt\|$.

2. Robot: I am fine and you?

3. Human: I am also ok.

From this short dialogue, in this particular experiment, we analyzed several times just the sentence number 3. To guarantee that the subject uttered the phrase with the same emotion in all cases, the user in this case was trained by listening several times a database of sentences produced by vocal actors, a psychological assessment on this same base of sentences was performed by Paixao et. at. in [64]. If nothing is said, the process does not fail, in the current configuration robot is waiting 10 seconds until the human speaks something. If after the 10 seconds the human does not speak, robot says a message “why are you so quiet? I feel abandoned, I am sad now”, and becomes sad. If the human speaks as expected, then each second of the phrase is recorded on a different wav file. The phrase is then mounted together (all pieces of 1 second each) and re-recorded in a wav file.Finally it is necessary to detect 3 variables from each wav file. These variables are described on table 2.4.

Praat toolkit [7] is used by us, in order to detect these evidences. This toolkit contains scripts that are easily integrated with other coding languages. After detecting these features (As for example in figure 2.19), they serve as input to our vocal expression Bayesian classifier that will be explained on section 3.3. The Bayesian classification model for auditory perception will be presented in next chapter (chapter 3).

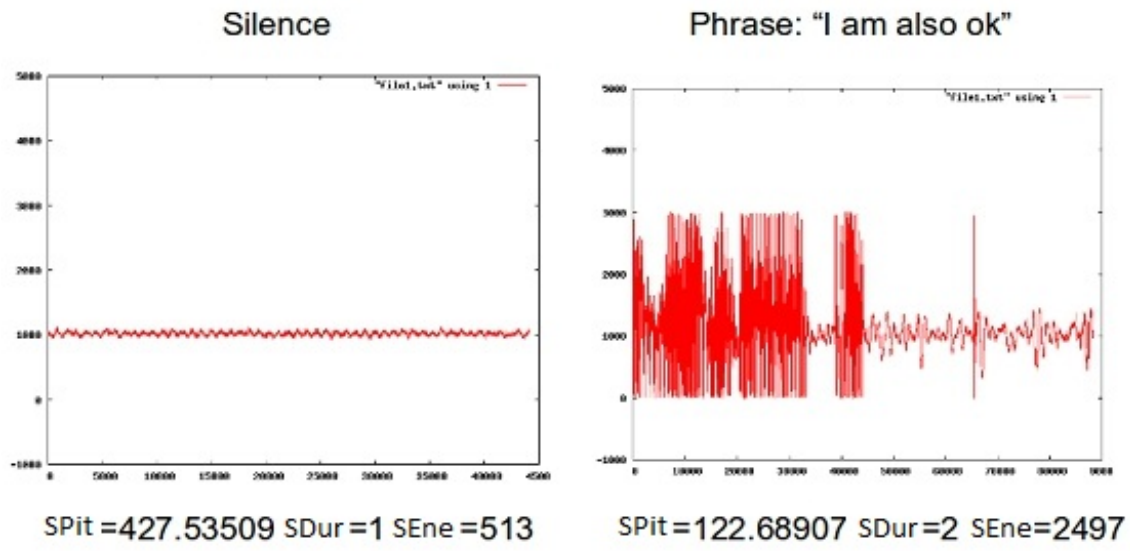


Figure 2.19: Samples results of our method for auditory perception/feature extraction. Left image presents 1 second of silence. Right image presents a 2 seconds phrase “I am also ok”. Notice that, on both waves, the amplitude of the signal is limited by both a maximum and a minimum thresholds. Respectively under each of the images: the detected audio features given by equations and methods referred in table 2.4.

2.4.2 Assessments for the Auditory Sensory Processing

For the sound it was measured using the *standard deviation* (σ_X), *mean* ($\overline{X}$) *and median* ($\widetilde{X}$) where X is a random variable standing for each of our variables (*SPit*, *SDur* and *SEne*) during 100 iterations. Each group of 100 iteration were done in 3 different environments keeping the same phrase, the same performed emotion (neutral) and the same user. To characterize the classes of environments, the amount of decibels (dB) was measured for each sample of the background noise and are presented in table 2.5, the voice of the user was also measured and the average intensity of the voice was 56 dB.

2.4.3 Results for Auditory Sensory Processing

According to what was defined for our assessments, the results were collected and can be seen in table 2.6. As expected, the standard deviation in a good environment was ac-

Table 2.5: DESCRIPTION OF THE THREE CLASSES OF AUDIO ENVIRONMENTS.

Environment	Description	Background Noise	Voice/Noise
Good	Alone with the robot in a room (low background noise)	27.7 dB	2.02
Medium	Other people talking normally in the same room (some noise)	32.7 dB	1.71
Bad	Five persons were recorder while talking loudly in the same room (a lot of background noise)	42.2 dB	1.32

ceptable, because even the same person when he/she repeats the same sentence with the same vocal expression does not produce exactly the same audio signal. On the medium environment, the standard deviation increased very little; the mean and median were quite similar to the good environment since we were using a microphone close to the mouth of the user, the influence in medium environment did not significantly affect the results of the sensory processing. However in the noisy environment the standard deviation increased a lot, specially for *SDur*. That's because our phrase ends automatically with a silence detector implemented based on the signal's amplitude, this was highly disturbed by the noisy environment. Also notice that the mean and median of Energy (*SEne*) significantly increased in the noisy environment. In noisy environment, the pitch values seemed to be completely wrong for some cases because the noisy signal may be composed by peaks of frequency. Thus, we concluded that we can use our system only in good or medium environments. Thus, all experiments over our classifiers were done in a "medium environment" for the feature extractors.

Table 2.6: RESULTS FOR AUDITORY SENSORY PROCESSING.

	$abs(SPit)$			$SDur$			$SEne$		
	σ_{SPit}	$\overline{SPit}$	$\widetilde{SPit}$	σ_{SDur}	$\overline{SDur}$	$\widetilde{SDur}$	σ_{SEne}	$\overline{SEne}$	$\widetilde{SEne}$
GE	10.7	124.5	133	0	3	3	105.1	1613.9	1590
ME	15.4	126.2	133	0.3	3.10	3	153.1	1623	1590
BE	103.2	181.7	139	1.77	4.36	3	235.4	1777.5	1707

In table 2.6, GE stands for good environment, ME stands for medium environment and BE stands for bad environment. Standard deviation, mean and median of *SPit*,

SDur and *SEne* are shown, respectively, over the three different environments as defined on the assessments. One hundred tests were done for each environment, keeping the same phrase and user.

Notice that *SDur* was measured in integer seconds. The same person produced the same sentence with the same number of (integer) seconds in all repetitions. Differences that are smaller than a second were ignored. This is due to the fact that our variable (*SDur*) is being trunked in integer numbers. That's why, although small differences existed, the standard deviation in the good environment became 0 (zero) for *SDur*.

2.5 Visuovestibular-based Gaze Control Experimental Case

The main objective of the work in this thesis is to address the emotions in a dialogue between human and robot. However, the use of Bayesian algorithms is also a strong point and important in our work. Therefore, in this section we dare to go beyond the scope of emotions, by presenting our work on using Bayesian networks in a system that combines visual and inertial information. This section may seams a little out of the scope, but if serves to reinforce and exemplify the use of Bayesian networks in autonomous systems using a different example of multimodality fusion.

The study case presented in this section is about visuo-vestibular gaze control. The purpose of this experimentation was reaching to a feasible Bayesian model for a robotic gaze control following the ideas presented on [4] and [5]. From the camera image we extract Fd^t and Fa^t which are the direction and the amplitude of the mean flow. From the IMU(Inertial Measurement Unit) we get the angles (Roll, Pitch and Yaw) that are further shown in this paper as R^t (for Roll), P^t (for Pitch), and Y^t (for Yaw).

The actuator control acts based on current angular position and on instantaneous flow information of the system. The pan-tilt unit are controlled with combined commands for target position and velocity. The motors move to the desired target position with the selected velocity and stop. The motor model takes this into account by having the current motor command depending on the current state and also on the probabilistic table filled out with the human reaction information.

The following variables were be used:

- S_t: is a tuple with the following four variables transformed in possible motor reactions
  - R^t: (roll) angle of the human-reaction for a given state
  - Y^t: (yaw) angle of the human-reaction for a given state
  - Fd^t: direction of the vector of the mean flow (comes from vectored product between u and v) (Radians)
  - Fa^t: amplitude of the vector of the mean flow
- M_t: is a movement variable with the following scope (UP, DOWN, LEFT, RIGHT, STOP)
  - The five states of M_t are concluded by doing atomization of the raw values in the following variables
    - pan motor velocity: $\mathcal{P}_\omega$ — pan motor target position: $\mathcal{P}_\theta$
    - tilt motor velocity: $\mathcal{T}_\omega$— tilt motor target position: $\mathcal{T}_\theta$
- H_t: is the human reaction to be learned (UP, DOWN, LEFT, RIGHT, STOP)

To simplify notation, state variables are grouped in a vector $S = (R^{0:t}, P^{0:t}, Y^{0:t}, Fd^{0:t}, Fa^{0:t})$ and motor variables are considering to be in the range U,D,L,R,S after atomization from $M = (\mathcal{P}_\omega, \mathcal{P}_\theta, \mathcal{T}_\omega, \mathcal{T}_\theta)$. The Bayesian program that show the relation between these variables is shown in fig. 2.20.

2.5.1 Experimental Paradigm and protocols

Apparatus and Stimuli

We used a HMD (Head Mounted Device) to pass the visual stimulus from the robot to the human subject. Our robotic head (fig.2.21) is a common platform consisted by many sensors, but basically those that we used in this work were: a stereo camera, a pan-tilt unit and a inertial sensor, all sensors were attached statically and being so every motor command sent to the pan-tilt unit will reflect on IMU (Inertial Measurement Unit) and also in the camera images, consequently in the calculated optical flow and inertial data.

Each device has a correspondent software module and they were integrated with our robotic software platform derived from CARMEN [58]. CARMEN is the Carnegie

Figure 2.20: BayesianProgram

Program
- Description
  - Specification
    - Variables:
    - $S^{0:t} = (R^{0:t}, Y^{0:t}, Fd^{0:t}, Fa^{0:t})$: state variable
    - $M^{0:t}$: motor movement variables with scope {U,D,L,R,S}
    - $H^{0:t}$: human reaction variables with scope {U,D,L,R,S}
    - $Fd^{0:t}$: flow direction (Rad)
    - $Fa^{0:t}$: flow magnitude (pixels)
    - Decomposition:
    - $P(S^t \wedge M^t \wedge H^t \wedge Fd^t \wedge Fa^t =$
    - $P(R^t.P(y)^t)$
    - $.P(H|R \wedge y)$
    - $.P(Fd|H \wedge R \wedge y)$
    - $.P(Fa|Fd \wedge H \wedge R \wedge y)$
    - $.P(M|Fa \wedge Fd \wedge H \wedge R \wedge y)$
    - Parametric forms:
    - $P(M^t) = P(H^{0:t}|S^t)$
  - Identification:
  - Bayesian human-learned gaze control.
- Question:
- $P(M^t|S^t \wedge H^{0:t})$

Mellon Robotic Toolkit, it is an open-source collection of software for mobile robot control. CARMEN is modular software designed to provide basic navigation primitives including: base and sensor control, logging, obstacle avoidance, localization, path planning, and mapping.

Subjects

Five human subjects with normal working visual and vestibular systems. In those subjects with visual distortion this should be compensated by using glasses or lens, thus the distortion perform no impact to the experiment. We mixed male and female. The subjects were four naive and one author.

Figure 2.21: Robotic Head

Protocol

By using the HMD (virtual reality), we gave to the human eyes the robot images, the visual connection becomes direct between the robot and the human. We can not inject artificial inertial sensor data into human brain. Thus, what is possible to be done is a indirect correlation where during the tests, the human will use it's own vestibular system while the robot will use the artificial inertial system. Gray scale images are the input, with several visual detectable features on the environment. Visual features are necessary by the human brain to have notion of motion. If a human is moving side by side in front of a perfectly white wall, once the acceleration stabilizes subject will have no sensation of motion. However if this same wall is full of visual detectable features, human will naturally detect the motion only by the visual influence. We have the same response on artificial optical flow algorithms, and that's why we are interested on considering the optical mean flow as a artificial visual ego-motion notion measurement variable.

2.5.2 Results

A first version of human-learning was implemented, using keyboard to control the robot while monitored by a human (human in the loop way of learning like in [68]). We still want to improve our way of learning by developing a helmet equipped with camera and IMU and then detecting real human neck movements.

Consider that we numbered our random variables as follows:

1. is $Roll^t$, subvariable of Imu^t variable

2. is $Pitch^t$, subvariable of Imu^t variable

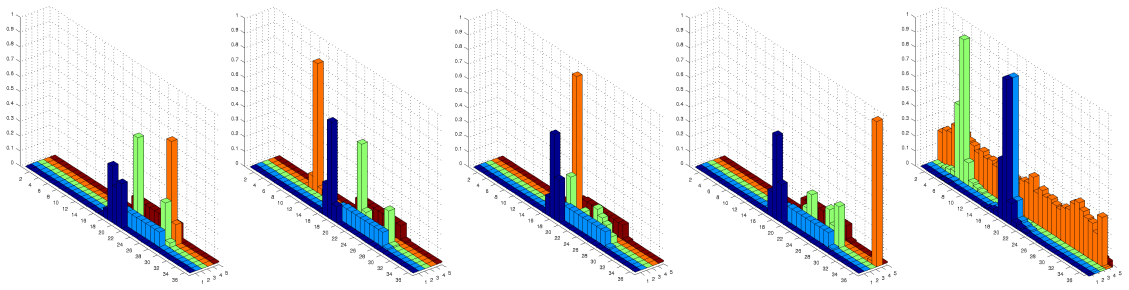


Figure 2.22: Probability Table - Learned data for UP, DOWN, LEFT, RIGHT and STOP movement.

3. is Yaw^t, subvariable of Imu^t variable
4. is Fd^t, the Flow Direction
5. is Fa^t, the Flow Amplitude

The learned table is a 4D probability table with dimensions [36x5x10x5] in our test, we plotted this in five 3D graphics in order to be possible to visualize them.

It is possible to observe that for the UP, DOWN, LEFT and RIGHT movements, the main categorizing random variable is Fd^t, in the other hand for the STOP movement Fd^t is very confusing, thus Imu^t will be much more useful categorizing this decision.

Testing the reaction of the system

Fake stimulus were injected into the system to measure if the robot's reaction would be like expected for that stimulus. As human trained the system, we know (approximately) which stimulus to create and which reaction to expect. For example if we train a walking robot not to fall from a step, we can put a step in front of it and our expectation will be that the robot do not fall. In our case we trained the head to be centered and then we give stimulus simulating that the head would moving to one or another side "forever" during each test. We also gave stimulus for the system to believe the head was flying up like a rocket and also falling down in free fall. It was performed 100 trials with different stimulus for each expected reaction.

One stimulus for each reaction would be the trivial case to categorize, but for this preliminary results we had approximately 98% of correct decisions in 500 different stimulus to be categorized in 5 movements.

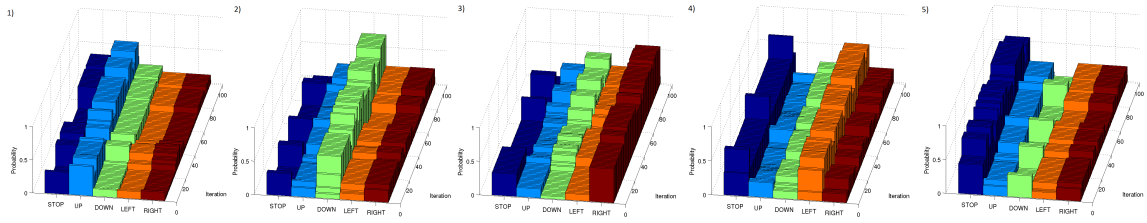


Figure 2.23: System reaction - 1) Falling down in free fall "forever" (simulation) 2) Launched up as a rocket to the sky "forever" (simulation) 3) Translating to left side "forever" (simulation) 4) Translation to right side "forever" (simulation) 5) Stop (Imu^t vary even stop because of Mag^t mainly).

2.6 Conclusions

In this chapter, methods for feature extraction were presented. As a initial step, a method of face attention was described in section 2.1. This method has as the objective to keep the robot targeting the camera to the user. The face attention method was also published by us in [87]. The Action Units extraction from face images was explained in section 2.2, and the process of producing the pseudo-face with the detection of important points was addressed. The pseudo-face method was detailed described, and submitted to publication as mentioned in: Chapter 1, section 1.8, point 8. Furthermore, the Action Unit detection method was used, but not detailed described, in our work [75]. In section 2.3, a dynamic background segmentation was proposed. It has as the objective to remove the background of the image capture by the camera of a mobile robot. This method was published by us in [73]. Procedures for auditory feature extraction were described in section 2.4. This extraction methods were used in our recent works, as in [76]. Moreover, section 2.5 presents an example of multimodality fusion by Bayesian models, it applies for a visuovestibular-based gaze control experimental case.

Summarizing, this chapter covered methods for: the robot to keep engaged to the human face, the robot to be able to detect the action units from face, the robot can subtract the background to avoid noisy input images, the robot can extract features from audio. Moreover, an example of a Bayesian system was presented. At the next chapters, Bayesian networks are going to be strongly present among our classifiers and synthesizers.

Chapter 3

Modeling for Emotion Analysis

3.1 Overview

After detecting all the features from audio signal, and from the visual signal (image), in this chapter we are going to explain what to do with these features in order to achieve a classification. Two dynamic Bayesian networks are presented in this chapter, one for auditory emotion classification, another for visual emotion classification. The purpose of the multimodal emotion analysis is that each channels complements the other, if the recognition of one fails, the other can sustain the correct classification and vice-versa. As shown in figure 3.1, the whole process starts with the feature extraction that was presented in the previous chapter. Later, each one of the dynamic Bayesian classifiers is able to convert the feature vector, of it's respective modality, into a histogram of 5 confidence scores (probabilities). To clarify, lets state that we call *dynamic Bayesian classifier* a classifier that uses a dynamic Bayesian network. The scope of both classifiers are among the five considered expressions *{Anger, Fear, Sad, Happy, Neutral}*.

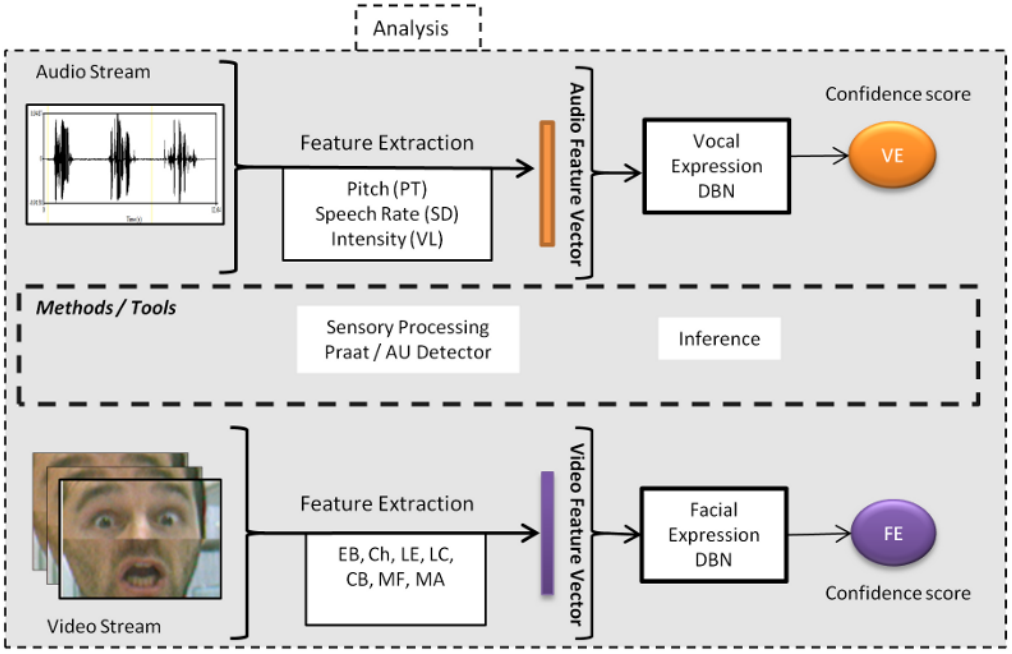


Figure 3.1: The analysis structure: Notice that two modalities signals are processed and then classified.

3.2 Modeling for Analysis of Facial Expression

3.2.1 Facial Expressions Classification dynamic Bayesian network

Once the feature extraction problem is solved, our robot must then classify the facial expressions. The *dynamic Bayesian network* (figure 3.2) is proposed to classify the facial expressions performed by the human interlocutor. Based on the psychological definitions found in [20], we mounted table 3.1 that presents the considered Action Units and association to each facial expressions.

When designing the classifier, we made our own interpretation of FACS (Facial Action Coding System) [20], which drives a set of random variables different from those defined by other researchers. Each facial expression is composed by a specific set of Action Units. Each of these Action Units is a distortion on the face induced by small muscular activity. Normally, a well determined set of face muscles is associated to a specific Action Unit, which can give the idea that all these basic distortions are independent. Nevertheless, some of these Action Units are antagonistic. One clear and understandable

Table 3.1: DISCRIMINATION OF THE AUs THAT ARE PRESENT IN EVERY ONE OF THE FACIAL EXPRESSIONS.

