Artificial Intelligence for Human-Robot Interaction
AAAI Technical Report FS-17-01

Towards Semantic Multimodal Emotion Recognition for Enhancing Assistive Services in Ubiquitous Robotics

Naouel Ayari, Hazem Abdelkawy, Abdelghani Chibani, Yacine Amirat
LISSI Laboratory
University of Paris-Est Créteil (UPEC),
Vitry-sur-Seine France
{naouel.ayari,hazem-khaled-mohamed.abdelkawy,abdelghani.chibani,amirat}@u-pec.fr

Abstract

In this paper, the problem of endowing ubiquitous robots with cognitive capabilities for recognizing emotions, sentiments, affects and moods of humans, in their context, is studied. A hybrid approach based on multilayer perceptron (MLP) neural network and n-ary ontologies for emotion-aware robotic systems is proposed. In particular, an algorithm based on the hybrid-level fusion, an expressive emotional knowledge representation and reasoning model are introduced to recognize complex and non-observable emotional context of the user. Empirical experiments on real-world dataset corroborate its effectiveness.

Introduction

One of the main challenging research problems in assistive robotics is to endow ubiquitous robots with ability to proactively taking part on some tasks to help human and provide them services according to their context (Chibani et al. 2013). The concept of context awareness is very important in the design of cognitive capabilities of assistive robots or agents (Dey and Abowd 2000).

Considering *emotional* or *affective* aspects is fundamental for natural assistive interaction where robots act as companion entities that can support conversation, understanding, and responses (Samani and Saadatian 2012). The common techniques of emotion recognition are based on data-driven techniques, such as the Artificial Neural Network (ANN) (Makioka et al. 2016), the Fuzzy logic (Nicolai and Choi 2015), the Hidden Markov Model (HMM) (Gorlova and Gultyaeva 2016), etc. However, generally these techniques are deeply linked with their training data. It becomes difficult to discern between the meanings of an emotion or to recognize a non-observable emotion. This latter can't be interpreted in an accurate way without considering its context.

The development of robotic systems with the capability of affect/emotion recognition requires sophisticated and novel approaches. These systems need to be endowed with advanced emotional knowledge representation and reasoning capabilities. To enable an efficient recognition of emotions in a dynamic ambient environment, an exhaustive emotion-sensing and multimodal fusion are required. Building efficient cognitive models for emotion-aware robotic systems requires a suitable architecture.

In this paper, a hybrid approach based on multilayer perceptron (MLP) neural network and n-ary ontologies for emotion-aware robotic systems is proposed to enhance assistive services in intelligent ambient environments. The main contributions of the paper is a cognitive architecture for emotion-aware robotic systems to endow robots with the capabilities of sensing and understanding an emotional context for better decision-making. In particular, the novelty concerns: (i) a model based on the hybrid-level fusion for online multimodal emotion recognition where temporal synchronization is guaranteed. This model extracts different modalities, such as, audio, text, facial expression, culture, age and gender, and exploits the multilayer perceptron neural network algorithm to recognize the observed emotions, (ii) an expressive model for commonsense knowledge representation and reasoning on emotions. This model exploiting the narrative knowledge representation language (NKRL) (Zarri 2009), allows, on the one hand, a complete, coherent and expressive emotional knowledge representation, and on the other hand, inference of new contextual emotions and better decision-making. N-ary ontologies in which this model is based, overcome the problems encountered in the emotional context modeling approaches based on binary ontologies.

The paper is structured as follows: First, we review the related works concerning emotion recognition in the robotics field. Then, we describe the proposed approach for multimodal emotion recognition and the proposed approach for emotional context management based on n-ary ontologies. We evaluate the performance of the proposed approach with extensive experiments on real-world dataset and we discuss the obtained results. This paper is concluded with a short review of the proposed approach and a summary of the ongoing works.

Related work

Emotions, sentiments, affects, and moods are tightly coupled terms which were differentiated in (Jacko 2012). Six universal human emotion/affect states are defined in (Darwin, Ekman, and Prodger 1998) to be happiness, fear, sadness, surprise, disgust, and anger.

