Expert Systems With Applications 66 (2016) 42–61

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multimodal emotion recognition with evolutionary computation for human-robot interaction

Luis-Alberto Perez-Gaspar, Santiago-Omar Caballero-Morales*, Felipe Trujillo-Romero

Technological University of the Mixteca, Road to Acatlima K.m. 2.5, Huajuapan de León, Oaxaca, Mexico,69000, Mexico

ARTICLE INFO

Article history:
Received 4 December 2015
Revised 12 August 2016
Accepted 13 August 2016
Available online 3 September 2016

Keywords:
Emotion recognition
Principal Component Analysis
Hidden Markov Models
Genetic Algorithms
Artificial Neural Networks
Finite state machines

ABSTRACT

Service robotics is an important field of research for the development of assistive technologies. Particularly, humanoid robots will play an increasing and important role in our society. More natural assistive interaction with humanoid robots can be achieved if the emotional aspect is considered. However emotion recognition is one of the most challenging topics in pattern recognition and improved intelligent techniques have to be developed to accomplish this goal. Recent research has addressed the emotion recognition problem with techniques such as Artificial Neural Networks (ANNs)/Hidden Markov Models (HMMs) and reliability of proposed approaches has been assessed (in most cases) with standard databases. In this work we (1) explored on the implications of using standard databases for assessment of emotion recognition techniques, (2) extended on the evolutionary optimization of ANNs and HMMs for the development of a multimodal emotion recognition system, (3) set the guidelines for the development of emotional databases of speech and facial expressions, (4) rules were set for phonetic transcription of Mexican speech, and (5) evaluated the suitability of the multimodal system within the context of spoken dialogue between a humanoid robot and human users. The development of intelligent systems for emotion recognition can be improved by the findings of the present work: (a) emotion recognition depends on the structure of the database sub-sets used for training and testing, and it also depends on the type of technique used for recognition where a specific emotion can be highly recognized by a specific technique, (b) optimization of HMMs led to a Bakis structure which is more suitable for acoustic modeling of emotion-specific vowels while optimization of ANNs led to a more suitable ANN structure for recognition of facial expressions, (c) some emotions can be better recognized based on speech patterns instead of visual patterns, and (d) the weighted integration of the multimodal emotion recognition system optimized with these observations can achieve a recognition rate up to 97.00 % in live dialogue tests with a humanoid robot.

1. Introduction

Human perception regarding emotion detection is a natural process that involves social interactions. Technological advances are in constant evolution and the future aims to a life with robotic agents capable of understanding people. Technology is growing very fast and research on the development of service robots for elderly people or people with motor disabilities (Odashima et al., 2008; PAL Robotics, 2015) is being performed and reported on the literature.

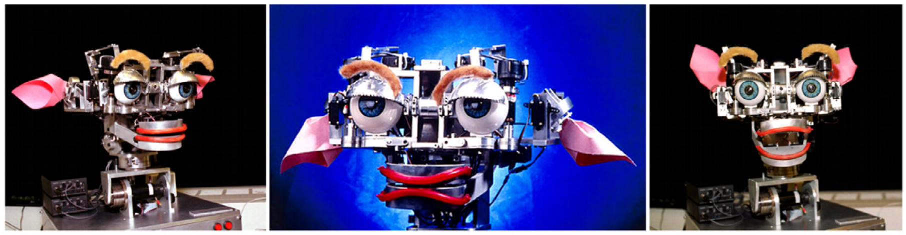
In recent years, research on the affective relation between a robot and a human has become an important subject. Developments on artificial intelligence have been focused on trying to emulate how humans interact with each other. Emotions are fundamental for social interaction and a robot with this capability can take the role of a companion entity that can support conversation, understanding, and responses aimed to improve the well-being of the human user. Emotion recognition can lead to the development of systems for more natural, understandable and intuitive communication (Samani & Saadatian, 2012).

Among the robotic systems that integrate emotion recognition for an interaction task, the following can be mentioned:

- Robot Kismet (Breazeal, 2003): This social robot was developed by Cynthia Breazeal at the Massachusetts Institute of Technology (MIT). This robot was developed to study how emotions expressed by a robotic system changed the perception and interaction with human users. Kismet was able to recognize affective intentions through the voice (Anger, Fear, Happiness, Tiredness,

* Corresponding author.
E-mail addresses: luis_335450@hotmail.com (L.-A. Perez-Gaspar), scaballero1979@yahoo.com (S.-O. Caballero-Morales), ftrujillo@mixteco.utm.mx (F. Trujillo-Romero).

http://dx.doi.org/10.1016/j.eswa.2016.08.047
0957-4174/



(a) Kismet

(b) Jibo

(c) MOUE

Fig. 1. Examples of artificial systems with emotion recognition.