	Upper Face			Lower Face			
	EyeBrows	Cheeks	Lower Eyelids	Lip Corners	Chin Boss	Mouth Form	Mouth Aperture
Neutral	-	-	-	-	-	-	-
Happy	-	AU6	-	AU12	-	-	AU25
Sad	AU4	-	-	AU15	AU17	-	-
Fear	AU1	-	-	-	-	AU20	AU25
Anger	AU4	-	AU7	-	AU17	AU23	AU24

example is the case of two Action Units related with the lips corners: AU12 and AU15. When performing AU12, lips corners are pulled up. Oppositely, when performing the AU15, the lip corners are pulled down. Therefore, if by one way the movements of the lip corners can be considered independent because they are performed by distinct muscle sets, by another, when analyzed visually they are antagonistic, and exclusive.

We assumed that the state space is discrete, and in this case, hidden Markov models (HMM) can be applied. As defined by Pearl at [35], hidden Markov model can be considered to be an instantiation of a dynamic Bayesian network and though exact inference is feasible. Based in these principles, belief variables were defined and a dynamic Bayesian classifier of facial expressions was developed.

Facial Expression dynamic Bayesian network We took advantage of the antagonism extant in some AUs to reduce the size of the dynamic Bayesian network. Though, instead of using the 11 AUs as leafs for our DBN (Dynamic Bayesian Network), we propose 7 (seven) variables. These variables groups the related antagonist and exclusive Action Units. The structure of the network of two levels is illustrated in figure 3.2, also in this figure, the time influence that characterizes this network as a dynamic Bayesian network is represented. Similar representation of DBNs were used by Martinez and Sucar at [55], where dynamic naive Bayesian classifiers (DNBC) are proposed for gesture recognition, and also by Kafai and Bhanu at [37], where dynamic Bayesian networks for vehicle classification in video is proposed.

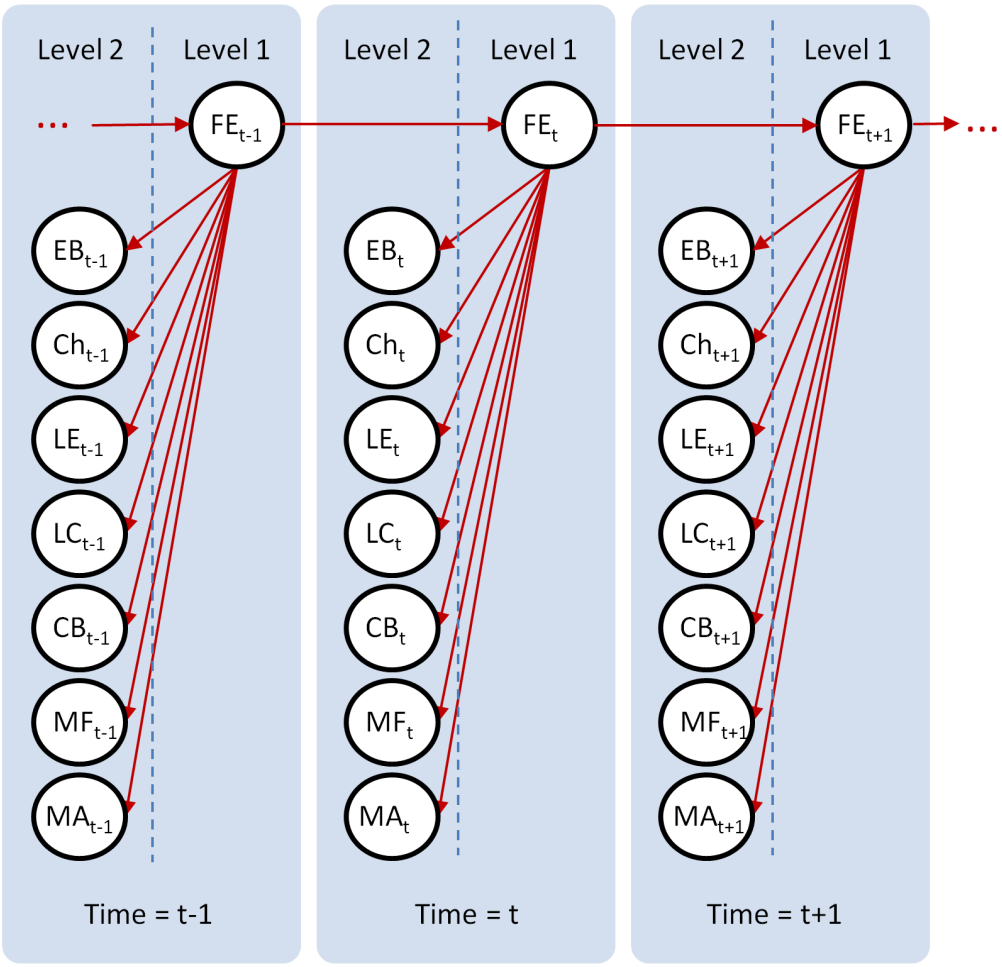


Figure 3.2: Facial Expression Dynamic *Bayesian network, t*hree time intervals are shown ($t-1, t, t+1$).

In the *dynamic Bayesian network*'s first level there is only one node. The global classification result obtained is provided by the belief variable associated to this node: $FE \in \{Anger, Fear, Sad, Happy, Neutral\}$, where the variable name stands from Facial Expression. Considering the structure of the dynamic *Bayesian network*, the variables in their second level have as parent this one in the first level: FE.

In the second level there are seven belief variables:

- $EB \in \{AU1, AU4, none\}$ is a belief variable related with the *Eye-Brows* movements. The events are directly related to the existence of AU1, and AU4.

- $Ch \in \{AU6, none\}$ is a belief variable which is related with *Cheeks* movements; more specifically, the events indicate if the cheeks are raised (AU6 is performed).
- $LE \in \{AU7, none\}$ is a belief variable which is related with the *Lower Eyelids* movements; AU7 is associated to lower eyelids set to up.
- $LC \in \{AU12, AU15, none\}$ is the belief variable associated with the movements of the *Lips Corners*. When the corners did not perform any movement then the event *none* has a high probability. The event *AU12* has a big probability when the corners of the lips are pulled up. If the lip corners moves down the event *AU15* must have a big probability.
- $CB \in \{AU17, none\}$ is the belief variable collecting the probabilities related with the *Chin Boss* movements. The event *none* is related with the absence of any movement, while the event *AU17* had a great probability when the chin boss is pushed upwards.
- $MF \in \{AU20, AU23, none\}$ is the belief variable associated with the Mouth's Form. The events *AU20* and *AU23* indicated, respectively, if the mouth is horizontally stretched or tightened.
- $MA \in \{AU24, AU25, none\}$ is the belief variable associated with the Mouth's Aperture. The events *AU24* and *AU25* are related, respectively, with lips pressed together or with lips relaxed and parted.

The movements performed by the human in one area of the face can slightly affect muscles on other area, however, this influence is very small and cannot be detected by the cameras of the robot. Thus, conditional independence among the 7 proposed variables was assumed.
The following equations illustrate the joint distribution associated to the Bayesian Facial Expressions Classifier.

$$\begin{aligned} &P(FE,EB,Ch,LE,LC,CB,MF,MA) = \\ &P(EB,Ch,LE,LC,CB,MF,MA|FE) * P(FE) = \\ &P(EB|FE) * P(Ch|FE) * P(LE|FE) * P(LC|FE) * P(CB|FE) * P(MF|FE) * P(MA|FE) * P(FE) \end{aligned} \tag{3.1}$$

The last equality is written assuming that the belief variables in the second level of the dynamic *Bayesian network* are independent.

From the joint distribution, the *posterior* can be obtained by the application of the Bayes rule as follows:

$$P(FE|EB,Ch,LE,LC,CB,MF,MA) = \frac{P(EB|FE)*P(Ch|FE)*P(LE|FE)*P(LC|FE)*P(CB|FE)*P(MF|FE)*P(MA|FE)*P(FE)}{P(EB,Ch,LE,LC,CB,MF,MA)} \quad (3.2)$$

From the Bayesian marginalization rule we can calculate:

$$P(EB,Ch,LE,LC,CB,MF,MA) = \sum_{FE} P(EB|FE)*P(Ch|FE)*P(LE|FE)*P(LC|FE)*P(CB|FE)*P(MF|FE)*P(MA|FE)*P(FE) \quad (3.3)$$

3.2.2 Inference Learning

After defining the structure of the Bayesian network, it is necessary to provide the probabilities: priors and likelihoods. For the facial expression classifier these likelihoods are provided as histograms tables. To build these distribution tables the Cohn-Kanade database was used [39]. The initial prior was defined through a uniform distribution. As there are five events associated to the belief variable in the Bayesian networks' first level, the priors are $P(FE = anger) = ... = P(FE = neutral) = 0.2$. These priors are changed dynamically: that is, systematically, after each classification, the posterior is transformed in the new prior of the Bayesian network. As opposite to what appends with prior, the likelihood was not changed dynamically and remained the same over time. In table 3.2 is presented an example of the likelihood histogram obtained as a result of the learning.

Table 3.2: FACIAL EXPRESSION LIKELIHOODS.

		FE				
		Anger	Fear	Happy	Sad	Neutral
EB	none	0.01	0.01	0.97	0.01	0.97
	AU1	0.01	0.01	0.01	0.01	0.01
	AU4	0.97	0.01	0.01	0.01	0.01
	AU1+4	0.01	0.97	0.01	0.97	0.01
Ch	none	0.99	0.99	0.01	0.99	0.99
	AU6	0.01	0.01	0.99	0.01	0.01
LE	none	0.01	0.99	0.99	0.99	0.99
	AU7	0.99	0.01	0.01	0.01	0.01
LC	none	0.98	0.98	0.01	0.01	0.98
	AU12	0.01	0.01	0.98	0.01	0.01
	AU15	0.01	0.01	0.01	0.98	0.01
CB	none	0.01	0.99	0.99	0.01	0.99
	AU17	0.99	0.01	0.01	0.99	0.01
MF	none	0.01	0.01	0.98	0.98	0.98
	AU20	0.01	0.98	0.01	0.01	0.01
	AU23	0.98	0.01	0.01	0.01	0.01
MA	none	0.01	0.01	0.01	0.98	0.98
	AU24	0.98	0.01	0.01	0.01	0.01
	AU25	0.01	0.98	0.98	0.01	0.01

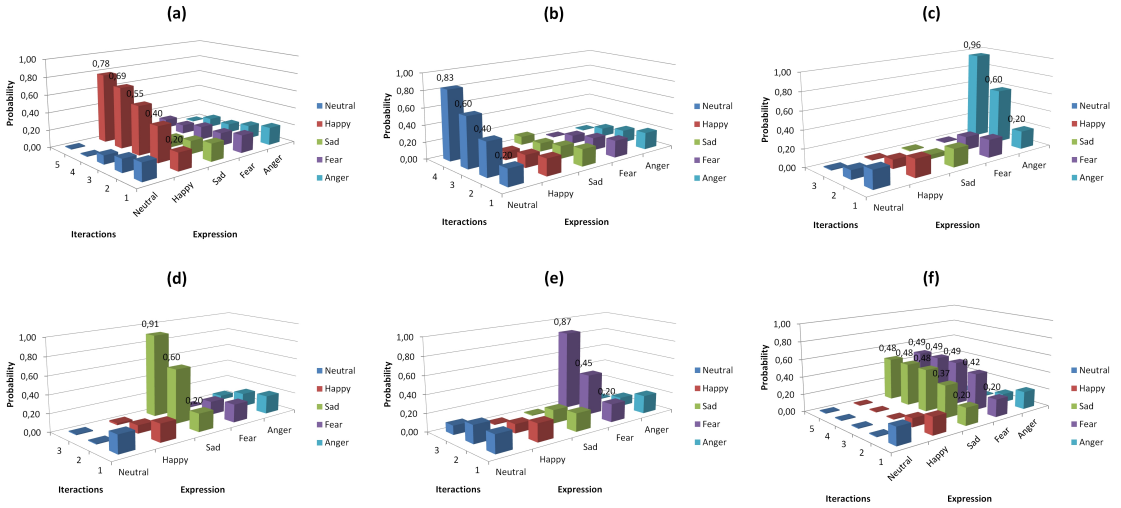


Figure 3.3: Results from facial expression classifier: camera grabbing was set to 5 fps, therefore, the iteration axis represents the 5 (or less) utterances that happens inside one second. The expression axis is the selected scope of possible expressions. Notice that the sum of probability at each iteration among the five possible expressions is always 1. In examples (a), (b), (c), (d) and (e), respectively, inputs were given for *happy*, *neutral*, *anger*, *sad* and *fear;* the dynamic Bayesian network was capable of classifying the expected expression with a fast convergence. In (f), an example of ambiguity and misclassification is shown, where the expected result was *sad* but the result of classification was *fear*.

3.2.3 Results of Facial expressions dynamic Bayesian network

As a consequence of network to be dynamic, convergence happens along the time, the resultant histogram from the previous frame is passed as prior knowledge for the current frame. We limited the maximum number of frames for convergence as 5. If the convergence reach to 80% before 5 frames, the classification is considered complete. If not it keeps converging up to the fifth frame. If the fifth frame is reached and no value is higher than 80%, the classifier selects the highest probability value (usually refered as the *"maximum a posteriori decision"* in Bayesian theory) as a classified result.

Expected results for the *analysis* part are a correct classification of facial and vocal expressions according to what is expected. Convergence is also expected to appear as time passes, since the *Bayesian Network* is Dynamic. Figure 3.3 shows experiments that took a few iterations with the following constant evidences:

(a) "$EB = none, Ch = AU6, LE = none, LC = AU12, CB = none, MF = none, MA = AU25$";

(b) "$EB = none, Ch = none, LE = none, LC = none, CB = none, MF = none, MA = none$";

(c) "$EB = AU4, Ch = none, LE = AU7, LC = none, CB = none, MF = AU23, MA = AU24$";

(d) "$EB = AU4, Ch = none, LE = none, LC = AU15, CB = none, MF = none, MA = none$";

(e) "$EB = AU1, Ch = none, LE = none, LC = none, CB = none, MF = AU20, MA = AU25$".

Notice that the convergence happened fast, after the second iteration the best result was already visible. The classification was considered completed when the percentage was higher than 80% for one of the expressions, or when it reached 5 iterations. Usually the convergence happened in less than 5 iterations, like in examples of figure 3.3 (b), (c), (d) and (e). A misclassification is presented on figure 3.3 (f), the expected expression was "*sad*", however it was a case where the *sensory processing* phase failed, thus it became ambiguous between "*sad*" and "*fear*" and the result was a misclassification to "*fear*".

3.2.4 Developed tools for assessment of classifiers

To be suitable for the comparison methodology proposed in [67], we needed first to measure *the percentage of correct classifications* (hit-rate) for both classifiers (audio and video). This was done by comparing the classified result to the expected result. We created graphical interfaces (see figure 3.4) to help accomplish this task. On those interfaces, for both audio and video; the system tester could see the result of classifications on real time, and click on what he/she expected as the result. When the system tester clicked on the expected result; a benchmark routine saved both the real time classification and the expected expression in a file, for further statistical calculation of *the percentage of correct classifications* (hit-rate).

3.2.5 Benchmark Over the Facial Expression Emotion Classifier

The system was initially trained for the possible facial expressions. Then, according to what was defined on the assessments, a battery of tests was performed over the facial expression classifier. The results are shown in figure 3.5. Notice that our classifier is

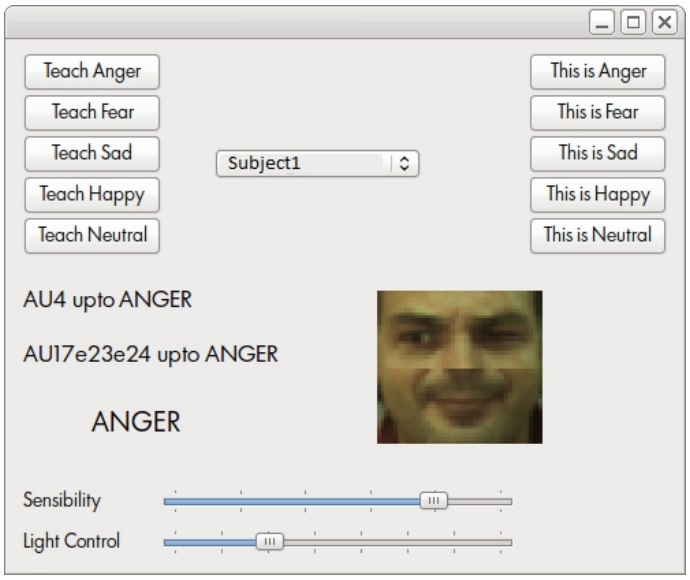


Figure 3.4: QT GUI for evaluation of the facial expression classifier.

dependent of the AU Detector, which for instance is dependent of the OpenCV haar-like features face detector. Nevertheless, the results were higher than 79% correct.

3.2.6 Assessment of Automatic Emotion Recognition from Face Images

Since there is no common benchmark for systems of "automatic emotion recognition from face images", we will use here a comparison methodology proposed on [67] to show the advantages of our classifier. The following questionnaire is used in table 3.6, this questionnaire was defined in [67].

1. Is the input image provided automatically?
2. Is the presence of the face assumed?
3. Is the performance independent of subject's sex, physiognomy, age, and ethnicity?
4. Can variations in lighting be handled?
5. Can rigid head movements be handled?
6. Can distractions like glasses and facial hair be handled?
7. Is the face detected automatically?

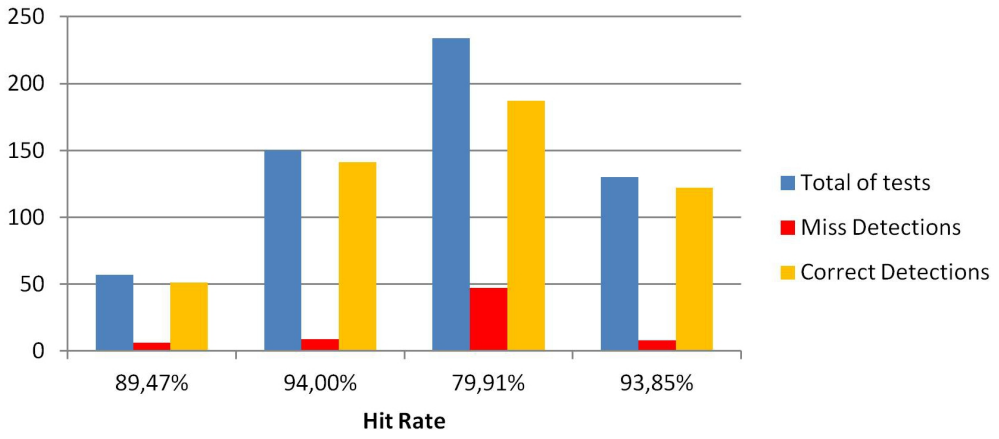


Figure 3.5: Four batteries of tests were done with clicking on the expected classification while saving the current classification. In these four batteries, the expressions on all of them were randomly chosen, the only difference between them was that they were performed in different days, so illumination conditions changed. The average percentage of correct classifications was 89.27%.

8. Are the facial features extracted automatically?

9. Can inaccurate input data be handled?

10. Is the data uncertainty propagated throughout the facial information analysis process?

11. Is the facial expression interpreted automatically?

12. How many interpretations categories (labels) have been defined?

13. Are the interpretation labels user profiled?

14. Can multiple interpretation labels be scored at the same time?

15. Are the interpretation labels quantified?

16. Is the input processed in fast or real time?

3.3 Modeling for Analysis of Vocal Expressions

3.3.1 Overview

In this section we focus on auditory analysis as the sensory stimulus. Our goal scenario is to have one robot interacting with one human through the vocalization channel. Notice that vocalization is far beyond speech; while speech analysis would give us what was said, vocalization analysis gives us how was said. A social robot shall be able to perform actions in different manners according to its emotional state. Thus we propose a novel Bayesian approach to determine the emotional state the robot shall assume according to how the interlocutor is talking to it. Results on section 3.3.5 shows that the classification happens as expected converging to the correct decision after two iterations. This research focus was also explored by us in [76].

3.3.2 Variables

The set of variables mentioned on table 3.3, compose the auditory feature vector, and they are detailed described as follows:

• PT *{low, normal, high}:* This random variable contains 3 possible categories for the input coming from the continuous variable *SPit*, defined in section 2.4. The voice pitch changes significantly along a sentence and an important part of our voices are unpitched, however, since the conversation follows a pre-defined story board, the mean pitch of the sentence will help to distinguish the emotional state that was there when this very sentence was spoken.

• SD *{short, normal, long}:* Is a random variable that contains 3 possible categories for the input coming from the continuous variable *SDur*, defined in section 2.4. This variable contributes to the classification: Ex. When a person speaks the same sentence with a happy emotion it usually speaks faster than with a sad emotion. For some emotional states the duration might be exactly the same, but then the other variables will contribute for the disambiguation.

• VL *{low, normal, high}:* Stands for Volume Level. It is a random variable that

Table 3.3: DISCRIMINATION OF THE AUDITORY VARIABLES THAT CHARACTERIZE THE VOCAL EXPRESSIONS.

Variable	Stands For	Scope				
VE	Vocal Expression	Anger	Fear	Happy	Sad	Neutral
PT	Pitch	low	normal	high	low	normal
SD	Sentence Duration / Speech Rate	normal	long	normal	short	normal
VL	Volume Level / Energy	high	normal	high	low	normal

contains 3 possible categories for the input coming from the continuous variable *SEne*, defined in section 2.4.

3.3.3 Model

According to Bayesian theory, Bayesian models are characterized by assigning probabilities to any degree of belief about the state of the world. Bayesian statistics sets how new data should be combined with prior beliefs and how data from several modalities should be integrated. It is defined in Bayesian decision theory the manner of combining beliefs with our objectives to make optimal decisions.