Many researchers have focused, in recent years, on services robots and how to endow them with human emotions

capability (Perez-Gaspar, Caballero-Morales, and Trujillo-Romero 2016). Endowing robots with emotion recognition allows the development of more intuitive, natural and understandable communication between humans and robots. These communications may take several forms. The first form concerns collaborative HRI involving persons and robots that are working together to complete a common task (Scheutz, Schermerhorn, and Kramer 2006). The second form concerns the assistive HRI where robots are designed to assist person (Conn et al. 2008). Most of these studies explore pattern recognition techniques where these techniques are deeply linked with their training data and allow only the recognition of observed emotions. Some research works on automated affect/emotion recognition have investigated approaches based on unimodal techniques where only single source is used such as facial expressions (Cid et al. 2013), voice (Jacob 2016), textual expressions (Kalchbrenner, Grefenstette, and Blunsom 2014), body language (McColl and Nejat 2014), etc. Other research works have focused on multimodal techniques for affect/emotion recognition using visual and spoken information (Hu et al. 2015). The use of multimodal sources over a single source allows, on the one hand, to estimate affective/emotional state using the remaining modalities when one modality is not available, and on the other hand, to provide increased robustness and performance in terms of emotion recognition due to the complementarity and diversity of information when multiple modalities are available.

Different ontologies were proposed in the literature, as an alternative to data-driven approaches, with the aim of modeling emotion and affect related issues. In text analysis area, a semantic lexicon of feelings is presented in (Cambria, Olsher, and Rajagopal 2014). Recently, an ontology, called Emotion Ontology, which covers all aspects of emotion, affect and mental states from the neuroscience point of view, is proposed in (Hastings et al. 2012). More recently, EmotionsOnto, the upper ontology under development, is proposed in (Gil et al. 2015) for describing emotions and their recognition systems. To our knowledge, no emotion ontology-based approach for robotic systems was proposed in the literature. Developing cognitive robots with the capability of managing natural interactions with humans according to their emotional contexts needs that robots and all entities populating the intelligent ambient environment share semantically the same knowledge. Indeed, an ontology should cover all commonsense concepts of humans states and entities populating the environment that existing ontologies lack and thus can no longer be reused in robotic systems. Furthermore, using commonsense knowledge allows the ontology to be extended automatically through other existing commonsense ontologies.

Cognitive architecture for emotion-aware robotic system

Since the cognitive architecture combines information from different sources such as robots, devices, everyday objects, and humans with several inference techniques, integrating them into one uniform system is an important issue. In particular, it consists of the integration at the representation level to manage the rich semantics of natural interactions with humans and to ensure that all entities populating the intelligent ambient environment share a commonsense knowledge and common language. These latter guarantee the semantic interoperability of the entities and enabling them to communicate between each other. A robot cognitive architecture was introduced in previous work (Ayari et al. 2015) with an emphasis on systems aspects according to multi-agents and service-oriented computing paradigms, such as, the commonsense knowledge representation model, reasoning in open world techniques, and communication capabilities. In this paper, a focus is addressed to the perception and the recognition of human's emotional context enabling robots to better decision making. An overview of the proposed architecture is shown in Fig. 1.

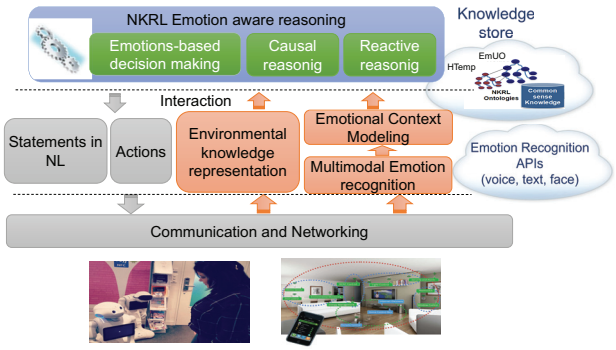

Figure 1: Cognitive architecture for emotion-aware robotic system

The *communication and networking service* enables the entities populating the environment to connect and subscribe to the cloud services using the standardized middleware technologies (XMPP, REST, etc.). The communication service enables also the basic exchange capability between any entity by more focusing on the encoding of messages' content, which is defined by elements such as lexicon, grammar, speech acts, and semantics.

The *knowledge base* connects most of the services such as the knowledge representation services and the reasoning services, enabling cognitive agents or robots to be endowed with large general purpose commonsense knowledge for human-centered applications. The ontology representing commonsense knowledge relies on a central server and is shared between all the entities populating the environment when each entity has its own instantiations of dynamic knowledge.

