Identification of Multimodal Signals for Emotion Recognition in the Context of Human-Robot Interaction

Andrea K. Pérez[1], Carlos A. Quintero[1](✉), Saith Rodríguez[1], Eyberth Rojas[1], Oswaldo Peña[1], and Fernando De La Rosa[2]

[1] Universidad Santo Tomás, Bogotá D.C., Colombia
{andrea.perez,carlosquinterop,saithrodriguez, eyberthrojas,oswaldopena}@usantotomas.edu.co
[2] Universidad de los Andes, Bogotá D.C., Colombia
fde@uniandes.edu.co

Abstract. This paper presents a proposal for the identification of multimodal signals for recognizing 4 human emotions in the context of human-robot interaction, specifically, the following emotions: happiness, anger, surprise and neutrality. We propose to implement a multiclass classifier that is based on two unimodal classifiers: one to process the input data from a video signal and another one that uses audio. On one hand, for detecting the human emotions using video data we have propose a multiclass image classifier based on a convolutional neural network that achieved 86.4% of generalization accuracy for individual frames and 100% when used to detect emotions in a video stream. On the other hand, for the emotion detection using audio data we have proposed a multiclass classifier based on several one-class classifiers, one for each emotion, achieving a generalization accuracy of 69.7%. The complete system shows a generalization error of 0% and is tested with several real users in an sales-robot application.

1 Introduction

In the last decades, multiple research groups have focused in developing the technology for constructing humanoid robots that interact actively in different aspects of daily life activities, which would generate a high social and industrial impact. One renowned initiative is RoboCup [1], which was established at the end of the last decade and where universities from 45 different countries participate every year. Such initiative has different research areas, one of them is human-robot interaction. Robocup, together with other well-known initiatives such as [2–4], show the relevance given among the academia and scientific community to achieve robots that can cooperate with humans in daily-life activities, in a medium-term period. However, such challenge implies the interaction of multiple disciplines and integration among all the different technical advances. Then, the common goal is to develop intelligent systems robust enough so that

C. Brito-Loeza and A. Espinosa-Romero (Eds.): ISICS 2018, CCIS 820, pp. 67–80, 2018.
https://doi.org/10.1007/978-3-319-76261-6_6

anyone can interact with robots in a natural manner, i.e., the robot should be able to understand different types of human communication.

Therefore, this work focuses in the problem of human emotion detection by a robot. Specifically, in the context of social interaction such as a sales agent robot promoting a product. In such environment, the robot starts the conversation presenting the product and consulting the interest of the potential client (user) through an emotion recognition system that is executed remotely while the person answers to a series of questions. Depending on the detected emotional state of the user a command is sent to the robot via Ethernet, which alters the responses according to the user's emotional response.

The emotion recognition system identifies four emotions: happiness, anger, surprise and neutral; which represent the possible emotional state of a user in front of a sales agent. The proposed system implements a multimodal classifier composed by the combination of two unimodal classifiers, one per type of input data: facial expressions and voice.

The rest of the article is organized as follows. The next section presents a summary of works related to emotion recognition systems. Section 4 contextualizes the proposed methodology under an application case of study. In Sect. 4 the proposed system's overall architecture in the human-robot interaction context is explained in detail. Section 5 details the selection process for the architecture and unimodal classifier parameters. In Sect. 6 the obtained results with the audio and video classifier are shown and analyzed; subsequently, the results obtained of the multimodal classifier are shown. Finally, in Sect. 7 we present the conclusions of the preceding results.

2 Related Work

In recent years, human-robot interaction (HRI) has been highly studied among researchers in artificial intelligence and computer science. In [5], the authors present a recompilation of the main problems to solve in HRI, the fundamental problems and a discussion regarding the short-term challenges that need to be addressed. This article first describes the history context of human-robot interaction, beginning in late 90's and early 2000's where multiple disciplines such as robotics and psychology noticed the importance to work together. Subsequently, the HRI problem is described as the interaction between one or more humans with one or more robots; understanding the interaction as an intrinsic part of robotics, mainly because robots are built to assist humans in their tasks.

Additionally, Goodrich [5] defines five features to consider while designing solutions in HRI context: autonomy level, type of information exchange, equipment structure, adaptability, learning and training for humans and robots, and the task. Finally, a series of challenges are described for HRI, one of them is assistive robotics, which seeks to provide physical, mental, or social support to persons in such need. This challenge includes special attention in features intrinsic in human communications, such as speech recognition, multimodal interaction and cognitive analysis.