In our model of auditory perception, the vocalization analysis classifies a vocal expression. The robot needs to be capable of classifying among the possible vocal expressions, which are in the same scope as the facial expressions: *{Anger, Fear, Happy, Sad and Neutral}*.

In [50, 49, 48] vocal tract length normalization was extensively studied and the pitch feature was used for it. There are several methods [89], [103], [11], [30] to extract pitch: zero-crossing, autocorrelation function, cepstrum, average magnitude differential function, comb transformation, FIR filter method of periodic prediction.

To classify the vocal expressions performed by the human, a dynamic Bayesian network was developed. The structure of this network of two levels is illustrated in figure 3.6. A vocal expression will be classified after a sentence finish. In other words, for the

dynamic Bayesian network, the time 1 is just after sentence 1 is completed, time 2 is just after sentence 2 is completed; and so on. This is independent of each sentence real time duration. The state switching is not currently necessary since the sentences are recorded separated. In order to switch the state, we are planning to use a silence detector according to what was presented in [31]. If the silence period is bigger than 3 seconds, the state is then switched.

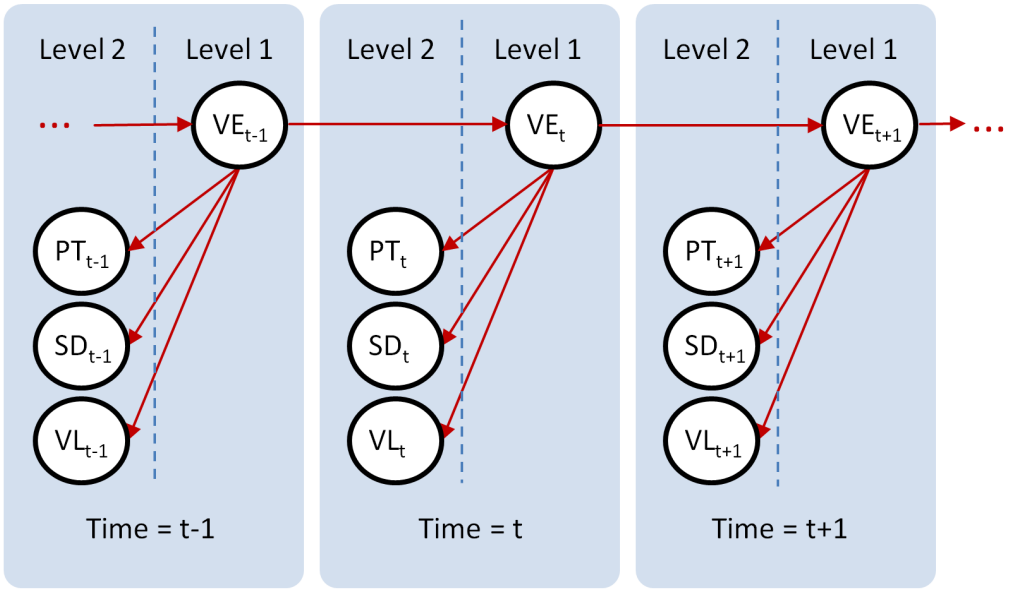


Figure 3.6: Dynamic *Bayesian network* for *Auditory Perception, t*hree time intervals are shown $(t-1, t, t+1)$.

The global classification result obtained is provided by the belief variable associated with the top level node: $VE \in \{Angry, Fear, Happy, Sad, Neutral\}$, where the variable name stands from Vocal Expression.

The joint distribution is illustrated by the following equation:

$$\begin{aligned} &P(VE,PT,SD,VL) = \\ &P(PT,SD,VL|VE) * P(VE) = \\ &P(PT|VE) * P(SD|VE) * \\ &P(VL|VE) * P(VE) \end{aligned} \tag{3.4}$$

The last equality can only be done if it is assumed that belief variables PT, SD

Table 3.4: VOCAL EXPRESSION LEARNED HISTOGRAM LIKELIHOODS.

		VE				
		Anger	Fear	Happy	Sad	Neutral
PT	high	0.10	0.34	0.10	0.80	0.10
	normal	0.10	0.33	0.80	0.10	0.80
	low	0.80	0.33	0.10	0.10	0.10
SD	long	0.80	0.10	0.25	0.10	0.10
	normal	0.10	0.80	0.50	0.10	0.10
	short	0.10	0.10	0.25	0.80	0.80
VL	high	0.80	0.34	0.10	0.10	0.10
	normal	0.10	0.33	0.80	0.10	0.80
	low	0.10	0.33	0.10	0.80	0.10

and VL are independent.

The *posterior* can be obtained, as follows:

$$P(VE|PT,SD,VL) = \frac{P(PT|VE) * P(SD|VE) * P(VL|VE) * P(VE)}{P(PT, SD, VL)} \tag{3.5}$$

From the marginalization rule we can calculate:

$$P(PT, SD, VL) = \sum_{VE} P(PT|VE) * P(SD|VE) * P(VL|VE) * P(VE) \tag{3.6}$$

3.3.4 Inference Learning

Bayesian systems require that probabilities (priors and likelihoods) are provided. The histogram table of likelihoods for vocal expression analysis are presented in table 3.4. There are a considerable number of probabilities whose value is near zero but not null. It happens because we "force" it during the learning phase, when the histograms are built. The justification for this procedure is the following: "*it is considered that, if a occurrence is not observed in the learning phase it is because it has a low probability, not because it is impossible*".

3.3.5 Results of dynamic Bayesian network for Auditory Perception

The robot is able to infer over the likelihoods when interacting with the user. The expected results for the *analysis* part are a correct classification of facial and vocal expressions. Convergence is also expected to appear across time, since both *Bayesian Networks* are Dynamic.

Figure 3.7 shows results of the Bayesian inference during some iterations with the following constant evidences:

(a) "$Pitch = 140.264309$, $SentenceDuration = 2$, $VolumeLevel = 1260$";

(b) "$Pitch = 136.569794$, $SentenceDuration = 3$, $VolumeLevel = 1170$";

(c) "$Pitch = 120.473537$, $SentenceDuration = 2$, $VolumeLevel = 2147$";

(d) "$Pitch = 138.326496$, $SentenceDuration = 2$, $VolumeLevel = 865$";

(e) "$Pitch = 137.345883$, $SentenceDuration = 4$, $VolumeLevel = 1477$".

Notice that the convergence happened fast, after the second iteration the best was already visible. Usually the convergence happened in less than 5 iterations, like in examples of figure 3.3 (a), (c), (d) and (e). A misclassification is presented on figure 3.3 (f), the expected expression was "*happy*", however it was a case where the *sensory processing* phase failed, thus it became ambiguous between "*happy*" and "*neutral*" and the result was a misclassification to "*neutral*".

3.3.6 Assessment for Automatic Emotion Recognition from Audio Signal

For the automatic emotion recognition from audio signals, we lie on a lack of common benchmark when trying to assess this kind of systems. Thus, we used a comparison methodology proposed on [67] in order to show the advantages of our classifier. The following questionnaire is used in table 3.5, this questionnaire was defined in [67].

1. Can non professionally spoken input samples be handled?
2. Is the performance independent of subject's sex, physiognomy, age, and ethnicity?
3. Are the auditory features extracted automatically?

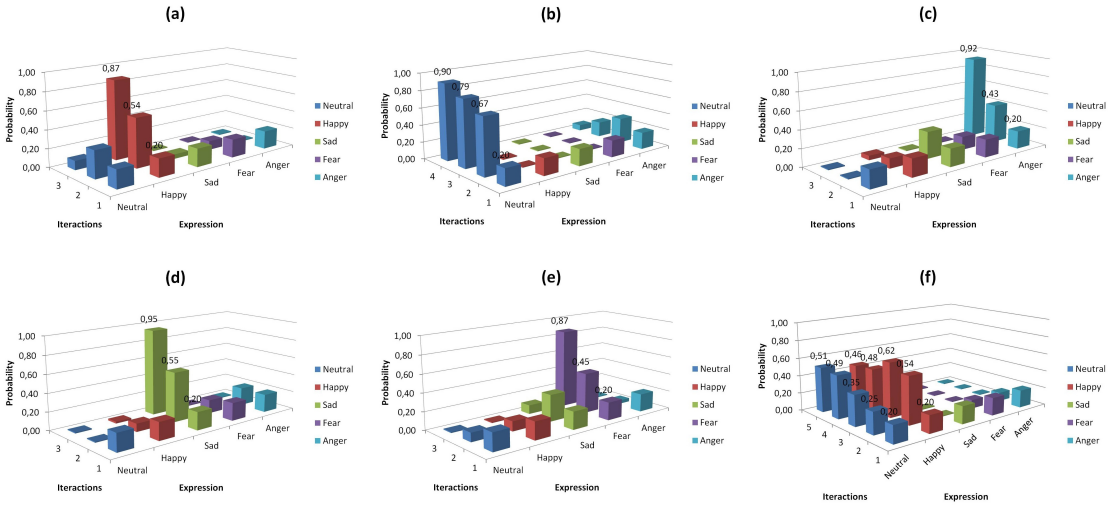


Figure 3.7: Results from analysis of vocalization: the sentence sound is recorded and divided second by second, thus, the iteration axis represents the 5 (or less) utterances that happen inside one sentence. The expression axis is the selected scope of possible vocal expressions. The sum of probability at each iteration among the five possible vocal expressions is always 1. In examples (a), (b), (c), (d) and (e), respectively, inputs were given for *happy*, *neutral*, *anger*, *sad* and *fear;* the dynamic Bayesian network was capable of classifying the expected expression with a fast convergence. In (f), an example of ambiguity and misclassification is shown, where the expected result was *happy* but the result of classification was *neutral*.

4. Are the pitch-related variables utilized?
5. Is the vocal energy (intensity) utilized?
6. Is the speech rate utilized?
7. Are pitch contours utilized?
8. Are phonetic features utilized?
9. Are some other auditory features utilized?
10. Can inaccurate input data be handled?
11. Is the extracted vocal expression information interpreted automatically?
12. How many interpretation categories (labels) have been defined?
13. Are the interpretation labels scored in a context-sensitive manner (application, user, task-profiled manner)?
14. Can multiple interpretation labels be scored at the same time?

15. Are the interpretation labels quantified?
16. Is the input processed in fast or real time?

3.3.7 Benchmark over the vocal expression emotion classifier

At first, the phrase was blocked so that the same phrase was repeated several times with different intonations. This procedure was done during the learning phase, when the user repeated the same phrase 50 times. From these 50 audio files, the features were extracted and this set of features was kept as the *trained set* for the current phrase. This *trained set* belongs to the user who trained it and was used for that user. A short dialog containing 9 phrases was used to guide the experimental tests, and, thus, the training procedure was repeated for each of the phrases used during the conversation. Therefore, a total of 450 sentences were used as the *trained set*.

After the learning phase, 129 sentences were tested into four batteries of tests. Nine in the first battery of tests, twenty nine in the second, forty one in the third, and fifty in the fourth battery of tests. These tests were done over the Vocal Expression Classifier; the results can be seen in figure 3.8.

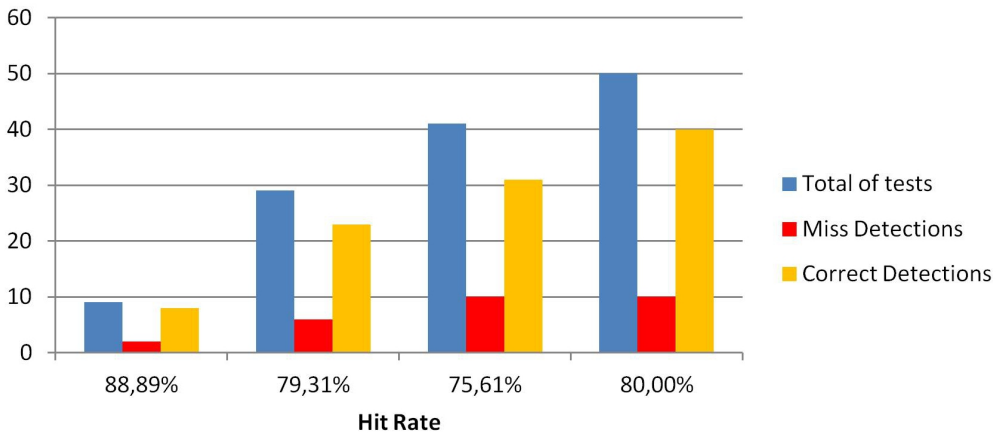


Figure 3.8: Four batteries of tests were done with clicking on the expected classification while saving the current classification. All the vocal expressions were randomly mixed during the tests while a person was speaking all the five considered possible vocal expressions. The average percentage of correct classifications is 80.92%.

Table 3.5: PROPERTIES OF STATE OF ART APPROACHES TO AUTOMATIC *EMOTION RECOGNITION* FROM AUDIO SIGNALS.

Reference	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Test results
Dimitrius' 06 [97]	×	U	•	•	•	•	•	•	×	×	•	U	×	•	•	×	Correct from 58.6% to 94.4% depending on the applied method
Nicolaou' 10[61]	×	T	•	•	•	U	U	U	•	×	•	3	×	×	×	U	Correct from 61.19% to 91.96% depending on the method and the expression
Our proposed approach	×	T	•	•	•	•	×	×	•	×	•	5	×	•	•	•	Correct 81%, mean of all expressions

Legend: •= "yes", ×= "no", U = unknown, T = handle speech samples of (known) subjects on which it has been trained.

3.4 Comparison of classifiers with state-of-art

3.4.1 Audio recognition comparison

According to Pantic et al. in [67], the auditory aspect of a communicative message carries various kinds of information. If we consider the verbal part (strings of words) only, without regarding the manner in which it was spoken, we might miss important aspects of the pertinent utterance and even misunderstand the spoken message by not attending to the nonverbal aspect of the speech. The problem of vocal affect analysis includes two subproblem areas: specifying auditory features to be estimated from the input audio signal, and classifying the extracted data into some affect categories. Pantic [67] also claims that the emotions of happiness, anger, fear, and sadness are the most commonly reported in the state of the art. The speech measures which seem to be reliable indicators of these "basic" emotions are the continuous acoustic measures, particularly pitch-related measures (range, mean, median, variability), intensity, and duration.

In order to accomplish a human-like interpretation of perceived vocal affective feedback, pragmatic choices (i.e., application, user, task profiled and time scale dependent choices) must be made regarding the selection of affective/attitudinal states and moods to

Table 3.6: PROPERTIES OF STATE OF ART APPROACHES TO AUTOMATIC *EMOTION RECOGNITION* FROM FACIAL IMAGES.

Reference	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Test results
Cohen' 02 [10]	•	•	U	U	×	U	×	•	×	×	•	7	×	×	×	×	12600 frames, 5 subjects, Correct:65%
Wuhan' 04 [101]	×	•	×	×	×	×	×	•	×	×	•	7	×	×	×	×	213 frames, 10 women, Correct 77%
Nicolaou' 10 [61]	×	•	U	×	•	×	•	•	×	•	•	6	×	•	•	×	Correct 91.76%
Pantic' 09 [66]	•	•	•	×	•	×	•	•	×	•	•	6	×	•	•	•	Correct from 61% to 93% depending on the expression
Gayatri' 08[65]	×	•	•	×	×	×	×	×	×	×	•	7	×	×	×	×	U frames, 40 subjects, 94.73 %.
Our proposed approach	•	×	•	×	•	×	•	•	×	•	•	5	×	•	•	•	560 frames, 1 subject, Correct 89.27%

Legend: •= "yes", ×= "no", U = unknown, T = handle images of subjects on which it has been trained.

be recognized by a vocal affect analyzer. Nevertheless, existing automated systems for auditory analysis of human affective feedback do not perform a context sensitive analysis of the input audio signal.

By using the comparison methodology proposed on [67], table 3.5 shows the properties of state-of-art systems and also our system. Other methods claimed to achieve a higher percentage of correct classifications, however they were not fully real-time, usually they used a database input of utterances instead of capturing directly from the microphone. Our approach captured directly from the microphone, and did every calculation in less than 1 second, being thus useful in our proposed structure and suitable for HRI applications.

3.4.2 Visual recognition comparison

By using the comparison methodology proposed on [67] and the percentage of correct classifications acquired by our tests, table 3.6 shows the properties of some of the current existent systems for the same purpose and we included our system on this table for a fair evaluation.

It is important to use this table comparison, because for example in [53] and [65], they claim to achieve a high percentage of correct classifications, however features were not extracted automatically. In [65], Gimp software distance measurement was used to manually extract the features and the Cohn-Kanade database [39] was used, so, in [65], lighting variations also cannot be handled and neither was it necessary to detect faces because all the images were already faces. In Wuhan'04 [101], the two eye pupils needed to be selected manually, no face feature extraction was done, JAFFE [38] database was used, and the system was not real time. A lot of progress was done in [66][61], and they are quite likely the state of art in this area. Our classifier is a simpler version, however, it has some advantages in some aspects that allows it to be well suited for our proposed structure. For example, we did not use a database of faces, we captured video stream from the robot camera and thus we did not assume that a face was there. We could deal with rigid head movements because the face feature extraction was done statically. The percentage of correct classifications of our facial expression classifier was acceptable and suitable for HRI applications.

3.5 Conclusions

This chapter presented and evaluated models for human emotion analysis. Section 3.2 introduced a novel Bayesian model for analysis of facial expression (published by us in [75]); while in section 3.3 a novel Bayesian model for analysis of vocal expressions was proposed (published at by us in [76]). The Bayesian networks were detailed described, also the respective equations for Bayesian inference. The learning procedures were explained in sections 3.2.2 and 3.3.4, respectively for facial and vocal expressions. Results for the classification of the proposed Bayesian networks were shown in sections 3.2.3 and 3.3.5. The Assessments, and comparison with state of the art, with similar methods,

shows the advantages of our classifiers in section 3.4. The classification methods and assessments proposal were published by us at [75]. Next chapter presents methods for synthesizing the desired expression over the robot. Moreover, fusion of the modalities into one single *human emotional state* will be explored in the next chapter, and a robotic personality emulation method will be presented and explained.

Chapter 4

Modeling for Emotion Synthesis

4.1 Overview

In the previous chapters, the feature extraction process and the classification process were covered. For the interaction be complete, the robot shall be able to express emotions back to the human. Since the system is multimodal, two channels are considered, auditory and visual channel. The auditory synthesis purpose is to endow the robot the capability to produce voice with pitch, speed and tone variations according to each emotion. The facial expression synthesis process purpose is to endow the robot the ability to express facial expressions. For both synthesizers, the scope is, once again: *{neutral, happy, sad, fear, anger}*.

This section explores our developments over the synthesis part of our framework for human-robot-interaction with emotions. The focus of our framework is on visuo-auditory perception and response. In other words, perception and response can be called analysis and synthesis; the analysis is responsible for the classification of human emotion and the synthesis is responsible for the synthetic expression that the robot must show. This research focus was also explored by us in [74].

For clarity, it must be stated that we consider the fusion to be as much a part of the synthesis as are the effectors. The result of both modalities were combined in this phase and can be used separately or together. There are nine possible combinations for the system:

1. Analyze *audio* then synthesize *audio;*
2. Analyze *audio* then synthesize *face;*
3. Analyze *face* then synthesize *face;*
4. Analyze *face* then synthesize *audio;*
5. Analyze *audio and face* then synthesize *audio;*
6. Analyze *audio and face* then synthesize *face;*
7. Analyze *audio* then synthesize *audio and face;*
8. Analyze *face* then synthesize *audio and face;*
9. Analyze *audio and face* then synthesize *audio and face.*

Taking advantage of this independence across the modalities, in chapter 3 we used option 1 and 3, respectively on the tests presented on figures 3.7 and 3.3. However, in this chapter we focus on option 9 which is the complete fusion.

When the robot is required to express an emotion, it shall express this emotion according to the way of expressiveness learned for that emotion. For this, reverse/inverted Bayesian models are proposed. The proposed models summarized by figure 4.1 allows the robot to express vocal and facial expressions with an automatic features generation of the required expression.

4.2 Facial Expression Synthesis

4.2.1 Reverse Model

Bayesian models are developed with a set of variables that depends to other variables. This dependence can be reverted when synthesizing, in other words, by using the histogram learned on the learning phase for Classification, one can take advantage of the extant information to produce the expressions according to what was learned before. The variables used in our model for synthesis of facial expressions are presented on table 4.1.

The Bayesian network presented in figure 4.2 is analog to the Bayesian network for classification, however the dependencies are in the opposite direction which means

Table 4.1: VARIABLES OF BAYESIAN MODEL FOR FACIAL EXPRESSIONS SYNTHESIS.

Variable	Scope	Description
FE	*{anger, sad, fear, happy, neutral}*	Stands for Facial Expression, it is a *random variable* among the scope
EB	*{none, AU1, AU4, AU1+4}*	Belief variable related with Eye-Brows movements. The four events are directly related with absence or with the existence of AU1, AU4 or of their "non-additive combination" (AU1+4).
Ch	*{none, AU6}*	Belief variable which is related with Cheeks movements; more specifically, the events indicates if AU6 is absent or if the cheeks are raised.
LE	*{none, AU7}*	Belief variable which is related with the Lower Eyelids movements; AU7 is the action unit associated with the raising of the lower eyelids.
LC	*{none, AU12,AU15}*	Belief variable associated with the movements of the Lips Corners. The event none must have a high probability when the corners do not perform any movement. The event au12 must have a great probability when the lip corners are pulled obliquely up and backwards. If the lip corners moves downwards the event au15 must have a great probability.
CB	*{none, AU17}*	Belief variable collecting the probabilities related with the Chin Boss movements. The event none is related with the absence of any movement, while the event au17 has a great probability when the chin boss is pushed upwards.
MF	*{none, AU20,AU23}*	Belief variable associated with the Mouth's Form. The events au20 and au23 indicates, respectively, if the mouth is stretched horizontally or, inversely, if the lips are tightened.
MA	*{none, AU24,AU25}*	Belief variable associated with the Mouth's Aperture. The events au24 and au25 are related, respectively, with the act of the lips are pressed together or lips are relaxed and parted.