The *environmental knowledge representation service* provides a set of common techniques, algorithms and technologies introduced in previous work (Ayari, Chibani, and Amirat 2013). It allows robots to understand statements by converting these latter into NKRL predicates occurrences. The knowledge representation services exploit the expressiveness of the n-ary ontologies on which the narrative knowledge representation language (NKRL) is based. It allows to overcome the problems encountered in the dynamic knowl-

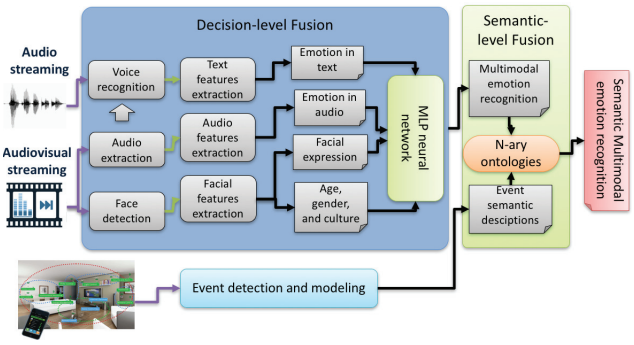

Figure 2: Semantic Multimodal emotion recognition system

edge representation approaches based on binary ontologies (Tenorth and Beetz 2015).

The *multimodal emotion recognition service*, proposed in this paper, ensures the human-emotion recognition from facial expressions, voice, and statements. By exploiting machine learning techniques, this service endows robots with the capability to fusion speech and face modalities for an efficient emotion recognition, cf. next section. The *emotional context modeling service*, proposed in this paper, allows the symbolic description of human emotions recognized by the emotion recognition service based on machine learning techniques and to associate them with their context for more accurate recognition. The expressiveness of the n-ary ontologies provided by the NKRL language is exploited for a richer emotional context description. The generated NKRL predicates occurrences within these services are stored in the knowledge base, and queried back, when necessary by reasoning techniques.

The *reasoning services* are based on the inference engine of Narrative Knowledge Representation Language (NKRL). This engine was extended, in previous work, by context-aware reasoning models including human preferences (Ayari et al. 2015), reasoning model based on collective intelligence (Ayari et al. 2016). In this paper, the inference engine is extended by adding new transformation and hypothesis rules for non-observable emotional context recognition, and emotion-based decision making model.

The main design principle of the proposed architecture is to integrate seamlessly the entities populating an ambient environment and to endow the decisional components of the robot with models of human emotions and human preferences in order to develop an effective cognition for a robot that is able to serve and interact seamlessly with humans.

Semantic multimodal emotion recognition

To handle the complexity of the daily living environments and to more accurate emotion recognition, an online multimodal emotion recognition model is proposed in this work. This model is based on hybrid fusion techniques that combine information of multiple modalities at different levels. In this paper, hybrid fusion consists of combining, at the low level, features with decisions fusion techniques based on MLP neural network. At the high level, the result of these techniques is exploited by the proposed semantic representation and reasoning models based on NKRL. The architecture of the proposed multimodal emotion recognition system is shown in Fig. 2.

Multimodal emotion recognition

Features extraction: Humans eye gaze, brows, lips, and face muscles motions and positions can be used as good features for recognizing the facial expression. Focusing on words used, their syntactic structure, their meaning, and the manner with which they are produced via the intensity and quality of the voice appears as a good features vector for recognizing emotions on speech. As emotion is defined as immediate reaction following an event, it is important to take into account what is happen around the human. In this work, speech and audio-visual modalities are taken into account for the expressiveness of the information that they contain. These information are relevant to the multimodal emotion recognition at the low level. Many features influence the emotional meaning. In this work, three features are extracted from audio-visual data such as age (Murry and Isaacowitz 2016), gender (Hess, Adams Jr, and Kleck 2004), and culture (Jack et al. 2009) that are considered as a vector of "emotion analysis" features. These latter improve the emotion recognition.

Hybrid fusion model based on Multilayer Perceptron (MLP) neural network: To estimate emotion probability from the combination of the multimodal prediction of the classifiers such as textual classifier, audio classifier and face classifier with the features age, gender and culture, Multilayer Perceptron (MLP) neural network is exploited. The Multilayer Perceptron (MLP) neural network consists of several layers of neurons which are completely connected. Formally, the layer is denoted by l_i, where $l_i = \{n_k^i\}; i \in 1..L$, i the number of the layers, L the last layer, and k the number of the neurons n of the layer's l number i. Each neuron n_k^i, except the input layer neurons n_k^1, has an input, denoted by x_k^i and an output, denoted by y_k^i. Three categories of these layers are distinguished as follows:

1. The input Layer: is the first layer denoted by $l_1 = \{n_k^1\}$ where, n_k^0 are the neurons representing the prediction of each classifier, such as emotion in text prediction, emotion in audio prediction, facial expression prediction, age, culture, and gender;
2. The hidden layers are the remaining layers denoted by $l_i = \{n_k^i\}; i \in 2..L-1$ where i denotes the number of the hidden layers and n_k^i the neurons calculated from previous layers. In this work,
3. The output layer: is the last layer denoted by $l_L = \{n_k^L\}; k \in 1..9$ where n_k^i, the neurons representing the emotions classes, are considered as output. In this paper, nine (9) classes of observed emotions are distinguished: *anger, contempt, fear, happiness, neutral, sadness surprise, disgust, and energy*.