Disgust, Surprise, Sadness, Interest and Calm) and express different emotions as presented in Fig. 1(a).
- Jibo (Chambers et al., 2015): This social robot was an extension of Kismet and it was conceived as a companion robot for commercial purposes. As presented in Fig. 1(b) Jibo *could* perform different tasks like taking photographs, reminding important events, telling stories and establishing video conference between relatives using Wi-Fi.
- Model Of User's Emotions (MOUE) (Lisetti et al., 2003): This was an intelligent interface for distance patient monitoring. This system was able to capture physiological signals and emotional gestures through a bracelet connected to a computer and a web-cam. Then this data was collected and sent to a central computer for processing. As presented in Fig. 1(c) this system enabled question-based interaction with an animated character (avatar) that acted as a mirror by reflecting the facial expressions performed by the patient. The interface was able to process the emotions of Neutral, Anger, Fear, Sadness and Frustration.

The development of robotic systems with the capability of emotion recognition depends on the use and adaptation of pattern recognition techniques. Among the most common techniques the following can be mentioned: Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Principal Component Analysis (PCA), Fuzzy Logic, Hidden Markov Models (HMMs) and Linear Discriminant Analysis (LDA). In Tables 1 and 2 a review of works on emotion recognition is presented. This review includes information regarding the number of emotions that were detected, the patterns (voice, facial expressions) that were considered, and the recognition techniques that were applied.

Further improvement on emotion recognition can be achieved by the development of multimodal systems that integrate expressions and voice patterns. In (Song et al., 2008) a multimodal system was built with Tripled HMMs (THMMs) for the recognition of the following emotions: Surprise, Happiness, Anger, Fear, Sadness and Neutral. The use of THMMs was performed to synchronize voice and facial features in the time domain. For the vision system, the distances between eyes, eye-nose, mouth-nose and width of the mouth were considered as the most important features. For the speech system, 48 prosodic and 16 formant features were extracted. The recognition rates were 87.40% and 81.45% for the vision and speech systems respectively. However, when an integration of both systems was realized an increase in the recognition rate was accomplished, leading to a final recognition rate of 93.31%. Similar improvements were reported in (Busso et al., 2004; Haq et al., 2008) for the recognition rate of four and seven emotions respectively. In Table 3 a review of works on the development of multimodal systems is presented.

From the works presented in Tables 1, 2 and 3 some observations can be highlighted:

- In general, the recognition rates of the vision systems are higher than those of the speech systems (Anagnostopoulos et al., 2015). However a multimodal system can achieve higher recognition rates than those of the individual vision or speech systems (Busso et al., 2004; Song et al., 2008).
- Most of the emotion recognition works on facial expressions considered specialized databases like FEEDTUM (Filko & Martinovic, 2013; Pal & Hasan, 2014), JAFFE (Gosavi & Khot, 2013; Kaur, Vashisht, & Neeru, 2010; Pooja & Kaur, 2010; Rasoulzadeh, 2012; Thuseethan & Kuhanesan, 2014), FACES (Tayal & Vijay, 2012), CK+ (Chaturvedi & Tripathi, 2014), and RaFD (Ilbeygi & Hosseini, 2012). These works also reported the highest emotion recognition rates.
- While most of the emotion recognition works on facial expressions consider six emotions (Chaturvedi & Tripathi, 2014; Gosavi & Khot, 2013; Ilbeygi & Hosseini, 2012; Pal & Hasan, 2014; Rasoulzadeh, 2012; Thuseethan & Kuhanesan, 2014), most of the works on speech recognition consider four (Caballero, 2013; Firoz-Shah et al., 2009; Lee et al., 2004; Wu & Liang, 2011; Yu et al., 2001) and five (Austermann et al., 2005; Chaavan & Gohokar, 2012; Pao et al., 2007; Yu, 2008) emotions.
- Most of the speech corpora used for emotion recognition are available in languages different from Spanish. Hence, speech