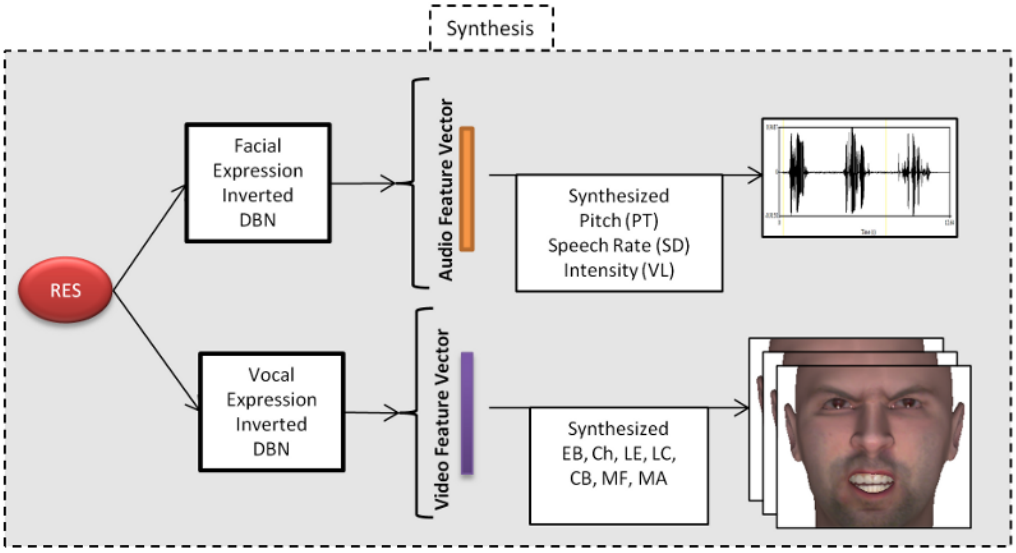


Figure 4.1: The synthesis process. From robot emotional state (*RES*) to synthetic face and voice.

that for a given expression we want to know what are the correspondent variables, at level 2, that shall be synthesized for the desired expression.

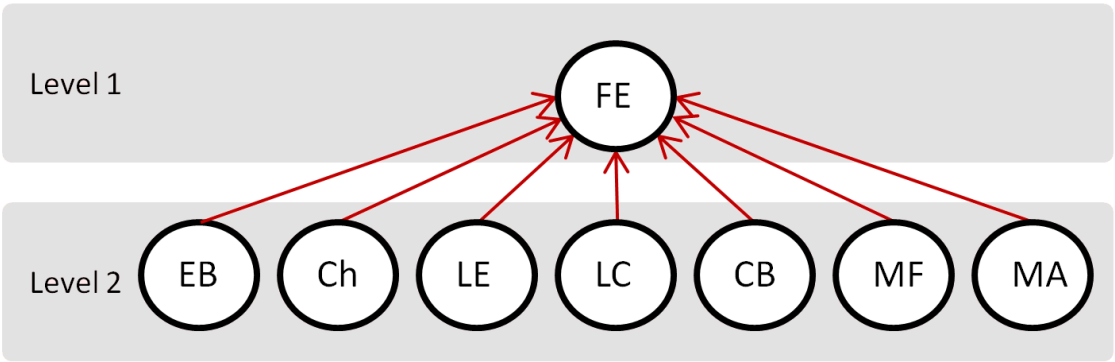


Figure 4.2: Bayesian Model for Synthesis of Facial Expressions.

The following equations illustrates how the model can be reverted:

$$\begin{aligned}
P(EB|FE) &\propto P(EB) * P(FE|EB) \\
P(Ch|FE) &\propto P(Ch) * P(FE|Ch) \\
P(LE|FE) &\propto P(LE) * P(FE|LE) \\
P(LC|FE) &\propto P(LC) * P(FE|LC) \\
P(CB|FE) &\propto P(CB) * P(FE|CB) \\
P(MF|FE) &\propto P(MF) * P(FE|MF) \\
P(MA|FE) &\propto P(MA) * P(FE|MA)
\end{aligned} \tag{4.1}$$

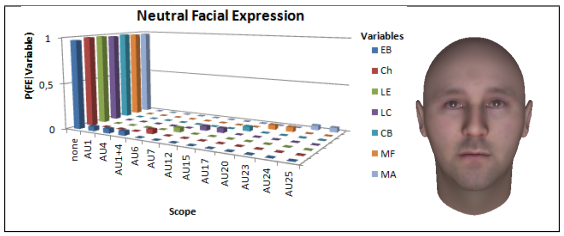
4.2.2 Inference Learning

It is possible to realize that now new likelihoods are needed ($P(FE|EB)$, $(P(FE|Ch)$, $(P(FE|LE)$, $(P(FE|LC)$, $(P(FE|CB)$, $(P(FE|MF)$, $(P(FE|MA)$). However, no new learning is needed to be done here because we already have this data.

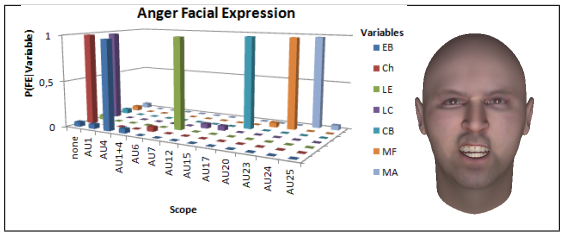
By reorganizing the data from table 3.2 for an specific desired Facial Expression, showing all the possibilities for all the variables, we obtain the a new matrix shown in figure 4.3. Notice that the variables have different scopes, thus we merged all the scopes and the probability can be minimally $0,01$ if the value is in the scope of the variable, however when the value is not on the scope of that variable the probability is 0 (zero). The output of the synthesis model are the most probable variables taking into account the desired Facial Expression.

4.2.3 Results

The produced result is an artificial face image. In our case, the face is similar to a human face; moreover, it need to be able of producing the defined emotions. It is also possible to use different head models as input to the produced face. These head models were previously generated from a set of voluntary persons. Examples of the same face performing the five covered expressions are presented in figure 4.4. Each head model is characterized for its particular face and head shape (see figure 4.5).



(a)



(b)

Figure 4.3: Likelihood for synthesis of: (a) *Neutral* and (b) *Anger*, facial expressions.

Figure 4.4: Facial Expression Synthesis: Example of different expressions *{neutral, happy, sad, fear, anger}*.

Figure 4.5: Facial Expression Synthesis: Example of different head models.

Models of human faces were created, and they were morphed according to *RES* previously defined by the fusion. These models were mesh files which were generated based on three pictures from a person. The Face-Gen 3D head modeler [34] was used to create the meshes. Later these meshes were imported to our OpenCV [33] application and we synthesized expressions. The level of emotion being expressed is denoted by a

variable that can vary from 0% to 100%. See example for the anger expression in figure 4.6.

Figure 4.6: Facial Expression Morphing: from neutral (anger=0%) to angry (anger=100%).

4.3 Vocal Expression Synthesis

Analog to what was done for the facial expression synthesis, the Bayesian dependencies can be reverted when synthesizing the audio also. The auditory features can be produced based on the desired auditory expression to put during a spoken phrase. By taking the histogram learned on the learning phase for classification, we reuse the extant information to produce the auditory expressions. The variables used in our reverted model for synthesis of vocal expressions are the same as described in subsection 3.3.2, table 3.3.

4.3.1 Reverse Model

Figure 4.7 present the graph of the auditory synthesis, which has the inverted dependencies as it is possible to notice by the direction of the arrows. This inverted dependency leads to the equations of the model that endows us to infer what is the auditory features present in a chosen vocal expression.

The following equations illustrates the auditory synthesis model:

$$\begin{aligned} P(PT|VE) &\propto P(PT) * P(VE|PT) \\ P(SD|VE) &\propto P(SD) * P(VE|SD) \\ P(VL|VE) &\propto P(VL) * P(VE|VL) \end{aligned} \tag{4.2}$$

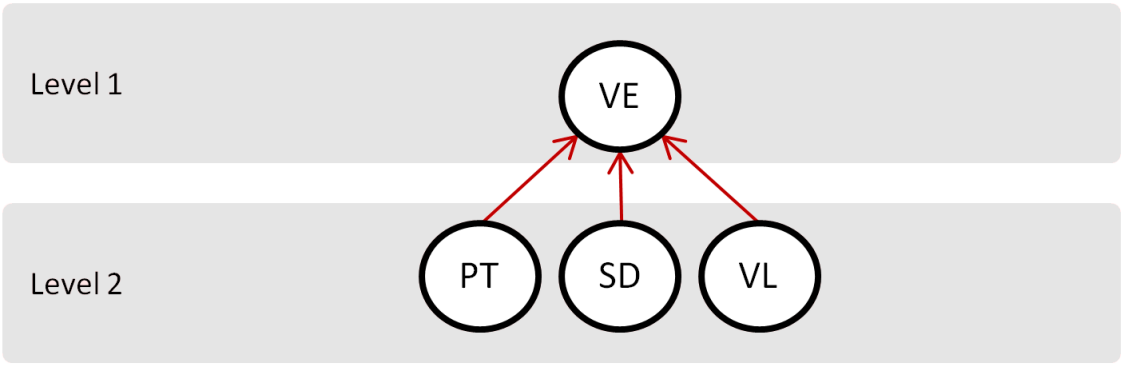


Figure 4.7: Bayesian Model for Synthesis of Facial Expressions.

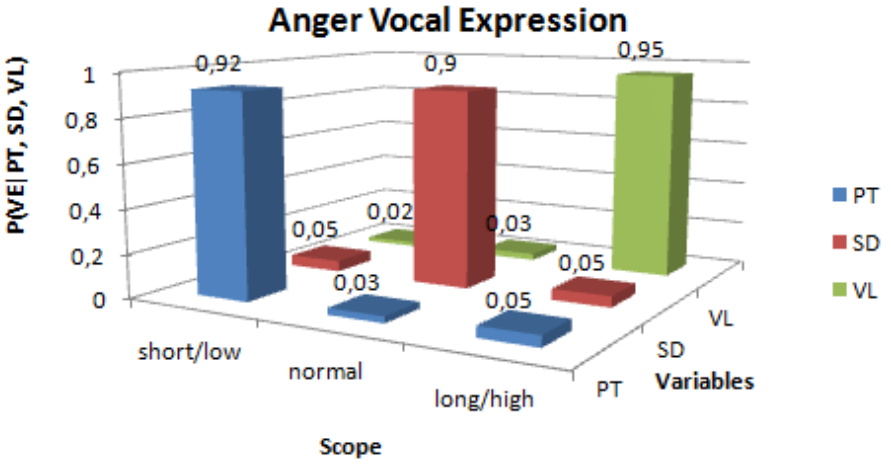


Figure 4.8: Likelihood for synthesis of anger vocal expression.

4.3.2 Inference Learning

The new likelihoods are: $P(VE|PT)$, $(P(VE|SD)$ and $(P(VE|VL)$. By reusing the information presented in table 3.3 for an specific Vocal Expression, in this example *Anger*, we have the features as shown in figure 4.8. The output of the synthesis model are the most probable values for the features to be synthesized, taking into account the selected Vocal Expression.

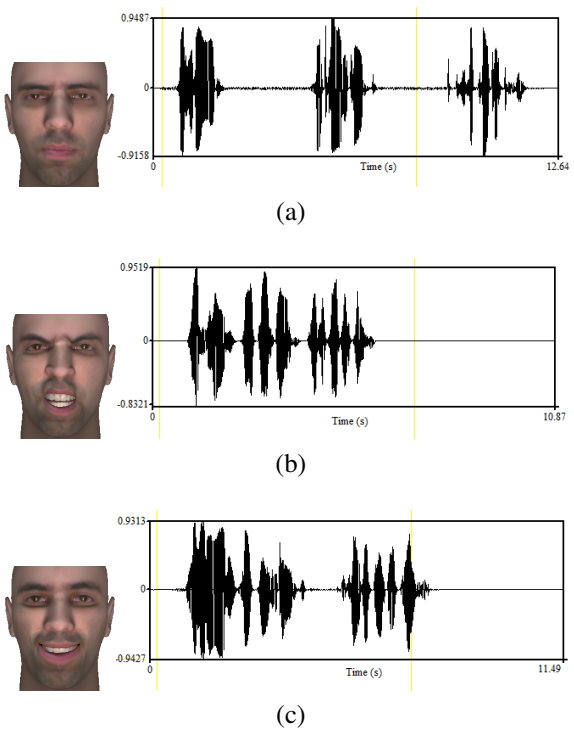


Figure 4.9: Synthesized vocal expressions, respectively: (a) Sad, (b) Anger, (c) Happy. The x axis represents the signal amplitude which is related to VL, the y axis is time.

4.3.3 Results

The same sentence is spoken with different intonations to form the vocal expression. This different intonation is noticeable in the parameters extracted from the wave. In figure 4.9 we present three wave forms of the same sentence synthesized among different vocal expressions. Figure 4.9 (a) is from a sad expression, notice that the speech rate is low, the word are more separated along time. Also, the *sad* amplitude is lower than the *anger* (b), but in this particular case it was higher than the *happy* (c). As expected from the likelihoods represented in figure 4.8, we can see in figure 4.9 (b) that the *anger* pitch was *low*, the *anger* speech rate is *normal* and the *anger* amplitude is *high*.

4.4 Emotional Vector

According to Libin et al. [45], recent research shows that people perceive and treat robots not just as machines, but also as their companions or artificial partners. Person–robot communication, viewed as a complex interactive system, is based upon three basic principles: interactivity, equifinality, and multimodality. Classification of artificial creatures from the robopsychologist's point of view divides them into two major groups: assisting robots, which are oriented toward industrial, military, research, medical, and service activities, and interactive stimulation robots, which are designed for social, educational, rehabilitation, therapeutic, and entertainment purposes. The latter class is considered the primary subject for the robotic psychology and robotherapy. These new fields consist of a concept that places the relationships between humans and robots into a psychological, rather than technological, context. Conceptual and experimental results of implementing the robotic psychology and robotherapy concept into the study of human–robot interactions concern basic operational definitions, theoretical framework, and the design of assessment tools.

After endowing the robot to analyze and synthesize facial and vocal expressions, we propose a Bayesian Mixture Model in order to fuse these both classifiers into one final human emotional state result. Moreover a Social Behavior Profile is given for the robot and this elements compose the emotional vector that defines what type of behavior the robot will have (see figure 4.10). We called this procedure emotional vector and it is used between analysis and synthesis. The structure of the Bayesian network is similar to the previous ones, however in this case the leaves of the graph are the output from the previous networks (FE and VE). Moreover, the social behavior profile comes into this model with the purpose is to endow different personalities to the robot, in it's simplest form, *SBP* would assume the value "*Agreeable*" what would lead the robot to do only human imitation.

4.4.1 Variables

Let's clearly describe all the variables used on our Bayesian equations from the input of classified expressions, through the Bayesian Mixture Model for fusion, to the process of decision of robot's response. These variables are listed and explained in table 4.2.

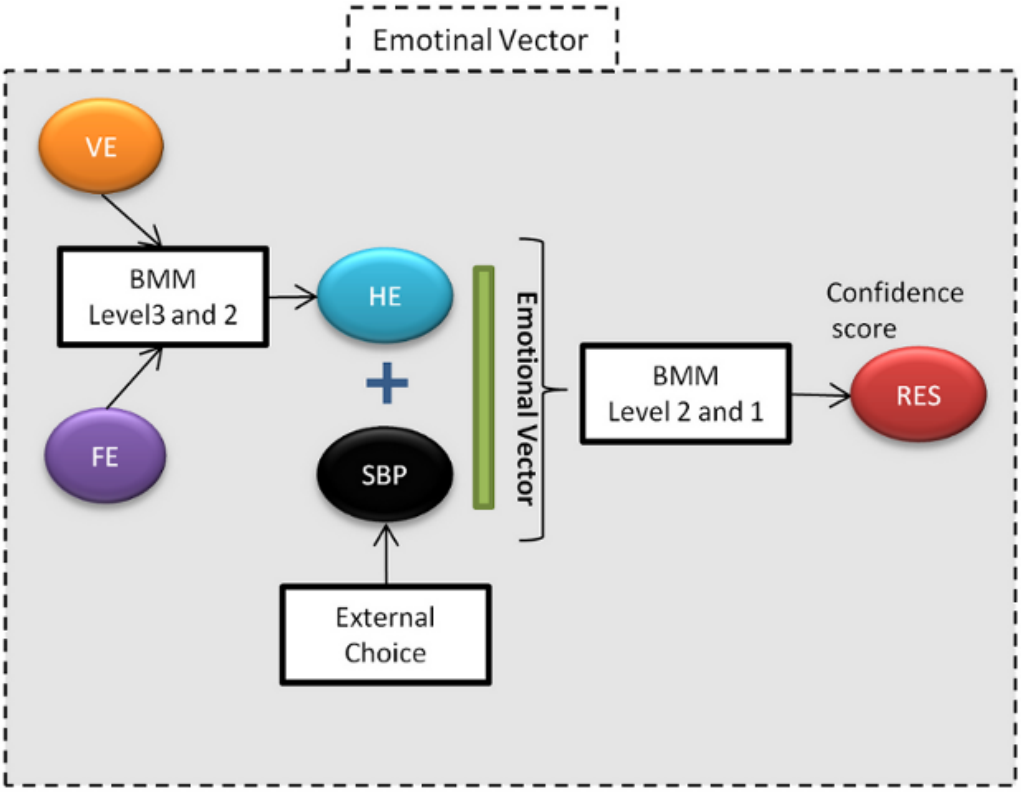


Figure 4.10: Emotional Vector Schema.

4.4.2 Model

The proposed Bayesian network can be seen in figure 4.11. According to the Bayesian theory, it is possible to state that the mixture is computed by the following equation:

$$\begin{aligned} &P(HE, FE_w, VE_w) = \\ &P(FE_w, VE_w|HE) * P(HE) = \\ &P((FE|HE) * fw) * (P(VE|HE) * vw) * P(HE) \end{aligned} \tag{4.3}$$

last equation is valid because we assumed that FE and VE are independent. The *posterior* can be obtained using the Bayes Formula as follow:

$$\begin{aligned} &P(HE|VE_w, FE_w) = \\ &\frac{P(VE|HE) * vw * P(FE|HE) * fw * P(HE)}{P(VE_w, FE_w)} \end{aligned} \tag{4.4}$$

Table 4.2: VARIABLES OF THE BAYESIAN MIXTURE MODEL.

Variable	Description
FE	Stands for Facial Expression, it is a *random variable* among the scope *{anger, sad, fear, happy, neutral};*
VE	Stands for Vocal Expression, it is a *random variable* among the scope *{anger, sad, fear, happy, neutral};*
HE	Stands for Human Emotion, it is a *random variable* among the scope *{anger, sad, fear, happy, neutral};*
SBP	Stands for Social Behavior Profile, it is a *random variable* among the scope *{anger, sad, fear, happy, neutral}*
RES	Stands for Robot Response with a Robotic Emotional State, it is a *random variable* among the scope *{Neurotic, Extroverted, Conscientious, Agreeable and Humorous};*
fw	Stands for of *Facial* expressions *Weight*, it is given by the level of confidence assumed for that classifier. $FE_w = P(FE).fw$
vw	Stands for of *Vocal* expressions *Weight*, it is given by the level of confidence assumed for that classifier. $VE_w = P(VE).vw$

from the marginalization rule we can calculate

$$P(VE_w, FE_w) = \sum_{HE} P(VE|HE) * vw * P(FE|HE) * fw * P(HE) \tag{4.5}$$

During the computation of the BMM, we have the *human emotion* represented by variable HE, it may vary among the scope *{Anger, Fear, Sad, Happy, Neutral}*. Next step is to compute the probability of response (RES) given the human emotion and given the robotic *SBP* that vary among the scope *{Neurotic, Extroverted, Conscientious, Agreeable and Humorous}*. The response (RES) also vary among the scope *{Anger, Fear, Sad, Happy, Neutral}*, and the robot will give the response according to the RES emotional state. The following equations are for the top of the network (level 1 and level 2), it illustrates the joint distribution associated to the Bayesian Fusion with the *SBP*:

$$
\begin{aligned}
&P(RES, HE, SBP) = \\
&P(HE, SBP|RES) * P(RES) = \\
&P(HE|RES) * P(SBP|RES) * P(RES)
\end{aligned}
\tag{4.6}
$$

last equation is only valid when it is assumed that the variables HE, and SBP are independent.

The *posterior* can be obtained from the joint distribution, using the Bayes Formula as follow:

$$
\begin{aligned}
&P(RES|HE, SBP) = \\
&\frac{P(HE|RES) * P(SBP|RES) * P(RES)}{P(HE, SBP)}
\end{aligned}
\tag{4.7}
$$

from the marginalization rule we can calculate

$$
\begin{aligned}
&P(HE, SBP) = \\
&\sum_{RES} P(HE|RES) * P(SBP|RES) * P(RES)
\end{aligned}
\tag{4.8}
$$

The result of the Bayesian inference is a probability vector with all the probabilities for all the possible responses. Usually what is done is the *maximum a posteriori decision (*more probable value is selected as result). In this particular case, after selecting what is the emotional state the robot shall express, a random number generator helps us to add some randomness in the process. At the end, the true response will be given by a random choice among the 5 possible sentences for each RES state of that stage of the conversation. We decided to add this randomness in order to decrease the predictability of the robot's responses by the part of the user.