In this work, the MLP neural network algorithm consists of:

- Three (3) hidden layers where each hidden layer has 100 nodes. The Dropout technique is exploited on 10% of the hidden nodes to prevent the model from over-fitting on the training data;
- The weight's initialization process that is based on *small normal distribution* with zero mean and 0.05 standard deviation;
- The neuron's output/input computing process where the forward propagation algorithm is exploited to compute the neuron's output/input with a nonlinear function as an activation function, denoted by $f(x)$. The Rectified Linear Units (ReLU) and Softmax functions are exploited;
- Errors-computing and weights-updating process where the back propagation algorithm (Hagan and Menhaj 1994) is exploited. The categorical cross entropy function is exploited as a cost function and it has been optimized by Adaptive Moment Estimation (Adam) (Kingma and Ba 2014).

Emotional context management based on n-ary ontologies

This section introduces the representation and associated inference methods for emotions and other human states information.

Emotional knowledge Representation: To describe emotions and affects in intelligent ambient environment, the ontological model based on the Narrative Knowledge Representation Language (NKRL) is exploited. The static characteristics concern here the commonsense knowledge, such as feeling, passion and dislike. The dynamic characteristics are used to describe what can be observed or inferred in a specific context, such as, "having a positive/negative experience". Static characteristics and dynamic characteristics are modeled, respectively, by using the $HClass$ ontology and the $HTemp$ ontology of NKRL language.

- **Toward emotion upper-ontology (EmUO):** The proposed ontology, extension of the *HClass ontology*, aims to represent human attributes, in particular, human states including affect, emotions, moods, etc. Unlike *emotion ontology (EMO)*, which covers all aspects of emotion, affect and mental states of humans from the neuroscience point of view (Hastings et al. 2012), the proposed *emotion upper ontology (EmUO)*, aims to cover all commonsense concepts of human states and robotic systems. Based on the automatic commonsense knowledge extension model introduced in (Ayari et al. 2015), the *(EmU) ontology* combines different sentiment and emotion lexica for their expressiveness such as *SenticNet3*[1] and *WordNet-Affect*[2] to infer the commonsense concepts of human states, cf. figure 3.
- **Emotional context modeling:** The $HTemp$ ontology is exploited to represent complex context using its n-ary schema. This latter is conceived as the formal representation of generic classes of elementary events such as "having a positive/negative experience", "threatening someone with violence", and "evaluating an artefact", etc. The NKRL templates are able to represent the full emotional context of human where elements like culture, religion, and other social rules should be taken into account.

[1]http://sentic.net/

[2]http://wndomains.fbk.eu/wnaffect.html

E.occ2: **EXPERIENCE**
SUBJ OLIVIER : SPECIF(ADP_building SHOWROOM_1)
OBJ SPECIF (feeling_ surprise_)
TOPIC SPECIF (robot_activity interacting_)
date-1: 14/02/2017 14:14:00
date-2:
Experience:Human/Social

Table 1: Representation of the emotional context: Olivier is surprised by the higher cognitive level of the robot

Let us consider the example where a companion robot, called *Pepper*, detects that a visitor, called *Olivier*, is surprised by its interaction capability. The semantic description of this emotional context is given in table 1 by the symbolic label of the predicative occurrence ($L =$ "$E.occ2$"). In this case, the conceptual predicate ($P =$ "$EXPERIENCE$") indicates the experience of the entity specifically a human. This human entity is represented with the role ($R =$ "$SUBJ$") and its argument ($a =$ "$OLIVIER$"), an instance of the $HClass$ concept "$human\_being$". The emotion is represented by the role "OBJ" and its argument *SPECIF(feeling_ surprise_)*. The question *"what is the meaning of someone is surprised?"* could be answered that is about *a good/positive feeling* or *negative feeling*. Due to the context of the user, a positive surprise can be differentiated from a negative one. The novelty of this work is the description of emotional context with n-ary ontologies that makes possible the recognition of non-observable emotions such *curiosity* and *attention*.