On the other hand, the researching relevance of HRI is reflected in the @Home league of RoboCup world initiative [6]. The main aim of such league is to develop technology cooperatively among the participant universities to create a robot that is able to do the domestic chores and have a natural interaction with humans. To achieve such objective, research needs to focus in navigation and mapping in dynamic environments, artificial vision and object recognition in natural lighting environments, object manipulation, and human-robot interaction and cooperation.

One approximation to the HRI problem is from the machine learning perspective. From the aforementioned works, it can be concluded that it is paramount to develop intelligent systems that allow a natural interaction between the human and the robot. According to this, [7] presents a framework for learning desirable tasks in HRI, based on the illustration of the task to the robot and its decomposition in sequences of spatial points. Such approach is called KLfD and the proposed model consists of an organized set of groups of spatial points. Another approach is presented in [8], which focuses on the recognition of human hand gestures in the HRI context. The presented methodology consists of combining an algorithm to recognize hand gestures with two classifiers, one built for hand skeleton recognition (HSR) and the other is based on a SVM.

As previously mentioned, in HRI different approaches are considered in which several challenges are tackled, this is exemplified in the aforementioned articles. Furthermore, in [9] an emotion multimodal recognition system is proposed for human-robot interaction. This work is developed under a RDS (Robotics Dialog System) and implements two modules for emotion recognitions, one for the voice (GEVA) and one for the facial expression (GEGA). These modules are based on a fast recognition algorithm for objects and a toolbox for expressions recognitions (CERT).

In this area, other approaches are found such as the one in [10], in which Machine Learning techniques are used for recognition of integrated audio-video signals, where supervised learning is usually used. However, recently, recommendation systems are being used in robotics, specifically in human-robot interaction.

Several works have been developed either in facial gestures recognition or emotion recognition. However, in [11] an integrated approach is presented, where facial and emotion recognition is implemented simultaneously. This approach, used Soft computing, specifically Fuzzy rule based system (FBS), to reduce road traffic accidents due to driver somnolence. In order to do so, this work proposes to include the facial gesture and recognized emotion of the driver, so that if any fatigue signal is detected in the driver, the vehicle switches to autopilot mode. The solution architecture consists of a camera for image acquisition, face recognition analysing images in RGB, then feature extraction for eye movement and lips. This is the input for the emotion and gesture recognition systems. The former systems where developed using Fuzzy inference, which allows to distinguish between 4 emotional states (happy, angry, sad and surprised). As a result, a precision of 91.66% is obtained for facial gestures recognition and 94.85% for the combination of facial gestures and emotion recognition.

3 Social Robot: Application Case Study

Figure 1 shows the general scheme for our case study to validate the proposed emotion recognition system. The scenario consists of a social robot that acts as a sales agent. The idea is that the robot will attempt to sale a specific product to the human and during their interaction, it will constantly detect the user's emotional state, using her facial gestures and voice, to adapt to the user's response. Similar to traditional *call centers*, the system will use the user emotions, detected through her answers during the interaction, to guide the sales process (i.e., to choose the verbal script to persuade the user to perform the purchase).

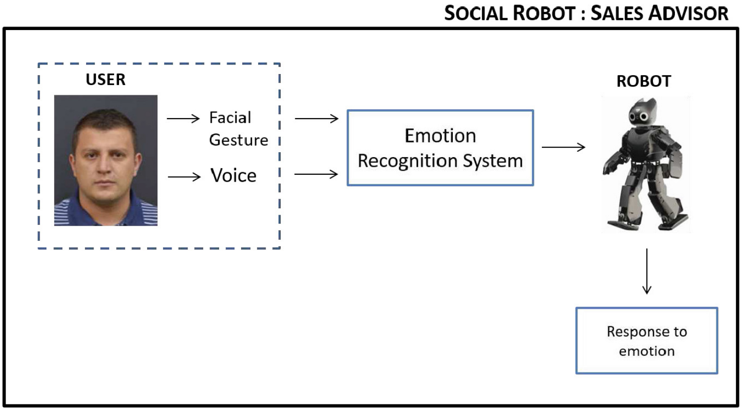


Fig. 1. General scheme of the proposed test case scenario for the detection of human emotions: robot sales advisor

4 Automatic Multimodal Emotion Recognition System Design

Our emotion recognition system is based on an automatic classifier that combines two different and independent classifiers, one for each input type, namely voice and facial gesture. The idea is that the classification system will detect the user's emotions through their interaction using data that comes from a video stream and also data collected with a microphone.