4.4.3 Inference Learning

After endowing the robot to analyze and synthesize facial and vocal expressions, the proposed Bayesian Mixture Model (that we called "Emotional Vector") serves for to fusing both classifiers into one final human emotional state result. This BMM also need to be

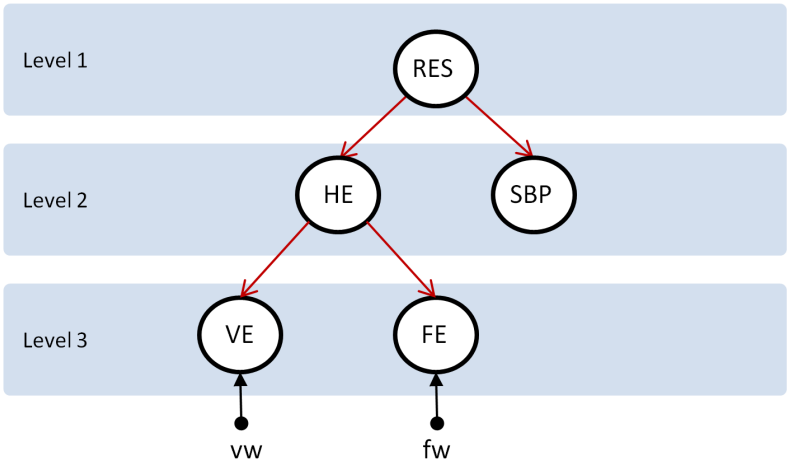


Figure 4.11: Bayesian Network for fusion of the two used modalities and also fusion with the given *social behavior profile.*

filled out with the respective likelihoods, several questions will be solved by this BMM learning as for example: How a neurotic robot shall respond to an anger question? How an extroverted robot shall respond to a happy question or statement? These possibilities and others are covered with this learning, and the possibility of having of a robot able to do this is endowed by our probabilistic system.

In this case, the Bayesian network priors and likelihoods are given from level 2 to level 1, and from level 3 to level 2. *RES* stands for response also for Robotic Emotional State, since each response caries a robotic emotional state with it. In order to build the likelihood histogram distribution tables, several studies from psychology were consulted. More specifically, Deary et al. [19] defined the five-factor model of personality traits. It was mentioned by Banos et al. in [2]: "Positive psychology researchers posit that the capacity to experience positive emotions may be a fundamental human strength central to the study of human flourishing. Thus, a lot of efforts are devoted to the understanding and study of the factors that allow individuals to flourish". Based on this motivation, from the five-factor model of personality traits, we replaced the "open to experience" personality trait by the "humorous". The description of each personality trait is shown in table 4.4.

The initial prior was once again defined as an uniform distribution. The priors are $P(RES = anger) = ... = P(RES = happy) = 0,20$. In this network the priors are also

Table 4.3: *RES* LEARNED HISTOGRAM OF LIKELIHOODS.

		RES				
		Anger	Fear	Happy	Sad	Neutral
HE	Anger	0.50	0.12	0.12	0.12	0.12
	Fear	0.12	0.50	0.12	0.12	0.12
	Happy	0.12	0.12	0.50	0.12	0.14
	Sad	0.12	0.12	0.12	0.50	0.12
	Neutral	0.14	0.14	0.14	0.14	0.50
SBP	Neurotic	0.77	0.20	0.01	0.20	0.01
	Agreeable	0.20	0.20	0.20	0.20	0.20
	Conscientious	0.01	0.20	0.01	0.40	0.77
	Extroverted	0.01	0.20	0.39	0.10	0.01
	Humorous	0.01	0.20	0.39	0.10	0.01

Table 4.4: ADAPTED FROM [19], DESCRIPTION OF OUR ADAPTATION FROM THE FIVE-FACTOR MODEL PERSONALITY TRAITS.

Personality Trait	Characteristics
Neurotic	Sensitive, emotional and prone to experience feelings that are upsetting
Extroverted	Outgoing and high-spirited. Prefers to be around people most of the time.
Conscientious	Well-organized. Have high standards and always strive to achieve the goals.
Agreeable	Compassionate, good-natured, and eager to cooperate and avoid conflict.
Humorous	Has the capacity to makes someone laugh or be amused.

(a) FE=90% happy, VE=80% happy, SBP=Agreeable, fw=100%, vw=100%
(b) FE=80% sad, VE=90% sad, SBP=Agreeable, fw=80%,vw=90%
(c) FE=90% anger, VE=90% sad, SBP=Agreeable, fw=100%, vw=20%
(d) FE=90%happy, VE=90% happy, SBP=Neurotic, fw=100%, vw=100%

Figure 4.12: Image inputs for the BMM tests: the classified FE, VE, the weights fw, vw and the selected SBP.

re-alimented dynamically, where the posterior from time $t-1$ becomes the prior on time t. As opposite to what happens with the prior, the likelihood remained the same over time. In table 4.3 the histogram of likelihood obtained as a result of learning is presented. It is possible to notice that, for example, a happy HE "calls" for a happy response without losing dependence from the SBP.

4.4.4 Results for Emotional Vector

We have performed several tests with different input audio and vocal expressions and different social behavior profiles. To illustrate some of these experiments, we present in figures 4.12,4.13 and 4.14 four of our tests. For clarity, we present one of the frames belonging to the image of the performed facial expression, the vocal expression and the selected social behavior profile is only mentioned in the text.

The inputs are shown in figure 4.12; the result of the mixture model is given after 3 utterances as shown in figure 4.13; the respective synthesized facial expressions are shown in figure 4.14. The same expression is also synthesized over the audio output.

Notice that we are only showing the *maximum a posteriori decision* of the input, however it takes into consideration the complete probability distribution function (PDF) of VE and of FE. Moreover we presented three results of the *Agreeable SBP* because it is the case where it is more easy to understand (*Agreeable* lead to human imitation), in a graphic, whether the results from the HE are correct or not.

The vw and fw stands for vocal expression weigh and facial expression weight,

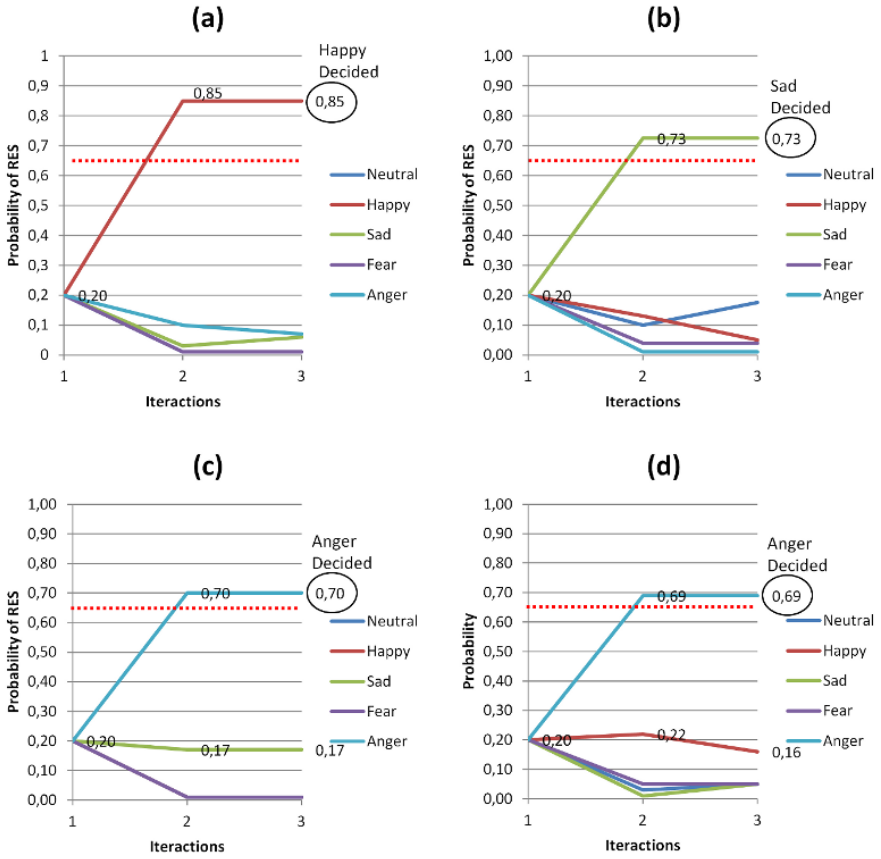


Figure 4.13: Results of the Bayesian mixture model; probability of*RES*,in the case of *SBP* set to *Agreeable* ("a","b" and "c") goes up to a mixture value of both classifiers. In "d" the resultant probability was highly influenced by the *SBP* of *Neurotic*. In the case of this fusion, any value over 65% is considered to be a final decision of *RES* to express.

they are used as confidence level of the previous classifier. How sure one can be of that the classification coming from the facial expression classifier is correct? How percent can I trust in the vocal expression classifier? In the examples shown in figure 4.13 they were set randomly, however, for a better accuracy in the real interaction benchmarks were run over the classifiers to determine the confidence level of the formers.

In example (c) we show a divergence between auditory and visual classifier results, and how this difference can be compensated by the belief percentage of each classifier. In example (d) the robot *SBP* is set to *Neurotic*, so although the *HE* is definitely *happy*, the probabilities for happy in *Neurotic* from table 4.3 are very low. Thus the re-

(a) (b) (c) (d)

Figure 4.14: Output faces from the Emotional Vector tests.

sponse in this case outcomes as an angry response.

4.5 Conclusions

This chapter presented models for robot emotion synthesis. The facial expression reverse model was presented in section 4.2.1, while its respective learning and results were presented in sections 4.2.2 and 4.2.3. The vocal expression reverse model was presented in section 4.3.1, and its respective learning and results were presented in sections 4.3.2 and 4.3.3. This models were published by us in [74]. The presented methods bring the advantages of not only synthesize an emotion over the robot, but also this synthesis can be trainable/customizable according to the features correspondent to each expression. Moreover, in section 4.4, an the concept of the emotional vector is described. Inference learning for the emotional vector was explained in section 4.4.3, and results were presented in section 4.4.4. The emotional vector is used to endow personality to the robot. The combination of the resultant human emotion from the classification methods, with the selected social behavior profile leads to influence over the decision for the robot's response. An evaluation of the emotional vector impact over human robot interaction was published by us in [75], and also continuation of the work was submitted to publication as refered in Chapter 1, section 1.8, point 7 and 8. Next chapter will present a study case of the whole system, by using the feature extraction, the classification, the synthesis and the emotional vector to test different personalities on the robot.

Chapter 5

Dialog between Human and Robot

5.1 Overview

Several sub-problems need to be addressed in order to create a dialog (see figure 5.1) between human and robot. This area can be so large, that researches need to focus in one or another sub-area. Several different approaches exist in the literature, some of them address only semantics on chat bots [13][8]. The semantics works usually exclude the visual and auditory issues, exclude the emotional issues and concentrates only in text chat and vocabulary semantics. In the presented case, we did the opposite, we excluded the semantics concern, so we do not do semantic analysis at all. Our work is focused on endowing emotion analysis and behavior modification with a robotic social behavior profile. Notice that, for having an effective social behavior profile, some semantic knowledge

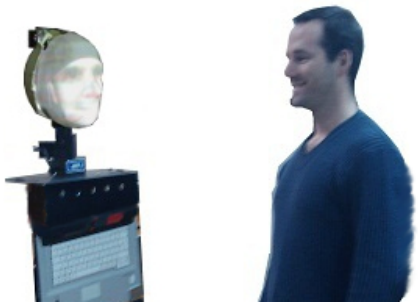


Figure 5.1: Robot to human dialog.

would be necessary. Our approach to deal with this issue was by running our experiments runs over a story board. With this approach, the semantic is known, so it does not need to be discovered automatically. The task is specific, and all the possible responses are given to the robot.

We believe that our contribution can be set together with any extant semantics text analysis program. Therefore, although we excluded the semantic part, the emotional system presented here could be, for example, attached to a system like [8]. Thus, endowing emotion recognition and personality behavior modifications to a chat-bot avatar. Our contribution can be embedded to a social robot with predefined tasks. In this case, regular tasks, performed by the robot, became more user friendly by the addition of emotive feedback. We strongly believe that our contribution is not meant to be alone in the future of social robots. All the work can be considered to be an emotive module or a component for a social robot. A complete social robot needs navigation solutions, needs recharging solutions, need control solutions, and perhaps need semantics text analysis solutions. In the prototype developed by us, we used basic extant algorithms to solve the navigation issues, the obstacle avoidance issues and the control issues. For contouring, or avoiding, the need of a semantic text analysis, we restrict our dialog to a given story board. In fact, every robot, even those who deal with semantics, has a database of sentences and has a database of limited words in its vocabulary. Even we humans have a limited database of sentences and words, in our brains. All these issues are known to be problems of a Social Robot. Although we are aware of that, the issues about semantic analysis are not addressed in this thesis.

5.2 Social Robot Dialog Issues

Once it is clear that we are not dealing with semantics, there are several dialog issues that we needed to cover and solve. A dialog is a conversation between two entities that talk with each other, one entity can speak one or more sentences, and the interlocutor can speak one or more sentences in response. So the first issue is how the robot knows that it is its turn to speak. Although the dialog is multi-modal and we have both auditory and visual channels, it is reasonable to realize that what commands the synchronization of the

dialog is the auditory channel.

5.2.1 Robot discovers its turn to talk

There are several assertions that need to be made in order to set the system to operate over a dialog.

- Assertion 1: the vocal channel commands the change of turns in the dialog.
- Assertion 2: the human can only speak one sentence in his/her turn.
- Assertion 3: the turn will change when silence is detected.

The robot than can speak one or more sentence, depending to what is in the story board (defined for each task). Artificial chat bots (text only) always responds something when the user finish a sentence. In the case of text only chat bots, the user sentence clearly finish when the human hits the "Enter key". However, when this concept is applied to a vocal chat robot, there is no such a deterministic certain way to define that the human ended his/her sentence.

According to Hoelper et. al. in [32], a sound can be characterized as being silence, unvoiced or voice. The classification of a wave sound can be performed by measuring the amount of energy in the signal. However, there is always some noise present in the background and the influence of the background noise may vary according to the microphone used. The amount of energy considered to be silence was adjusted empirically to our current microphone. Furthermore, we have implemented a silence detector based on [32] to determine the change of the turn. With this silence detector solution we covered the problem of changing the turns, because the human can speak for a long time or a short time, and the end of the sentence is detected by the silence detector. The robot will just answer after considering that the human finished it's turn, as shown in figure 5.2.

With the help of the silence detector, some special cases can be perceived, for example, when the human abandon the robot and there is no response at all, the robot will "understand" that the human is not responding. This perception influences the recognized auditory emotion. Based on the psychological concepts presented by Deary et. al. in [19],

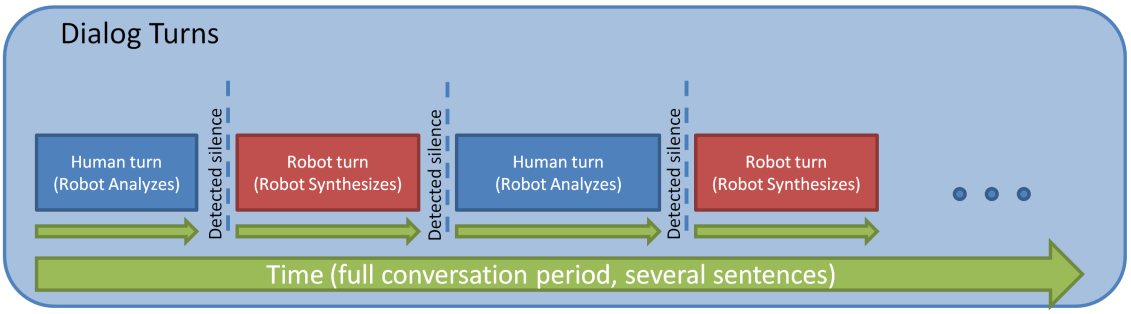


Figure 5.2: Dialog turns.

in our experiments we trained the system to associate silence and abandonment with the *sad* emotion.

Later the auditory detected emotion need to be synchronized with the visual detected emotion. This is done by the proposed visuoauditory fusion inference while analyzing. Notice the fact that inference steps in auditory analysis are not homogeneously separated in time. However, since the robot does not need to infer the emotion during the sentence, the proposed solution was to trigger the inference after each detected "end of sentence".

5.2.2 Synchronizing auditory and visual channel during analysis

Heretofore, we understand that there is no automatic semantic analysis, the story board is given and restricted, the auditory channel commands the dialog turns. The turn changes with the use of a silence detector. However, how do we synchronize this with the facial expressions recognition? The main issue here is that when the human is talking, it is difficult, or even impossible, to maintain a recognizable facial expression. Furthermore, the movements of talking lips interfere on the classification of the facial expressions.

The vocal expression is classified sentence by sentence, so, to have a synchronous classification of human emotion, the facial expression also need to be classified sentence by sentence during the dialog. However, the proposed facial expression classifier is capable to be classifying all the time, converging from one expression to another expression, and the user could have different expressions along the time he/she speaks the sentence.

The facial expression classifier can perform multiple classifications per sentence,

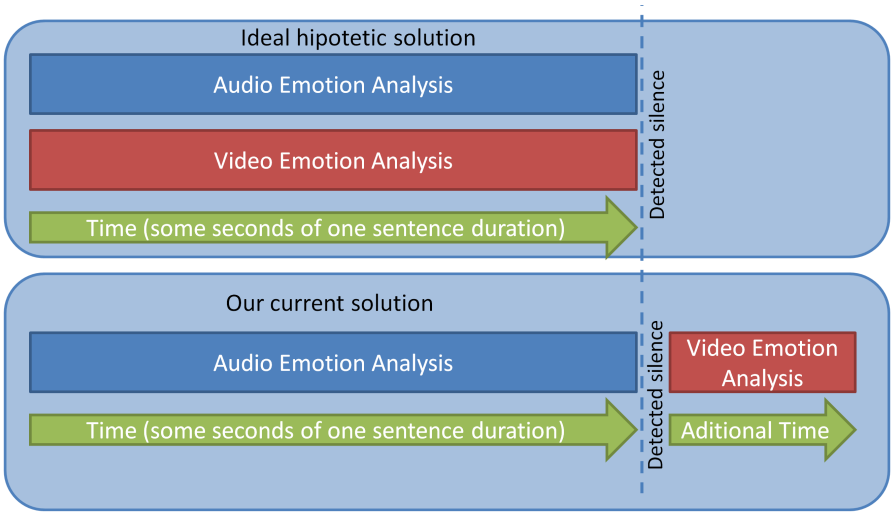


Figure 5.3: Audio and Video Analysis timing during a dialog.

but the vocal expression classifier can perform just one classification per sentence. We need to collect just one output per classifier, at a time, in order to merge them with the proposed mixture model. So the selected time is the slowest one, which is the auditory perception time. For the auditory perception classifier, each time instance is one sentence. Thus, another assertion needed to be done:

- Assertion 4: only one facial expression is taken into account, per each audio sentence.

The elected facial expression is the one which was classified more frequently during the time user spoke the sentence. Some problems were identified in this process, because for some facial expressions are difficult to be maintained during speech, for example, the fear expression. Humans cannot perfectly hold a facial expression while talking, to work around this problem, we decided to give the user some extra seconds to perform the facial expression associated to the previous spoken sentence.

- Assertion 5: some seconds are given to the user to perform the facial expression associated to the sentence, after the end of the spoken sentence, as shown in figure 5.3.

5.2.3 Synchronizing auditory and visual channel during synthesis

The input variable for Synthesis of Vocalization is the same as for the Synthesis of Facial Expressions (*RES*). However, here the vocalization synthesizer takes this input and continues the story board producing the desired output sound. The vocal expressions phrase database was a previously prepared database of a finite set of possible phrases that can be spoken.

Some visemes were used and are associated with phonemes for the English language according to:

1. BMP: (sounds like the “m” in mother, also covers “b”, or “p”),
2. EEh: (big "E" sounds like as in "Free"),
3. Er: (Sounds like as in "Earth"; may also be used for phonemes like "H" in "Hammer"),
4. Ay: (“I” sounds like as in "Cry" and "Pray"),
5. i: (fast "I" sounds as in "it", "Tip"),
6. Oh: (sounds as in "Flow"),
7. ooo: double “u”, “w” (sounds as in "Mood" and "Wild"),
8. YchJ: (as in "You, Cheese, Jacob"),
9. FV: (covers “f”, or “v”).

Moreover, lips synchronization was implemented over the avatar, nine visemes (each set in figure 5.4), associated to nine phonemes (see table 5.1), which were used according to what the avatar would speak. Since we were not doing phoneme recognition, this lips synchronization was only possible on the avatar responses where the phrases were known and not on the avatar which was mimetizing the human. Notice that, with these nine visemes, we are able to synchronize the lips at one single facial expression. So we expanded this model of visemes to cover lips synchronization with different facial expressions.