NKRL emotion-aware reasoning: The inference engine of NKRL language based on the "transformation and the hypothesis rules allows the inference of implicit relations between predicative occurrences and consequently the chronological/semantic emotional context when an event occurs. This inference engine helps in better understanding of human emotions in a given period. The NKRL inference rules can be conceived as implications that can be expressed as formula 1. The constraint $var_i \subseteq var_j$ corresponds to the variables defined in the antecedent that must also appear in the relative consequent according to the commonsense knowledge.

$$A(var_i) \Rightarrow Cs_i(var_j); var_i \subseteq var_j. \quad (1)$$

By exploiting the richness of the n-ary ontologies-based emotional context representation and the high-level inference mechanisms of the NKRL, new complex emotional

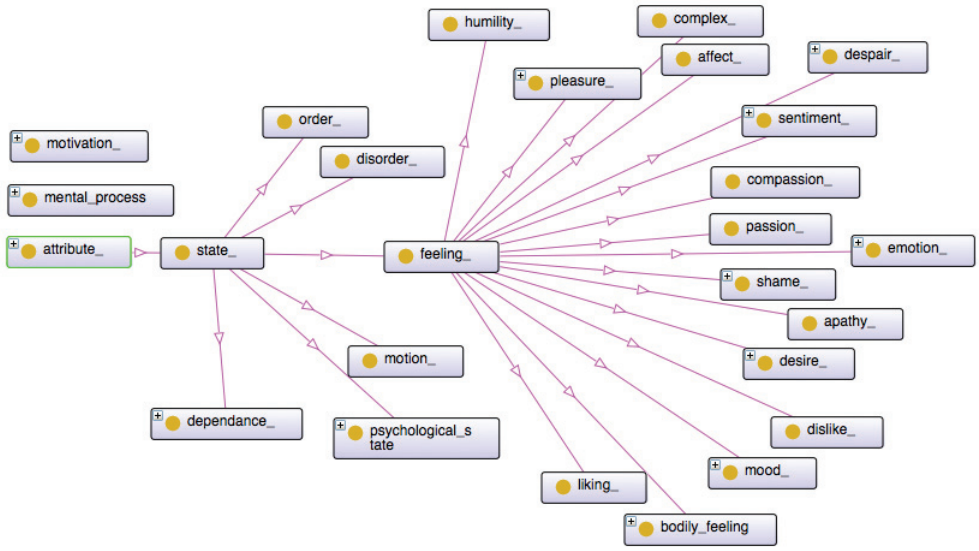

Figure 3: A segment of the *emotion upper ontology (EmUO) describing human state taxonomy*

contexts can be recognized allowing a better decision making in terms of assistance services.

The inference engine of the NKRL is used here to infer the feeling of the human following the occurrence of an event e at the instant t; the transformation rules $\mathcal{TR}_j$ of NKRL are used. Formally, an event e can make true a situation (i.e feeling) f starting from an instant t such as: $(f,t) \Leftarrow (e,t) \wedge \mathcal{RT}_j$

The table 2 shows an example of transformation rule allowing to infer that a human feels curious about something. The feeling of curiosity is a complex context that couldn't be observed. It needs the aggregation of different situations, attention to an activity, process information and persist on tasks that are represented respectively by the NKRL templates *Behave:Focus*, *Behave:Questioner* and *Behave:Participant*. The operator *COORD* appears in the antecedent of the transformation rule, to aggregate the three aspects of curiosity such as attention, process information, and persist on tasks, see table 2. The recognition of some situations such as the recognition of the attention to something is difficult. For example, the attention of a person to a smart device can be detected by looking for a long period this device.

Emotion-based decision making: The main objective of the introduced reasoning techniques is to create emotion-aware services for intelligent assistance. To model the emotional context of an AAL (Ambient Assisted Living) application, the different elements that affect the application and infer the context have to be identified. In this paper, the states of all humans populating the ambient environment are used for modeling the context. The *user feeling* is one of the contextual attributes. As the humans act in the dynamic environments, a description of the continuously updated environment is required. The reasoning core of the robot supports Spatio-temporal reasoning about the changing emotions of humans. In order to reason about which emotions are needed for performing an action, the robot *Pepper* has a semantic description of the tasks regarding the emotional context of the user. In this study, a task selection model is able to check which emotions can be acquired to achieve a task. Formally, a task selection consists of a tuple $\Pi = \langle K, E, T \rangle$ where K is a set of tasks, and E is a set of emotions evolving over the time T. Selecting a task k according to a particular emotion e_t at specific time t consists of determining the normalized score N_k:

$$N_k = \frac{K^s}{\sum {K^s}_i} \tag{2}$$

The task score K^s is a factor which is evolving over time based on various parameters: the prerequisites tasks score p_k of the next planned task, the current and previous emotion score respectively e_t and e_{t-1}, and the emotion transition factor ϕ_e where $\phi_e \in [-1, 1]$.

$$K^s = \frac{p_k}{e_t} + \phi_e \tag{3}$$

$$\phi_e = e_t - e_{t-1} \tag{4}$$

Evaluation

To evaluate the proposed approach for multimodal emotion recognition, several experiments were conducted. Some experiments have been addressed to evaluate the multimodal emotion recognition based on hybrid fusion techniques. Results are reported for validation and test sets of a 5-*fold cross-validation scheme*.

Dataset: multimodal emotion

A dataset is proposed in (Morency, Mihalcea, and Doshi 2011) to build a multimodal emotion sensing model based

$Antecedent$:
COORD(e_1 e_2 e_3)
e_1: BEHAVE SUBJ var1 :
OBJ conversation_
date-1
date-2
var1:human_being
Behave: Participant
e_2 : BEHAVE SUBJ var1 :
MODAL attention_
CONTEXT var2
date-1
date-2
var1:entity_
var1:human_being
Behave:Focus
e_3: BEHAVE SUBJ var1 :
MODAL SPECIF(quering_ information_)
CONTEXT var2
date-1
date-2
var1:entity_
var1:human_being
Behave: Questioner

$Consequent$:
f_1: Experience SUBJ var1:
OBJ SPECIF (feeling_ curious_)
TOPIC var2
date-1 t
date-2
var1:entity_
var1:human_being
Experience:PosotiveHuman/Social

Table 2: Example of transformation rules $\mathcal{RT}$

on visual, textual, audio and culture features. It contains 47 videos collected from Youtube which are addressing more than one topic like electronics products, politics, movies, and foods. The videos were found using the following keywords: opinion, product review, toothpaste, war, job, business, cosmetics review, baby product review, etc. (Morency, Mihalcea, and Doshi 2011). 27 males and 20 females with age varying from 14 years to 60 years were involved in the dataset. All subjects are speaking in English although they belong to different ethnic backgrounds like Asian, African, African-American and Caucasian. To avoid the issues of introductory titles and multiple topics, all videos are preprocessed as follows: (i) remove the first 600 frames from each video, (ii) divide each video into multiple segments, and (iii) label each segment with the corresponding emotion category.

To avoid that the model might end up learning the dependencies between consecutive segments and to guarantee fair evaluation, the whole divided dataset was shuffled randomly before splitting it into training and testing set.

Multimodal emotion recognition

A set of services for extracting features from audio-visual data are exploited. *Emotion API* of Microsoft cognitive service[3], *Alchemy API*[4] and *Vokaturi API*[5] are used to extract respectively facial, textual and vocal expressions. All of the services have been applied on the used dataset.

The MLP neural network is implemented based on Keras framework with tensor-flow as a backend. The experimental results on the used dataset are compared with results of the approaches proposed by (Poria, Cambria, and Gelbukh 2015) and (Pérez-Rosas, Mihalcea, and Morency 2013) using the same dataset.

Table 3 presents the performance of the multimodal emotion recognition using classification techniques by including the selected features such as age, gender and culture. It consists of measuring the mean accuracy of the $5 - folds$ cross validation. The reported results show that including the features improve significantly the recognition performance of emotions in the different techniques such as unimodal, bimodal, and multimodal. For example, the mean accuracy is enhanced, in the audio modality from 84.75% to 92.7% where the mean accuracy of the face modality is enhanced from 80.46% to 86.35%.

The proposed hybrid fusion model based on MLP neural network is able to recognize the basic emotions such as happy, sad, anger and neutral. By including the selected features, the performance of this model is enhanced in terms of mean accuracy from 81.65% to 85.45%. Comparing the modalities' performances, the audio modality has the highest accuracies, in terms of both average and best accuracy, that are respectively 92.7% and 98.25%. In contrast, during real-world interactions, noise can affect the audio modality's performance whose justifies the choice of multimodal approach.

Since the type of accuracy has not been identified in the benchmark if it was the best or the average accuracy, both of these latter are calculated in this work. By comparing the results of the state of the art fusion models obtained, in terms of the average accuracy, the proposed approachs multimodal performance is 85.45% that is not as good as the benchmark (88.6%). This shortcoming is due to the evaluation method where the proposed multimodal approach's performance is calculated over 5 folds cross validation unlike the benchmark that a traditional training/testing evaluation was applied by dividing the dataset into 70% for training and 30% for testing. The benchmark's evaluation method has an advantage to get better results since the training set is more than testing set. In terms of the best accuracy, the proposed approachs multimodal performance is 91.29%.