After the robot had analyze the FE and VE associated to one sentence, HE will be calculated as explained in section 4.4. Then *SBP* will be inserted as also explained in section 4.4. The synchronization problem during synthesis is treated in a different manner, because the robot, unlike the human, is completely able to perform the facial expressions

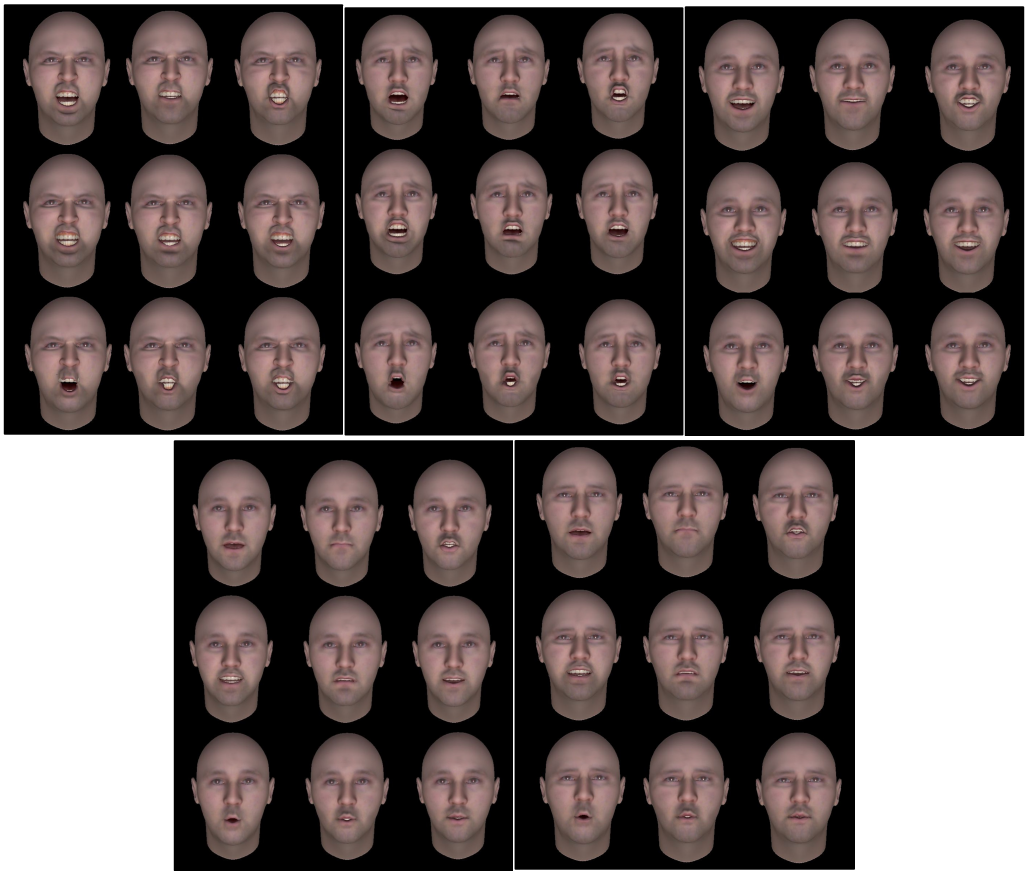


Figure 5.4: The five sets of visemes that allow the robot to speak while performing a facial expression.

while speaking. For this, 5 sets of 9 visemes each were prepared. The visemes shown in figure 5.4 are organized in each set according to table 5.1.

Still in figure 5.4, each set of visemes are organized from left to right as: Anger, Fear, Happy, Neutral and Sad. With these different meshes, robot has the ability to speak, synchronizing the lips and performing the facial expression at the same time. This is very difficult, sometimes impossible, to be performed by a human, but it is very easy to be perceived by a human.

- Assertion 6: unlike human, robot can speak and perform the facial expression at the same time, as shown in figure 5.5.

With all this assertions, the dialog procedure is covered. The human can speak

Table 5.1: PHONEMES CORRESPONDENT TO THE VISEMES OF FIGURE 5.4.

Ay	BMP	YchJ
eee	FV	i
Oh	ooo	Er

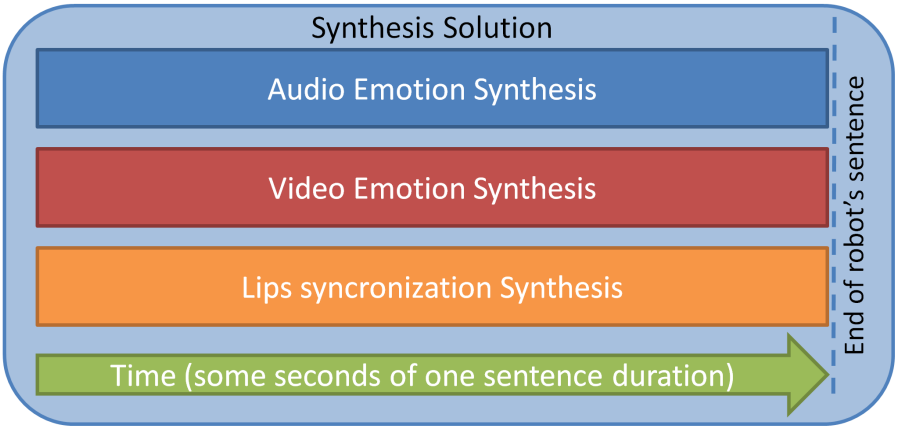
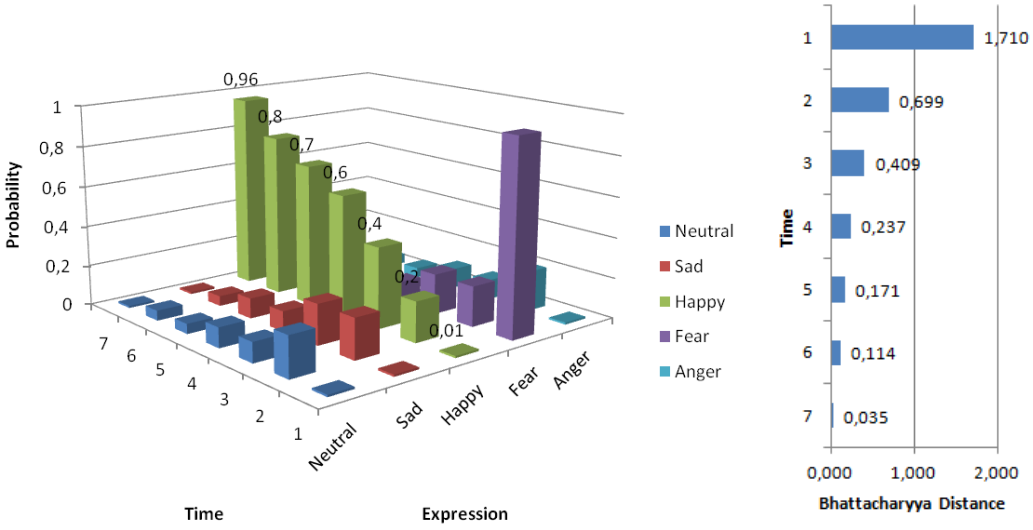


Figure 5.5: Audio and Video Analysis timing during a dialog.

with the robot, performs a facial expression that will be associated with one spoken sentence. At each sentence, robot is able to classify facial and vocal expression. It is also capable to identify when is its turn to speak. Furthermore, robot can speak while performing vocalization and facial expression at the same time. The robot does not deal with semantic, but is capable of responding the same idea with different emotions, what means different vocal tones, different sentences (more polite or less polite, more funny or less funny) from a database of sentences. Alike the humans, the same information can be passed in different ways, with different emotions and different words, but it is still the same "raw" information. It is fascinating to notice that when the same information is passed in different ways, different human reactions are expected. This modification in human behavior is measured in section 5.8 according to the assessments defined in section 5.4.



(a) (b)

Figure 5.6: Bhattacharyya Distance reduces as $P(A)$ get closer to $P(R)$.

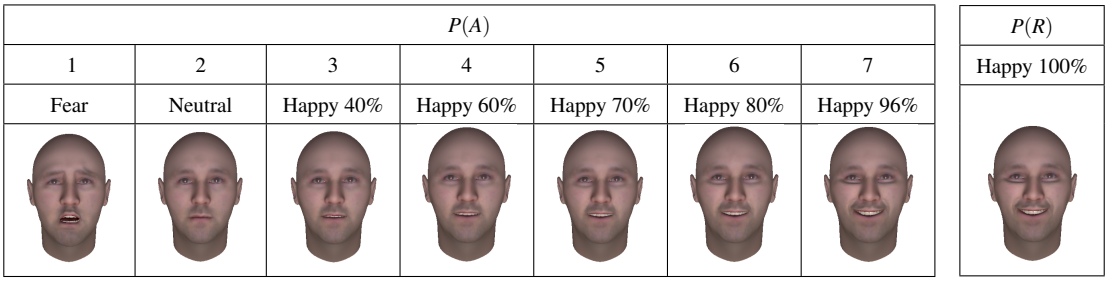
5.3 Expression Verification

5.3.1 Assessments for Robotic Expression Verification

It is expected that the robot presents an emotion according to the given *SBP*, thus it is needed a measure to know if this emotion was expressed correctly. Thus, we state error (E) as being the distance from the "output emotion" from the robot and the "expected emotion". This is valid for both audio and face synthesis. Then we used Bhattacharyya distance according with Appendix A.

5.3.2 Results of Robotic Expression Verification

Bhattacharyya Distance reduces as the histogram from the presented output of robot's expression $P(A)$ get closer to the expected expression $P(R)$ (see figure 5.7). Notice, in figure 5.6, that at time 1 a wrong expression was presented, thus the error increases to $1,71$. Any value of D_B higher than $0,699$ indicates that a wrong expression is being shown. As the expression converges to the expected one, D_B decreases up to near zero.



(a) (b)

Figure 5.7: (a) Input images for verification, each of the images originates $P(A)$ in a different time. (b) The ground-truth $P(R)$. The respective histograms are shown in figure 5.6(a) and the respective results of the verification are presented in figure 5.6(b).

5.4 Engagement assessments

In order to evaluate our proposed methods, we selected an study case to run the tests over. This study case lies in a physiotherapy robot, where the purpose is to give moral and contactless support to a subject while this persons performs his/hers exercises. Several studies regarding physiotherapy robots are frequently about exoskeleton robot to physically support the exercises (as for example in [96]). The idea of setting up a social robot to do moral support is motivated by the fact that many patients needs encouragement. This encouragement is usually done by the physiotherapist, however with the help of such a robot, physiotherapist could have more free time to engage with multiple patients. The robot by it self will then keep attention on one of the patients, controlling and motivating this patient of continuing the exercise list provided by the physiotherapist until the professional returns. Since our purpose is to know how the robotic expressiveness is influencing the interactive experience, it is necessary to define assessments to quantize how the human is engaged in this experience.

There is no standard benchmark for this type of system. However, there are extant ideas for assessments that we reinterpret, by defining our own assessments.

The desired measurements must be related to the engagement of the human during the conversation. Touch-less interfaces are one of our constraints, though we selected the variables listed bellow:

1. Time between phrases (TBP): This variable is measured in seconds and it is anno-

tated along the time.

2. Total Dialog Time (TT): This variable is measured in seconds and it is annotated after the entire dialog.

3. The amount of Happiness (AH): This variable is an integer and represents the amount of times that human expressed happiness during the dialog.

The TBP, TT and AH are annotated during the experiments by an external agent that observes the dialog. This external agent is called engagemeter and currently the annotations for TT and TBP are done manually based on a recorded video of the dialog. The AH and the Error (see 5.3.1 for Error) are automatically calculated. We expect in near feature to have a fully automatic engagemeter.

5.5 Experimental Scenario Details

Since the response's sentences are determined according to the task, in this study case the robot uses the dialog defined on figure 5.8. At each cycle, the exercise and the amount of times change according to what the physiotherapist had previously defined (see figure 5.8). The emotion, facial expressions and voice intonations vary according to the patient emotive action/reaction. Later results about the patient feedback will be presented. Different persons may need different motivations, so the robot *Social Behavior Profile* is configurable to be adjusted according to the profile or the mood of the patient.

At each utterance the robot classifies the expression of the human, than it behaves according to the selected *SBP* to suit the subject mood or profile. Later the effects of this behavior will be measured according to our proposed assessments presented in section 5.4.

5.6 Relation between SBPs and emotions

The *SBP* of the robot shall be selected according to the personality of the user, in this study case, the patient. Each *SBP* is related with a way of expressiveness, therefore, there is a

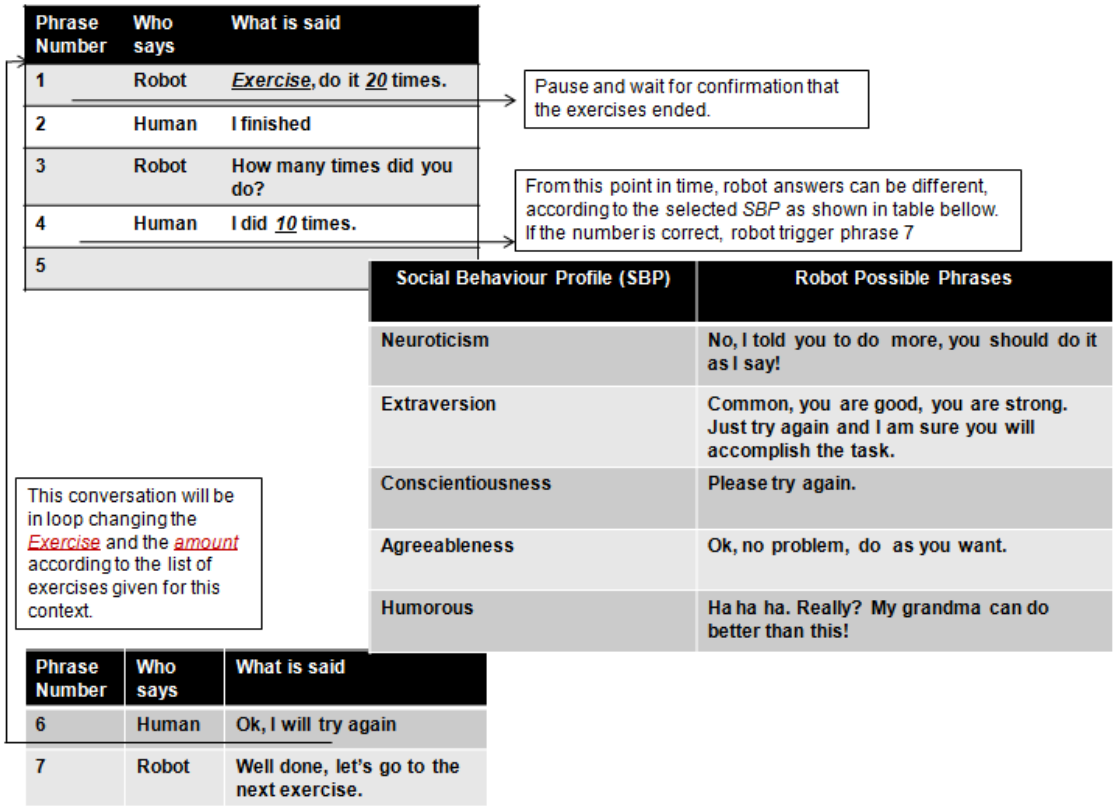


Figure 5.8: Stimulating exercises — In this study case, the robot is an assistive robot that gives instructions during physiotherapy exercises, each utterance repeats until the patient confirms that he did the proper amount of exercises.

strong relationship between the *SBP* of the robot and which emotion it will express. This relationship is not a direct association, however the robot uses the emotive expressiveness to behave in a certain way that fits the current *SBP*. In the list below we present the *SBP* scope and it's relationship with the expression.

- The "Neurotic" *SBP* is characterized by being prone to experience feelings that are upsetting; thus the robot express *anger* or *fear*, context dependent.
- The "*Extroverted*" *SBP* is outgoing, high-spirited, prefer to be around people most of the time; therefore it is associated with the *happy* expression mostly.
- The "*Conscientious*" is well organized, have high standards and always strive to

achieve the goals, it is cold, strait forward and thus it is related with the *neutral* expression. This social behavior profile is where robots without emotions fits.

- The *"Agreeable" SBP* is characterized by good-natured and eager to cooperate and avoid conflict; therefore, all the 10 variables (video plus sound) will be imitated from what was interpreted from the human. The *agreeable SBP* is basically *human-imitation* over the two channels, together with the agreeable phrase defined on the task dependent context.

- The "*Humorous*" *SBP* is very imaginative and willing to consider new ways of doing things, it is funny and may perform jokes to cheer up the interlocutor. Therefore it is related to *happy*.

In our case, each emotion that the robot expresses includes the multi-modal channels that we are using, and additionally a pre-defined context of conversation as presented on subsection 5.5.

5.7 Platform Setup

Our robotic platform was built upon SoPHIE (Social rObotic Platform for Human Interactive Experimentation, see also description of this platform in our work [70]), and the base consists in a two wheel differential Scout robotic platform. A laptop support was added over it and this laptop is the computer responsible for all the processing, from the emotional processing to the movements control. Standard control algorithms were used to move the platform around. Moreover the neck of the robot is a pan-tilt unit with two degrees of freedom and over this unit is attached our retro-projectable expressive head. The retro-projectable mask was built for a better interactive interface and its schema is presented on figure 5.9. The full mounted robot, with clothes, is presented in figure 5.10a.

The robotic technology used as the experimental platform had an active vision system. This feature allowed the robot to move its head towards the tracked face before starting to move its body, thus, the robot avoided unnecessary movements with the body structure.

Figure 5.9: Retro-projected translucent mask.

Furthermore, as another platform; we developed a 3D virtual world as a "Blender game", where the corresponding basis of interactions could be used both over the real robot and/or inside the virtual world; see figure 5.10b. We generated 14 meshes of heads from 14 persons of our lab, so that we used the image of the real person on the avatar that mimics the human. Stereo vision systems were also a possibility that we tested but did not seek. According to the techniques presented on chapter 2, the background segmentation allowed by the stereo vision can be used to improve the choice of which user to interact with.

5.8 System Overall Results

We present in table 5.2, an example of a dialog flow from our tests, with the *SBP* set to *Neurotic*. Notice that even when the human performs happy, the response is mostly aggressive. In tables 5.3, 5.4, 5.5 and 5.6, examples of the same dialog with the SBP set to *Agreeable*, *Humorous*, *Conscientious* and *Extroverted*; respectively.

Analyzing the mentioned tables, it is possible to notice that, as expected, in this short dialogue, the robot behave in different ways according to the defined personality. For *Neurotic* it responds more often with anger while for *Humorous* and *Extroverted* the results were happy all the time. For *Conscientious* the probability tends more to the neutral expression and finally the *Agreeable* is pure human emotion imitation.

Table 5.2: EXAMPLE OF PERFORMED EMOTIONS DURING DIALOG WHEN *SBP* IS SET TO *NEUROTIC*.

Phrase Number	Who Says	What is said	Facial Expression	Vocal Expression
1	Robot	Exercise one, do it 20 times	neutral	neutral
2	Human	I finished	sad	sad
3	Robot	How many times did you do?	anger	neutral
4	Human	I did 10 times	neutral	neutral
5	Robot	No, I told you to do more, you must do as I say!	anger	anger
6	Human	Ok, I will try again	happy	happy
7	Robot	So, let's continue.	anger	anger

Table 5.3: EXAMPLE OF PERFORMED EMOTIONS DURING DIALOG WHEN *SBP* IS SET TO *AGREEABLE*.

Phrase Number	Who Says	What is said	Facial Expression	Vocal Expression
1	Robot	Exercise one, do it 20 times	neutral	neutral
2	Human	I finished	sad	sad
3	Robot	How many times did you do?	sad	sad
4	Human	I did 10 times	neutral	neutral
5	Robot	No, I told you to do more, you must do as I say!	neutral	neutral
6	Human	Ok, I will try again	happy	happy
7	Robot	So, let's continue.	happy	happy

Table 5.4: EXAMPLE OF PERFORMED EMOTIONS DURING DIALOG WHEN *SBP* IS SET TO *HUMOROUS*.

Phrase Number	Who Says	What is said	Facial Expression	Vocal Expression
1	Robot	Exercise one, do it 20 times	neutral	neutral
2	Human	I finished	sad	sad
3	Robot	How many times did you do?	happy	happy
4	Human	I did 10 times	neutral	neutral
5	Robot	No, I told you to do more, you must do as I say!	happy	happy
6	Human	Ok, I will try again	happy	happy
7	Robot	So, let's continue.	happy	happy

Table 5.5: EXAMPLE OF PERFORMED EMOTIONS DURING DIALOG WHEN *SBP* IS SET TO *CONSCIENTIOUS*.

Phrase Number	Who Says	What is said	Facial Expression	Vocal Expression
1	Robot	Exercise one, do it 20 times	neutral	neutral
2	Human	I finished	neutral	neutral
3	Robot	How many times did you do?	neutral	neutral
4	Human	I did 10 times	neutral	neutral
5	Robot	No, I told you to do more, you must do as I say!	neutral	neutral
6	Human	Ok, I will try again	happy	happy
7	Robot	So, let's continue.	neutral	happy

(a) (b)

Figure 5.10: a)One version of our robot: Scout based platform and a head with 2 degrees of freedom. b)Our virtual world is another option of interaction instead of the robot. A real person can look to the camera, and speak at the microphone; where an avatar mimics this person and the other avatar simulates the robot.

A series of tests was performed over our study case presented in subsection 5.5, the tests were run among a number of male subjects. The time between phrases (TBP) and the total time (TT) were measured. Moreover, let's state clearly that the TBP may have different meaning according to the phrase.

Analyzing figure 5.8, where we generalized the conversation, notice that:

1. The time between phrase 1 and 2 ($TBP(1,2)$) is not applicable for interaction purposes, because it is the time the user takes to complete the exercise,

2. The $TBP(2,3)$ is dependent of our silence detector[1], since phrase 3 is an response of the robot,

3. $TBP(3,4)$ is relevant, it depends only of the human reaction,

[1]The silence detector is a technical feature that detects the end of human speech and will not be discussed in this paper.

Table 5.6: EXAMPLE OF PERFORMED EMOTIONS DURING DIALOG WHEN *SBP* IS SET TO *EXTROVERTED*.

Phrase Number	Who Says	What is said	Facial Expression	Vocal Expression
1	Robot	Exercise one, do it 20 times	neutral	neutral
2	Human	I finished	sad	sad
3	Robot	How many times did you do?	happy	happy
4	Human	I did 10 times	neutral	neutral
5	Robot	No, I told you to do more, you must do as I say!	happy	happy
6	Human	Ok, I will try again	happy	happy
7	Robot	So, let's continue.	happy	happy

4. $TBP(4,5)$ is subsidiary of our silence detector,

5. $TBP(5,6)$ is relevant, it depends only of the human reaction,

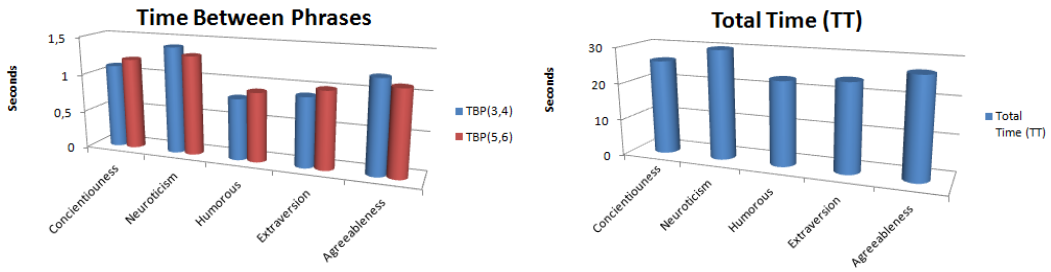
6. $TBP(4,7)$ when the person did the exercise correctly; its premise is the same as $TBP(2,3)$ and $TBP(4,5)$.

Considering this, we present figure 5.11 with the average of $TBP(3,4)$, $TBP(5,6)$, TT. Figure 5.12 shows the result of *AH* for the different Social Behavior Profile of the robot.

A study on tolerable waiting time (for Internet interactive systems) was done by Nah et al. in [60], and according to what was defined by Nah we can state that:

- 1 second is about the limit for the user's flow of thought to stay uninterrupted, even though the user will notice the delay. Normally, no special feedback is necessary during delays of more than 0.1 but less than 1.0 second, but the user does lose the feeling of operating directly on the data.

- 10 seconds is about the limit for keeping the user's attention focused on the dialogue. For longer delays, users will want to perform other tasks while waiting



(a) (b)

Figure 5.11: In (a) is presented the average of $TBP_{3,4}$ (left columns) and $TBP_{5,6}$ (right columns); notice that the response time in Humorous and Extroverted are faster than the others. (b) presents the total time of the entire conversation (TT) and it reflects the same conclusions.

for the computer to finish, so they should be given feedback indicating when the computer expects to be done.

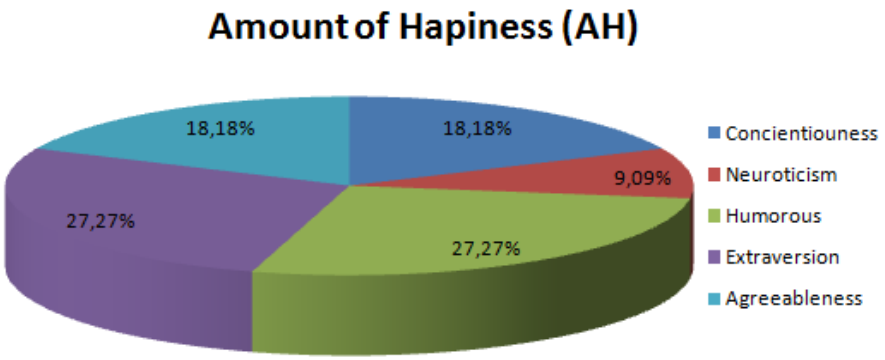


Figure 5.12: Percentage of the amount of happiness from the human (AH), for the different *Social Behavior Profiles* of the robot.

Considering the tolerable computer response time defined by Nah, the results presented in figure 5.11 are clearly fast enough for HRI applications.

The results presented in figure 5.12 means that the satisfaction level of the human regarding both *Humorous* and *Extroverted* is higher than the *Conscientious*. The *Conscientious* SBP represent a robot without emotions, thus it is possible to notice the importance of endowing emotions in this kind of interaction.

The robot's ability of expression has a major effect on the human reaction time during a dialog. By using the assessments defined in section 5.4, we conclude that the human answer faster when when *SBP* is set to *"Humorous"* and *"Extroverted*". The "*Conscientious*" *SBP* is the same as a robot without emotions and results show that in this configuration human response time and total time of the dialog is higher than for *"Humorous"* and "*Extroverted*" SBP. Moreover we conclude that people enjoy more talking to a funny robot and these results encourage to continue research in humorous robots.

Chapter 6

Overall Conclusions and Future Work

The main objective of this work was endowing robots with the capability to, not only recognize human emotions, but also synthesize its own emotions and emulate personality. In the previous chapters, all the necessary components for our multimodal emotion recognition and synthesis were described. Also, a dynamic background segmentation approach was proposed for the robot to be capable of localizing the person to engage. Furthermore, the emotional vector was proposed as a method to add a personality to the robot, increasing the satisfaction level of the user. Bayesian learning was explained for all used Bayesian networks. Two modalities were addressed with methods for analysis and synthesis of emotions, a method for Bayesian mixture of visual and auditory channel was presented. Dialog flow was explained and restrictions were made to allow the dialogue to be plausibly implemented. Classifiers were evaluated and compared with the state of the art. New assessments for measuring the value of the interaction were proposed based on psychological references. The whole system was tested and evaluated accorded with the defined assessments.

An important aspect for intelligence of the robot is the memory. Memory is composed of all knowledge that is stored over the lifetime of the system. One portion of memory dedicated to priors was manually filled before the system starts. A prior was what was believed to be the initial probabilities for a *Bayesian network*. Usually it was assumed to be an uniform distribution. Another portion of memory was dedicated to learning Bayesian classifiers of both visual and auditory channels, namely, the likelihood

was completed through learning. Learning was done by putting the system to run in a non-autonomous fashion while gathering the variables' values. During the learning phase, a human expert takes the "correct" decision or classification for the robot while the variables' values were gathered. In our emotional system, that has been mainly discussed in chapters 2,3 and 4, three macro phases for learning can be defined: learning for analysis, learning for synthesis and learning for the emotional vector. The learning for analysis and synthesis were separated for the visual channel and for the auditory channel. The learning for emotional vector was just one, since the same learning rule applied for both modalities together and one serves as support to the other. Our contribution goes beyond the new models of Bayesian Networks that are presented here; by also taking into account the interpretation of these models.

The work can be reimplemented with the information of the likelihoods tables presented along this thesis. Real time processing is very important for human robot interaction and most extant methods for this kind of perception are not real time. The impact of our methods was measured over a study case (in chapter 5) and it was concluded that our method has satisfactory time of response for human robot interaction applications. Moreover, different *social behavior profiles* of the robot were analyzed over the tests. The *social behavior profile* of the robot has a strong influence on the satisfaction level of the users.

Furthermore, we want to experiment using directional microphones and re-run the assessments defined on section 2.4. We intend to explore the Kinect sensor capabilities to do the zone of interaction segmentation and also to add gestures recognition to our future Social Robot. We want to improve the robotic platform by continuing already extant cooperation with companies (a prospect view of the new platform that are being built by IDMind can be see in figure 6.1). We intend also as future work to made a strong effort over implementation in order to made the system more robust, reliable and encapsulated.

As future/ongoing work, we are working into a project to develop a commercial product of a social robot. This project are scaled to the next 4 years and it will produces a robot that shall include several daily tasks for elderly companionship and support. Among other capabilities of the robot like: navigation, measurements of human behavior, daily

Figure 6.1: Social Robot Project and the prospected robotic platform.

tasks; the social aspects concerning emotions presented in this thesis are going to be included in this product.

In the elderly house, there is already a script of daily exercises. The SocialRobot project incorporates new exercises in this daily agenda, by using the approaches proposed on this thesis. The examples presented in figures 6.2, 6.3 and 6.4 are approved scenarios that shall be part of the SocialRobot project.

The main goal of the SocialRobot project is to provide an answer to the demographic change challenge, through knowledge transfer and the creation of strategic synergies between the project's participating academia and industry regarding the development of an integrated Social Robotics system (SocialRobot) for "Aging Well". The work focuses on bringing together the Robotic and Computer Science fields by integrating state of the art Robotic and Virtual Social Care Communities technologies and services to provide solutions to key issues of relevance for improved independent living and quality of life of elderly people and efficiency of care. The SocialRobot development will be based on a "human centered approach" in which the elderly individual needs and requirements

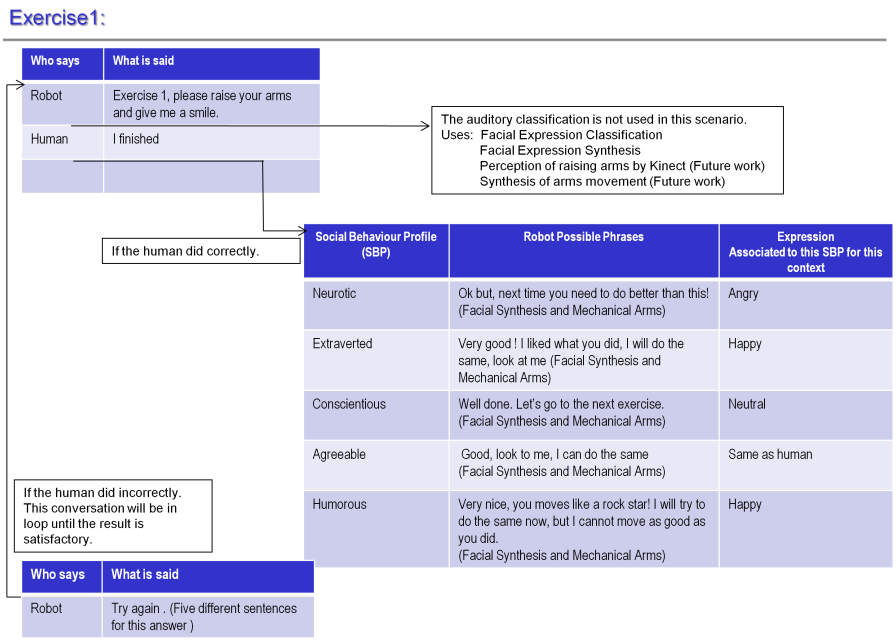


Figure 6.2: Exercise one, to be triggered by the social robot.

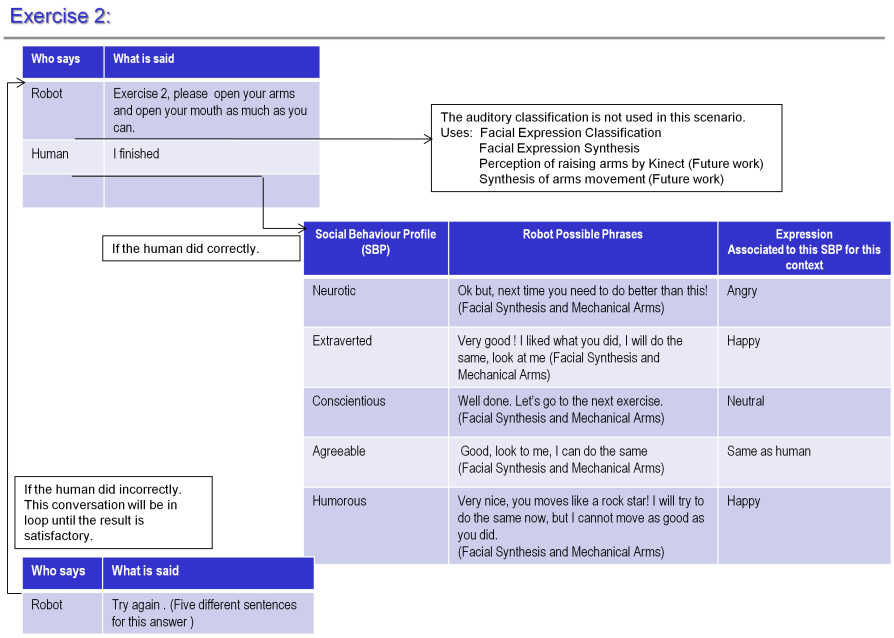


Figure 6.3: Exercise two, to be triggered by the social robot.

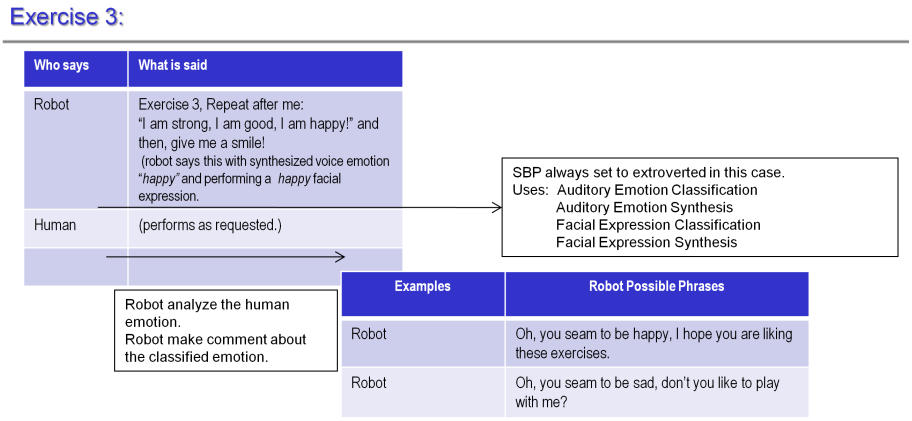


Figure 6.4: Exercise three, to be triggered by the social robot.

are met.

The major challenges to be addressed in the project include the adaptation of state of the art robotic mobile platforms and their integration with a virtual collaborative social network to provide:

- Detection of individual needs and requirements related to aging (e.g. physical mobility limitations or/and cognitive decline), and provision of support through timely involvement of care teams, consisting of different groups of people (family members, neighbors, friends) that collaborate dynamically and virtually; means independently of time and their physical locations; behavior analysis to adapt social relationships and contexts of the elderly people as they age;

- Navigate indoors and unstructured environments and provide affective and empathetic user-robotic interaction, taking into account the capabilities of and acceptance by elderly users.

From the perspective of the academic partners (University of Coimbra and University of Cyprus), the work focuses on bringing together the Robotic and the Computer Science fields through the integration of state of the art robotic and virtual social care community technologies and services. From the perspective of the industrial partners (IDMind and CITARD), this project is an opportunity to develop and provide an integrated state of the art solution with practical relevance to the market addressing key issues of rele-

vance for improved independent living and quality of life of the elderly people. From the perspective of the social care sector and society, the project will provide an innovative *Information and Communication Technology* (ICT) based system that promotes efficiency of elderly care and adaptability to different users.

The initial target group of SocialRobot is that big group of 'late elderly' (As per the World Health Organization, 2007 (http://www.who.int/en/) those are defined to be people in the age 75 or more). The SocialRobot is considering the 'late elderly' with light physical or psychological health problems who need limited support in an inside environment in order to carry out their daily life activities. The aim of choosing the specific target group is to empower and stimulate those people through the provision of ICT based care provision services to continue living as long as possible independent in their preferred environment.

In the last years it has been noted that the market of ICT for aging well, while growing fast, is still in a pre-mature phase and does not yet fully ensure the availability of the necessary ICT-enabled solutions. However the potential of the market is huge, because in many countries the older population has a larger buying power and aging is becoming a global phenomenon. The provision of successful ICT-based solution for aging care, to be applied European wide, is associated with a lot of challenges that can only be tackled by putting together and promote knowledge transfer and strong lasting research collaborations among different sectors in different European regions. Thus, diversities in cultures, elderly care models and practices need to be optimally considered. SocialRobot project addresses the aging care growing high potential market, in Europe and beyond, by providing a well defined know how transfer environment among different sectors (academia, industry, end users) and European regions (Cyprus, Portugal). The SocialRobot approach of ICT-based elderly care can only be achieved, when partners from academia and industry find together to develop, optimize and test such a solution by considering the benefit of the end user. The involvement of the academia sector (University of Cyprus and University of Coimbra) provides for Know-how regarding innovative ICT based research models and technologies in the Robotics. The industrial involvement will lead to an on time to market availability of the project's outcome, so that considerable reductions in care systems will be expected. Without the involvement of the academia,

the industrial sector would have to struggle through a lot of new care related technologies, means investing a lot without being able to foresee an on time commercialization of the project outcome. Furthermore, the consideration of end user requirements and needs, through the validation of the SocialRobot outcome in an end user environment,will help the academia to advance on testing and optimizing the *R&D* models and technologies. The strategic partnerships among academia and research institutes and the industry and the establishment of the know how transfer activities in this project, will be the key stone of introducing in the market the SocialRobot innovative solutions. The outcome will benefit of the end users, in particularly the elderly community, but also of all the different stakeholders related to elderly care provision, and as a consequence a positive impact into the society and economy.

Bibliography

[1] Hadi Aliakbarpour, Pedro Nunez, Jose Prado, Kamrad Khoshhal, and Jorge Dias. An efficient algorithm for extrinsic calibration between a 3d laser range finder and a stereo camera for surveillance. *in Proceedings of the 14th International Conference on Advanced Robotics, ICAR , Munich, Germany*, 2009.

[2] R. Banos, G. Garcia-Soriano, C. Botella, E. Oliver, E. Etchemendy, J. Breton, and M. Alcaniz. Positive mood induction and well being. In *Human System Interactions, 2009. HSI '09. 2nd Conference on*, pages 517 –519, May 2009.

[3] Pro Bayes. Probt - mastering uncertainty. *http://www.probayes.com*, 2012.

[4] Pierre Bessiere, Juan-Manuel Ahuactzin, Kamel Mekhnacha, and Emmanuel Mazer. *Bayesian Programming Book*. 2006.

[5] Pierre Bessiere, Christian Laugier, and Roland Siegwart. *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*. Springer, 2008.

[6] Kim Binsted, Helen Pain, and Graeme Ritchie. Children's evaluation of computer-generated punning riddles. *Department of Artificial Intelligence, University of Edinburgh*, 1997.

[7] Paul Boersma and David Weenink. Eletronic.

[8] Rollo Carpenter. Cleverbot, 1997.

[9] Gregory J. Chaitin. *Meta math!: the quest for omega*. Pantheon Books, the University of Michigan, 2010.

[10] I. Cohen, N. Sebe, A.Garg, M.S. Lew, and T.S Huang. Facial expression recognition from video sequences. *in proc. ICME*, pages 121–124, 2002.

[11] Perry R. Cook, Dexter Morrill, and Julius O. Smith. An automatic pitch detection and midi control system for brass instruments. *Invited for special session on Automatic Pitch Detection, New Orleans*, 1992.

[12] R. Cowie, E.Douglas-Cowie, K. Karpouszis, G. Caridakis, M. Wallace, and S. Kollias. Recognition of emotional states in natural human-computer interaction. *School of Psychology, Queen's University*, 2007.

[13] Inc. Crown Industries. A.l.i.c.e., 2012.

[14] N. Dahlbock, A. Jonsson, and L. Ahrenberg. Wizard of oz studies: Why and how. *In Proceedings of the International Workshop on Intelligent User Interfaces, Orlando, FL, ACM Press*, pages 193–200, 1993.

[15] Antonio Damasio. *The Feeling of what happens*. Harcourt, Inc - ISBN 978-0-15-601075-7, 2000.

[16] Antonio Damasio. *Looking for Spinoza*. Harcourt, Inc - ISBN 978-0-15-100557-4, 2003.

[17] C. R. Darwin. *The expression of the emotions in man and animals*. London: John Murray, 1872.

[18] Rolando Grave de Peralta Menendez, Jorge Manuel Miranda Dias, Jose Augusto Soares Prado, Hadi Aliakbarpour, and Sara Gonzalez Andino. Multiclass brain computer interface based on visual attention. *17th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2009.

[19] Ian J. Deary, Alexander Weiss, and G. David Batty. Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health

inequalities. *Journal of Psychological Science in the Public Interest intelligence, personality, and health outcomes*, 11-2:53–79, 2010.

[20] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. *Facial Action Coding System - The Manual*. A Human Face, 2002.

[21] Paul Ekman and W.V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ma: Malor Books, 2003, 2003.

[22] Paul Ekman and E.L. Rosenberg. *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press. Second expanded edition, 2004.

[23] Marc O. Ernst and Heinrich H. Bulthoff. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169, 2004.

[24] Kathinka Evers. The empathetic xenophobe: A neurophilosophical view on the self. *Centre for Research Ethics and Bioethics (CRB), Uppsala University. The text is adapted from Chapter 3 in Evers (2009): Neuroethique. Quand la matiere s eveille, Editions Odile Jacob, Paris, and was originally presented in an earlier version at College de France, Paris, 2006.*, 2009.

[25] Diego Faria, Jose Prado, Paulo Drews, and Jorge Dias. Object shape retrieval through grasping exploration. *In Proceedings of the 4th European Conference on Mobile Robots, ECMR09, Mlini/Dubrovnik, Croatia*, 2009.

[26] Joao Filipe Ferreira, Jose Prado, Jorge Lobo, and Jorge Dias. Multimodal active exploration using a bayesian approach. *in Proceedings of the 14th IASTED International Conference in Robotics and Applications, Cambridge MA, USA*, 2009.

[27] Sebastien George and Pascal Leroux. An approach to automatic analysis of learners social behavior during computer-mediated synchronous conversations. In Stefano Cerri, Guy Gouarderes, and Fabio Paraguacu, editors, *Intelligent Tutoring Systems*, volume 2363 of *Lecture Notes in Computer Science*, pages 630–640. Springer Berlin / Heidelberg, 2002.

[28] Rafael C Gonzales. *Processamento de Imagens Digitais*. 2003.