During the 5 folds validation set, the average processing time to recognize emotions for each fold was $0.35s$ which is well compatible with the dynamic nature of the ambient environment for real time recognition.

[3]https://azure.microsoft.com/fr-fr/services/cognitive-services/emotion/

[4]www.alchemyapi.com/

[5]https://vokaturi.com/

Techniques		(Pérez-Rosas, Mihalcea, and Morency 2013)	(Poria, Cambria, and Gelbukh 2015)	The proposed approach		
				Mean Accuracy		Best
				without features	with features	accuracy
Unimodal	T	70.94%	79.77%	73.24% (+/- 4.4%)	**77.19** % (+/- 2.05%)	80.70%
	F	67.31%	76.38%	80.46% (+/- 4.42%)	**86.35**%(+/- 4.68%)	92.98%
	A	64.85%	74.22%	84.75% (+/- 2.02%)	**92.7**%(+/- 3.46%)	98.25%
Bimodal	T+F	72.39%	85.46%	81.47% (+/- 3.7%)	**83.7**%(+/- 2.6%)	88.60%
	T+A	72.88%	84.12%	74.34% (+/- 3.25%)	**81.96**%(+/- 3.51%)	87.72%
	F+A	68.86%	83.69%	82.37% (+/- 4.6%)	**86.5**%(+/- 2.3%)	89.47%
Multimodal		74.09%	88.60%	81.65% (+/- 6.7%)	**85.45**% (+/- 2.9%)	**91.23**%

Table 3: Accuracy of state-of-the-art method compared of the proposed model with Youtube dataset. (T:text, F:face, A:Audio)

Figure 4: Scene extracted from the smart devices showroom

Non-observable emotion recognition

To validate the proposed approach for non-observable emotion recognition, a use case of a tour guide is studied. In this use case, Oliver is a guest who needs help from a robot called Pepper that will act as a tour guide to explain and show *Oliver* the equipment in the smart devices showroom, cf. figure 4. Pepper can detect and monitor Oliver's emotions continuously based on the proposed approach. In particular, it changes the discussion based on the detected emotions. During this use case, two complex emotions are recognized using the proposed approach, such as upset and confusion. For instance, the emotion "confusion" is recognized by analyzing the meaning of a query of the user during a discussion using NKRL language. The native purpose of this language aims to represent and to reason on the meaning of natural language sentences which allows the recognition of complex emotions. These emotions cannot be recognized by other approaches, such as data-driven approaches and logic-based approaches. Therefore, data-driven emotion recognition is not sufficient where ontology-based reasoning seems a good complementary.

The proposed use case shows how the proposed approach makes the Human-Robot interaction in a complex environment more naturally and more intelligent by recognizing non-observable emotions.

Conclusion

In this paper, hybrid approach based on MLP neural network and n-ary ontologies for emotion-aware robotic systems is proposed to enhance human-robot interaction in ambient intelligent environments. Its principle consists of, on the one hand, an efficient multimodal emotion recognition based on MLP neural network, and on the other hand, the expressive emotional knowledge representation and reasoning model based on NKRL language.

The performance of the multimodal emotion recognition based on MLP neural network model was enhanced in terms of accuracy by including the selected features: age, gender and culture. The proposed semantic multimodal recognition approach is able to recognize the non-observable emotions by associating the recognized emotions at the low level with their context.

The ongoing works address the extension of the proposed approach for managing the uncertainty of the multimodal emotion recognition in order to improve the recognition and to reduce the variance of the accuracy.

References

Ayari, N.; Chibani, A.; Amirat, Y.; and Matson, E. T. 2015. A novel approach based on commonsense knowledge representation and reasoning in open world for intelligent ambient assisted living services. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 6007–6013.

Ayari, N.; Chibani, A.; Amirat, Y.; and Matson, E. 2016. A semantic approach for enhancing assistive services in ubiquitous robotics. *Robotics and Autonomous Systems* 75:17–27.

Ayari, N.; Chibani, A.; and Amirat, Y. 2013. Semantic management of human-robot interaction in ambient intelligence environments using n-ary ontologies. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 1172–1179. IEEE.

Cambria, E.; Olsher, D.; and Rajagopal, D. 2014. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, 1515–1521. AAAI Press.