[29] J. Gratch, S. Marsella, and Petta. Modeling the cognitive antecedents and consequents of emotion. *Cognitive Systems*, 10(1):1–5, 2008.

[30] W. Hess. Pitch determination of speech signals. *Berlin: Springer Verlag*, 1983.

[31] C. Hoelper, A. Frankort, and C. Erdmann. Voiced/unvoiced/silence classification for offline speech coding. *in Proceedings of international student conference on electrical engineering (Prague)*, 2003.

[32] C. Hoelper, A. Frankort, and C. Erdmann. Voiced/unvoiced/silence classification for offline speech coding. *in Proceedings of international student conference on electrical engineering (Prague)*, 2003.

[33] Intel. Intel open source computer vision library. http://www.intel.com, 2006.

[34] Singular Invertions. eletronic, 2010.

[35] Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible*. Inference. Morgan Kaufmann Publishers, Inc. (Elsevier), revised second, 1988.

[36] Jose Prado Jorge Lobo, Joao Filipe Ferreira and Jorge Dias. Robotic implementation of biological bayesian models for visuo-inertial image stabilization and gaze control. *International Conference on Intelligent RObots and Systems, Nice, France*, 2008.

[37] Mehran Kafai and Bir Bhanu. Dynamic bayesian networks for vehicle classification in video. *IEEE Transactions on Industrial Informatics*, 2011.

[38] Miyuki Kamachi, Michael Lyons, and Jiro Gyoba. The japanese female facial expression (jaffe) database, 1998.

[39] Kanade, Cohn, and Tian. Cohn-kanade au-coded facial expression database, 2000.

[40] Alice S.M Kau, Elaine Tierney, Irena Bukelis, Mariah H Stump, Wendy R KAtes, William H Trescher, and Walter E Kaufmann. Social behaviour profile in young

males with fragile x synfrome: Charactetistics and specificity. *American Journal of Medical Genetics*, 126:9–17, 2004.

[41] C. D. Kidd and C. Breazeal. A robotic weight loss coach. *In Proceedings of the Twenty-Second Conference on Artificial Intelligence, Menlo Park, CA, AAAI Press*, 2007.

[42] S. Klemmer, A. Sinha, J. Chen, J. Landay, N. Aboobaker, and A. Wang. Suede: A wizard of oz prototyping tool for speech user interfaces. *In CHI Letters: Proceedings of the ACM Symposium on User Interface Software and Technology*, 2:1–10, 2000.

[43] Chul Min Lee, Shrikanth S. Narayanan, and Roberto Pieraccini. Classifying emotions in human-machine spoken dialogs. *ICME*, 2002.

[44] Peter A. Levine. *Waking the Tiger - Healing Trauma*. North Atlantic Books, 1997.

[45] A.V. Libin and E.V. Libin. Person-robot interactions from the robopsychologists' point of view: the robotic psychology and robotherapy approach. *Proceedings of the IEEE*, 92(11):1789 – 1803, nov. 2004.

[46] Rainer Lienhart and Jochen Maydt. An extendeed set of haar-like features for rapid object detection. *ICIP*, 2002.

[47] Jorge Lobo, Joao Filipe Ferreira, and Jorge Dias. Bioinspired visuovestibular artificial perception system for independent motion segmentation. *In 2nd International Cognitive Vision Workshop*, 2006.

[48] C. Lopes and Perdigao F. On the use of pitch to perform speaker normalization. *Proc. International Conf. on Telecommunications, Electronics and Control, Santiago de Cuba, Cuba*, 2002.

[49] C. Lopes and F. Perdigao. Vtln through frequency warping based on pitch. *ProcIEEE International Telecommunications Symp., Natal, Brazil*, 2002.

[50] C. Lopes and F. Perdigao. Vtln through frequency warping based on pitch. *revista da Sociedade Brasileira de Telecomunicacoes*, 18-1:86–95, 2003.

[51] Nicolas Lori and Alex Blin. Application of quantum darwinism to cosmic inflation: An example of the limits imposed in aristotelian logic by information-based approach to godel's incompleteness. *Foundations of Science*, 15:199–211, 2010.

[52] Nicolas F. Lori and Paulo Jesus. Matter and selfhood in kant's physics: A contemporary reappraisal. *Relations of the Self. Edmundo Balsemão Pires, Burkhard Nonnenmacher, and Stefan Büttner-von Stülpnagel (ed.). Imprensa da Universidade de Coimbra*, pages 207–226, 2010.

[53] M.J. Lyons, J. Budynck, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions PAttern Anal. Machine Intell.*, 21:1357–1362, 1999.

[54] Tarek M Mahmoud. A new fast skin color detection technique. In *World Academy of Science*, 2008.

[55] Miriam Martinez and L. Enrique Sucar. Learning dynamic naive bayesian classifiers. *Proceedings of the Twenty-First International FLAIRS*, 2008.

[56] J. Meessen, C. Parisot, C. Lebarz, D. Nicholson, and J.F. Delaigle. Smart encoding for wireless video surveillance. In *In SPIE Proc. Image and Video Communications and Processing*, volume 1, 2005.

[57] P. Menezes, F. Lerasle, and J. Dias. Towards human motion capture from a camera mounted on a mobile robot. *Image and Vision Computing*, 29-6:382–393, 2011.

[58] Michael Montemerlo, Nicholas Roy, and Sebastian Thrun. Perspectives on standarization in mobile robot programming: The carnegie mellon navigation (carmen) toolkit. *School of Computer Science Carnegie Mellon University Pittsburgh*, 2003.

[59] Kevin Murphy. Bayes net toolbox for matlab. *http://code.google.com/p/bnt/*, 2002.

[60] Fiona Fui-Hoon Nah. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour Information Technology*, 23(3):153–163, 2004.

[61] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *ICPR*, pages 3695–3699, 2010.

[62] L R Oliveira and Urbano Nunes. On integration of features and classifiers for robust vehicle detection. *IEEE Conference on Intelligent Transportation Systems (ITSC08)*, 2008.

[63] L R Oliveira and Urbano Nunes. On using cell broadband engine for object detection in its. *IEEE International Conference on Intelligent RObots Systems (IROS)*, pages 54–58, 2008.

[64] R. Paixao, L. Coelho, and J. Ferreira. Teste de reconhecimento paralinguistico das emocoes. *Psychologica*, 53:232–254, 2010.

[65] Gayatri Paknikar. *FACIAL IMAGE BASED EXPRESSION CLASSIFICATION SYSTEM USING COMMITTEE NEURAL NETWORKS*. PhD thesis, The Graduate Faculty of The University of Akron, 2008.

[66] Maja Pantic. Facial expression recognition. In *Encyclopedia of Biometrics*, pages 400–406, 2009.

[67] Maja Pantic and Leon Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of IEEE*, 91, NO. 9, September 2003.

[68] Xavier Perrin, Ricardo Chavarriaga, Roland Siegwart, and Jose del R. Millan. Bayesian controller for a novel semi-autonomous navigation concept. *ECMR*, 2007.

[69] Minh Tri Pham and Tat Jen Cham. Fast training and selection of haar features using statistics in boosting-based face detection. *In Proc. 11th IEEE International Conference on Computer Vision (ICCV'07)*, 2007.

[70] J. Prado, J.Lobo, and J.Dias. Sophie: Social robotic platform for human interactive experimentation. In *4th International Conference on Cognitive Systems, COGSYS 2010, ETH Zurich, Switzerland*, 2010.

[71] Jose Prado and Jorge Dias. Visuovestibular-based gaze control experimental case. *RECPAD 14a Conferencia Portuguesa de Reconhecimento de Padroes, Coimbra*, 2008.

[72] Jose Prado, Luis Santos, and Jorge Dias. Horopter based dynamic background segmentation applied to an interactive mobile robot. *14th International Conference on Advanced Robotics, ICAR09, Munich, Germany*, 2009.

[73] Jose Prado, Luis Santos, and Jorge Dias. A technique for dynamic background segmentation using a robotic stereo vision head. *RO-MAN 2009, 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama International Conference Center, Japan*, 2009.

[74] Jose Prado, Lakmar Seneviratne, and Jorge Dias. Synthesis of emotions on a human-robot-interactive platform. *in Proceedings of IASTED, Robo2011, The 16th IASTED International Conference on Robotics, Pittsburg, USA*, 2011.

[75] Jose Prado, Carlos Simplicio, Nicolas Lori, and Jorge Dias. Visuo-auditory multimodal emotion classifiers to improve human robot interaction. *to be submitted to International Journal of Social Robots*, 2011.

[76] Jose Augusto Prado, Carlos Simplicio, and Jorge Dias. Robot emotional state through bayesian visuo-auditory perception: focus on auditory perception. *In proceedings of - DOCEIS 2011 - Doctoral Conference on Computing Electrical and Industrial Systems*, 2011.

[77] Iain Matthews Ralph Gross and Simon Baker. Active appearance models with occlusion. *Image and Vision Computing*, 24:593–604, 2006.

[78] Joerg Rett. *Robot-Human Interface Using Laban Movement Analysis Inside a Bayesian Framework*. PhD thesis, University of Coimbra, 2009.

[79] Charles Rich and Candace Sidner. Robots and avatars as hosts, advisors, companions, and jesters. *Advancement of Artificial Intelligence. ISSN 0738-4602*, 2009.

[80] G. Ritchie. Prospects for computational humor. *In Proceedings of 7th IEEE International Workshop on Robot and Human Communication*, pages 283–291, 1998.

[81] S. Rougeaux and Y. Kuniyoshi. Velocity and disparity cues for robust real-time binocular tracking. *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 1997.

[82] Luis Santos, Jose Augusto, and Jorge Dias. Human robot interaction studies on laban human movement analysis and dynamic background segmentation. *IROS, The IEEE/RSJ International Conference on Intelligent RObots and Systems, St Louis, USA*, 2009.

[83] Marc Schroder. The semaine api: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction, Article ID 319406, 21 pages. doi:10.1155/2010/319406*, 2010, 2010.

[84] N. Sebe, M.S. Lew, I. Cohen, A.Garg, and T.S Huang. Emotion recognition using a cauchy naive bayes classifier. *in proc. ICPR*, 1:17–20, 2002.

[85] Bau-Cheng Shen, Chu-Song Chen, and Hui-Huang Hsu. Face image retrieval by using haar features. *Pattern Recognition ICPR*, pages 1–4, 2008.

[86] Carlos Simplicio, Jose Prado, and Jorge Dias. Comparing bayesian networks to classify facial expressions. *in Proceedings of RA-IASTED, The 15th IASTED International Conference on Robotics and Applications, Cambridge, Massachusetts, USA*, 2010.

[87] Carlos Simplicio, Jose Prado, and Jorge Dias. A face attention technique for a robot able to interpret facial expressions. In *DOCEIS 2010 - Doctoral Conference on Computing Electrical and Industrial Systems*, 2010.

[88] Carlos Simplicio, Jose Augusto Prado, and Jorge Dias. A face attention technique for a robot able to interpret facial expressions. *in Proceedings of the DoCEIS'10 - Doctoral Conference on Computing, Electrical and Industrial Systems. Lisbon.*, Springer - ISBN 978-3-642-11627-8, 2010.

[89] Mohan Sondhi. New methods of pitch extraction. *IEEE Transactions on audio and electroacoustics*, 16:262–266, 1968.

[90] Spinoza. *Ethics*. 1677.

[91] O. Stock and C. Strapparava. Getting serious about the development of computational humor. *In proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 59–64, 2003.

[92] Oliviero Stock and Carlo Strapparava. The act of creating humorous acronyms. *Journal of Applied Artificial Intelligence*, 19:137–151, 2005.

[93] Lambert M Surhone, Miriam T Timpledon, and Susan F Marseken. *Horopter*. VDM Publishing House, 2010.

[94] Barry-John Theobald, Iain Matthews, and Simon Baker. Evaluating error functions for robust active appearance models. *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1:149–154, 2006.

[95] Theodore Totozafiny, Olivier Patrouix, Franck Luthon, and Jean-Marc Coutellier. Dynamic background segmentation for remote reference image updating within motion detection jpeg2000. In *ICIP International Conference on Image Processing*, volume 1, 2008.

[96] B. C. Tsai, W. W. Wang, L. C. Hsu, L. C. Fu, and J. S. Lai. An articulated rehabilitation robot for upper limb physiotherapy and training. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 1470 –1475, oct. 2010.

[97] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162 – 1181, 2006.

[98] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[99] Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE CVPR*, 2001.

[100] Yongjin Wang and Ling Guan. Recognizing human emotion from audiovisual information. *ICASSP IEEE*, 2005.

[101] Wuhan. Facial expression recognition based on local binary patterns and coarse-to-fine classification. *Fourth International Conference on Computer and Information Technology (CIT'04)*, 16, 2004.

[102] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Machine Intelligence*, 24:34–58, 2002.

[103] S. Zielinskin. Papers from work on comb transformation method of pitch detection ("description of assumptions of comb transformation", "comb transformation - implementation and comparison with another pitch detection methods"). *Technical University of Gdansk*, 1997.

Appendix A

Comparing Histograms with Bhattacharyya distance

Let's call A the output emotion vector composed by 10 elements (3 variables from sound plus 7 from image, see 5.6 for description of those elements). When the robot synthesizes A, it does it according to the likelihood tables previously filled out. Later A is classified using the same Bayesian networks proposed for human emotion classification.

Therefore we have $P(A)$ which is a discrete probability distribution, of A, in the scope X=*{happy, sad, fear, neutral, anger}*. Let's call $P(R)$ the expected discrete probability distribution. Since it is expected one of the five expressions, $P(R)$ will always has 0.96 probability on the expected expression and 0.01 at the other four.

According to [57] the Bhattacharyya distance D_B is the best metric to compare histograms. Thus we selected this metric to compare our histograms and our error function is given by:

$$E = D_{B_{(A,R)}} = -ln \sum_{x \in X} \sqrt{P(A)_x P(R)_x}$$

It is difficult to assess how the interaction is improving. On literature, [91, 92]

claims that a interaction is better when the person consider the system to be funny. Specially those from the European project called “Hahacronym”, we found descriptions of results but no specific descriptions of assessments. However, it is clear that they performed an experiment with various persons, while an independent agent do a manual ranking of how happy was the person with the performance of that system. In [80] details of assessments are also clear where the system was shown to children and what was considered as a joke was also manually measured (by questionnaires after the dialog), they develop an assessment procedure for measure the “jokiness” of each response proposed on [6]. Previously on [6] it was measured the average of “jokiness”, “funniness” also “heard before” possible classifications for each text, according to their defined assessments. The “jokiness” could be scored from zero to one. For “funniness” the scope was defined from one to five. For “heard before” the range of the score was from zero to one.

Appendix B

Calculation of Bayesian Probabilities

There are several tools used to calculate the Bayesian probabilities, like for example [3] and [59], however this section presents some examples of how to compute the probabilities for a small Bayesian Program without using any external tool.

Let's use a simple example where we want to classify a facial expression among a binary scope *{happy or fear}* based only in 2 variables, *LC* (Lips Corners) and *MA* (Mouth Aperture). The real values of *LC* can be between 0 and 43, because the image trunked in the mouth only have $70x43$ pixels of resolution. The value of *LC* is the pixel height coordinate where the lip corner was found. The Mouth Aperture value may vary among 0 and 43 also. *MA* is the difference between the pixel height coordinate from bottom mouth and pixel height coordinate of top mouth.

In our example, the $P(A|B)$ becomes $P(FE|LC,MA)$ and it means the probability of every facial expression (of the defined scope) given the current values of *LC* and *MA*. Thus, it is necessary to calculate $P([FE = happy]|MA,LC)$, and also $P([FE = fear]|MA,LC)$.

The first step on the Bayesian Programming implementation is to perform an initial learning. The Bayes formula represented on figure B.1, only works with a likelihood to infer, a prior and a normalizing factor. The likelihood, in this case, is an histogram

Table B.1: SAMPLE OF LIKELIHOOD GATHERING DURING LEARNING OF SIMPLE EXAMPLE.

happy.txt		fear.txt	
LC	MA	LC	MA
1	1	20	7
1	2	21	7
2	3	22	3
1	3	19	10
2	3	22	7

distribution given by the learning.

B.1 Learning Phase

The purpose of the learning phase is to collect variable values that somehow characterizes what we want to classify. To gather the necessary data for later creating the likelihood histograms, it is necessary to save the current values of the variables, in this case *LC* and *MA* into some permanent area of storage, a file for example. Thus we run the program that is capable of detecting the points, and by using the interface created by us to teach the system, we click in a button saying what is the current expression. The user smiles for the camera and clicks in a button that says “this is happy”. When this button is pressed, the values of *LC* and *MA* are saved into a file called “happy.txt”. The same is done for the *fear* expression and the values are saved in a different file called “fear.txt”. The layout of these files is like what is presented in table B.1.

Notice in table B.1, for example, that the *happy* expression is being characterized by high lips corners (*LC*) and small mouth aperture (*MA*). The pixel 1 or 2 in the Lips Corners indicates that the lips are in position 1 or 2 among 43, what is high. This is according to what was expected, since when the person performs a *happy* expression, the lips corners elevates. The mouth aperture is also expected to be small, since when the person smiles, the mouth is not opened or maybe it is just slightly opened. An analogous interpretation can be done over the file “fear.txt”. Values for *LC* near 20 means that the lips corners are positioned close to the center of the image, vertically. While values near

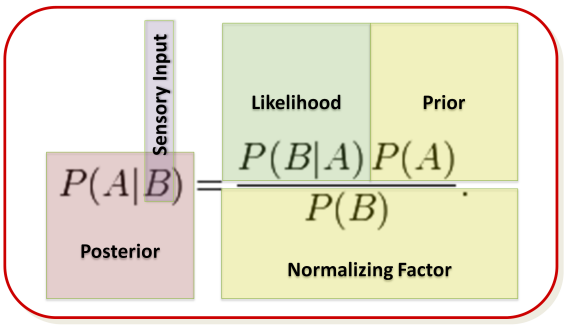


Figure B.1: Bayes Rule.

7 or 10 for the *MA* means that the mouth is opened as expected at the *fear* expression.

B.2 Inference

B.2.1 Likelihood calculation

After doing the learning, we already have the necessary data to start inferring. At the inference phase, system is running and the objective is to obtain $P(FE|[LC = known], [MA = known])$. In order to do the inference, first step is to calculate the likelihood based on the learned data. So we want to have values for $P(B|A)$ from the Bayesian Rule as shown in figure B.1. In our case, $P(B|A)$ is in fact $P(LC, MA|FE)$.

$$P(LC, MA|FE) = P(LC|FE) * P(MA|FE)$$

It is necessary to know what is the probability of the current values of *MA* and *LC* belong to each of the possible facial expressions. Thus, we have to calculate the probability of current value of *LC* and *MA* belongs to a given $[FE = happy]$ and probability of current value of *LC* and *MA* belongs to a given $[FE = fear]$. The probability is the number of matches of the current value among the samples collected during the learning, divided by the total samples of the learning. Therefore, $P(LC|[FE = happy])$ and $P(LC|[FE = fear])$ can be computed with data coming from two different histograms

(one for *fear*, another for *happy*).

A small example of likelihood for one frame:
Let's state an example of the likelihood calculation for a frame where the *LC* was detected with value 2, and *MA* with value 3. What is the probability of *LC* belong to *happy* expression according to learning files presented on table B.1? The $P(LC|[FE = happy])$ is $2/5 = 40\%$. And what is the probability of *LC* belongs to fear expression? The $P(LC|[FE = fear])$ is $0/5 = 0\%$. The $P(MA|[FE = happy])$ is $3/5 = 60\%$ and the $P(MA|[FE = fear])$ is $1/5 = 20\%$. Some adjust is necessary because we can never accept a probability to be zero. In this case we manually fix the lower limit of probability to be $0.01 = 1\%$. The matching exemplified here is static, however the matching can also be done with a threshold tolerance instead of exact match.

The final likelihood for this frame is then: $0,40 * 0,60 = 0,24$ for *happy*, and $0,01 * 0,20 = 0,002$ for *fear*. At this point it is already possible to guess that this frame will probably be classified as *happy*, however it will still be multiplied by the prior and normalized.

B.2.2 Prior calculation

The purpose of the prior is to have some knowledge about what is happening before accepting the current classification given by the likelihood. In practice, the prior is a distribution to be multiplied by the likelihood.

The prior represented by $P(A)$ in the Bayesian generic formula, in the case of our example is $P(FE)$. When nothing is know, since our distributions are discrete, prior is an uniform histogram distribution. If the previous example is the first frame of a video sequence, our prior will be 50% happy and 50% fear. We assume prior as being uniform at the first frame, later the classification of the previous frame will enter as prior information for the current frame. This approach (from hidden Markov models) of propagating the previous frame classification as prior of the current frame avoids that erroneous frames

interfere into the classification. After the user keeps the expression for some frames, it converges to the new expression. This is the big advantage of using dynamic Bayesian networks.

B.2.3 Normalization

After calculating the likelihood and multiplying by the prior, we already have our classification done. However it is not normalized. To be normalized the sum of all probabilities should be 1= 100%. According to Bayesian marginalization rule, the normalization factor $P(B)$, which in our example case is $P(LC,MA)$ can be expanded in to the form of:

$$P(LC,MA) = \sum_{FE} P(LC|FE) * P(MA|FE) * P(FE)$$

In this sum, each instance of FE is taken into account, this is $P(LC|[FE = happy])$ and $P(LC|[FE = fear])$. The same applies to $P(MA|FE)$ where the instances are: $P(MA|[FE = happy])$ and $P(MA|[FE = fear])$.

After calculating the normalization factor, the result will be histogram with all the probabilities for all the possible classification scope, normalized, where the sum is 100%.