Chibani, A.; Amirat, Y.; Mohammed, S.; Matson, E.; Hagita, N.; and Barreto, M. 2013. Ubiquitous robotics: Recent challenges and future trends. *Robotics and Autonomous Systems* 61(11):1162 – 1172.

Cid, F.; Prado, J. A.; Bustos, P.; and Nunez, P. 2013. A real time and robust facial expression recognition and imitation approach for affective human-robot interaction using gabor filtering. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, 2188–2193. IEEE.

Conn, K.; Liu, C.; Sarkar, N.; Stone, W.; and Warren, Z. 2008. Affect-sensitive assistive intervention technologies for children with autism: An individual-specific approach. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 442–447.

Darwin, C.; Ekman, P.; and Prodger, P. 1998. *"The expression of the emotions in man and animals"*. Oxford University Press, USA.

Dey, A. K., and Abowd, G. D. 2000. Towards a better understanding of context and context-awareness. In *Workshop on The What, Who, Where, When, and How of Context-Awareness*.

Gil, R.; Virgili-Gom, J.; Garca, R.; and Mason, C. 2015. Emotions ontology for collaborative modelling and learning of emotional responses. *Computers in Human Behavior* 51, Part B:610 – 617.

Gorlova, N. A., and Gultyaeva, T. A. 2016. Emotion classification on image using hidden markov models. In *2016 13th International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, volume 03, 1–1.

Hagan, M. T., and Menhaj, M. B. 1994. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks* 5(6):989–993.

Hastings, J.; Ceusters, W.; Mulligan, K.; and Smith, B. 2012. Annotating affective neuroscience data with the emotion ontology. In *Third International Conference on Biomedical Ontology*. ICBO. 1–5.

Hess, U.; Adams Jr, R. B.; and Kleck, R. E. 2004. Facial appearance, gender, and emotion expression. *Emotion* 4(4):378.

Hu, Y.; Ren, J. S. J.; Dai, J.; Yuan, C.; Xu, L.; and Wang, W. 2015. Deep multimodal speaker naming. *CoRR* abs/1507.04831.

Jack, R. E.; Blais, C.; Scheepers, C.; Schyns, P. G.; and Caldara, R. 2009. Cultural confusions show that facial expressions are not universal. *Current Biology* 19(18):1543–1548.

Jacko, J. A. 2012. *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications, Third Editiion*. Boca Raton, FL: CRC Press, 3 edition.

Jacob, A. 2016. Speech emotion recognition based on minimal voice quality features. *2016 International Conference on Communication and Signal Processing (ICCSP)*.

Kalchbrenner, N.; Grefenstette, E.; and Blunsom, P. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 655–665. Baltimore, Maryland, USA: Association for Computational Linguistics.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Makioka, T.; Kuriyaki, Y.; Uchimura, K.; and Satonaka, T. 2016. Quantitative study of facial expression asymmetry using objective measure based on neural networks. In *2016 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 1–4.

McColl, D., and Nejat, G. 2014. Determining the affective body language of older adults during socially assistive hri. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, 2633–2638. IEEE.

Morency, L.-P.; Mihalcea, R.; and Doshi, P. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, 169–176. New York, NY, USA: ACM.

Murry, M. W., and Isaacowitz, D. M. 2016. Age differences in emotion perception the effects of the social environment. *International Journal of Behavioral Development* 0165025416667493.

Nicolai, A., and Choi, A. 2015. Facial emotion recognition using fuzzy systems. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2216–2221.

Perez-Gaspar, L.-A.; Caballero-Morales, S.-O.; and Trujillo-Romero, F. 2016. Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Systems with Applications* 66:42–61.

Pérez-Rosas, V.; Mihalcea, R.; and Morency, L.-P. 2013. Utterance-level multimodal sentiment analysis. In *ACL (1)*, 973–982.

Poria, S.; Cambria, E.; and Gelbukh, A. F. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, 2539–2544.

Samani, H. A., and Saadatian, E. 2012. A multidisciplinary artificial intelligence model of an affective robot. *International Journal of Advanced Robotic Systems* 9(1):6.

Scheutz, M.; Schermerhorn, P.; and Kramer, J. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, 226–233. New York, NY, USA: ACM.

Tenorth, M., and Beetz, M. 2015. Representations for robot knowledge in the knowrob framework. *Artificial Intelligence*.

Zarri, G. 2009. *"Representation and Management of Narrative Information: Theoretical Principles and Implementation"*. Springer Verlag. Advanced Information and Knowledge Processing series.