In Fig. 2 we show a diagram of the overall emotion recognition system. Initially, the input is the data acquired by the robot while it interacts with the human. This input is multimodal since it comes from different sources an hence there is one classifier per source. In the end, the outputs of both individual classifier is combined to identify one unique emotion. To build the emotion recognition model, we have performed experiments to collect data of different users interacting with the robot and such data have been split into two disjoint sets: the training set and the testing set. The former will be used to create the classification module while the latter will be used to evaluate the model's accuracy. All the modules of the overall architecture will be explained in detail in the following sections.

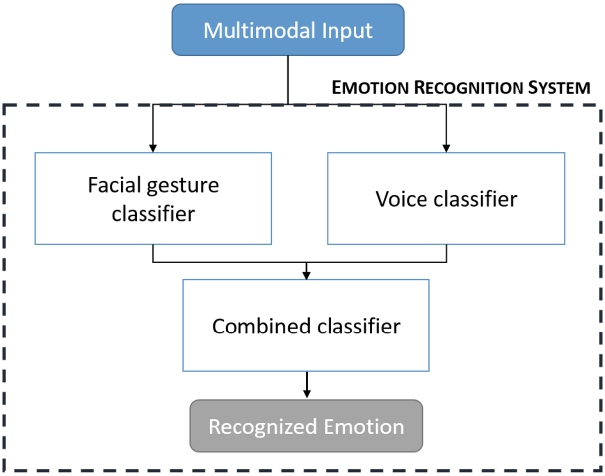


Fig. 2. Diagram of the proposed automatic emotion recognition system

4.1 Facial Gesture Classifier

We have selected a convolutional neural network (CNN) to implement a classifier capable of automatically recognize the user emotion based on a video data stream that captures the user facial gestures. In artificial vision systems particularly, deep learning and more specifically the CNNs have been successfully used, especially for their ability to automatically extract valuable features of input images for classification tasks [12,13].

Figure 3 illustrates the general scheme of the facial expression classifier. The video stream is considered as a set of independent image frames that are fed into the convolutional neural network, one at a time to be classified into one of the possible emotions.

Each frame is converted from its original RGB space to greyscale since the key features used to recognize the user's facial expressions are mostly related to local relationships such as borders and corners.

The CNN will output one detected emotion for each input frame of the video stream. All the detected emotions need to be merged together to define the whole video stream emotion using a voting procedure, which simply chooses the emotion decided by the majority. We define the video stream confidence q_{cv} found by the classifier as shown in Eq. (1), where FD is the number of frames labeled with the selected emotion and FV is the total number of frames of the video stream.

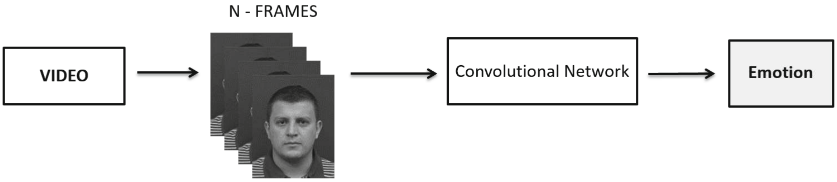


Fig. 3. General scheme of the facial expression classifier

$$q_{cv} = \frac{FD}{FV} \tag{1}$$

4.2 Voice Classifier

For the emotion recognition classifier that uses voice data we will use a similar approach to the one shown in [14] where One-class SVMs are used to discriminate one class each and then merged into one single output to create a multiclass classifier. Figure 4 shows the general voice-based classifier where a set of features need to be extracted from the audio source to feed the classifier. This feature extraction process, for the application at hand, is performed by computing the signal's Mel-frequency cepstral coefficients (MFCC): a set of coefficients that contain frequency relations of small windows that allows us to represent the important voice features from the user.

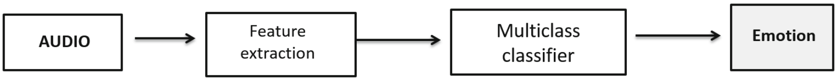


Fig. 4. Block diagram for the detection of emotions using audio data

The architecture for the multiclass classifier is shown in Fig. 5. An ensemble of one class classifiers process the features extracted from the original voice data and determine whether the given sample belongs to their class or not. Each SVM also outputs the classification margin for the current data sample, which provides an estimation of the classification confidence of the classifier for such data. In order to compare the classification margins output by each one-class classifier, we normalized the given margin $m(x)$ by the maximum margin obtained by each classifier during its training process $\hat{m}$. Therefore, the confidence value $q_{ca}(x)$ of each classifier when presented the input data x is as shown in Eq. (2).

$$q_{ca}(x) = \frac{m(x)}{\hat{m}} \tag{2}$$

The final output function for the multiclass classifier will select the class of the one-class SVM with highest confidence value.

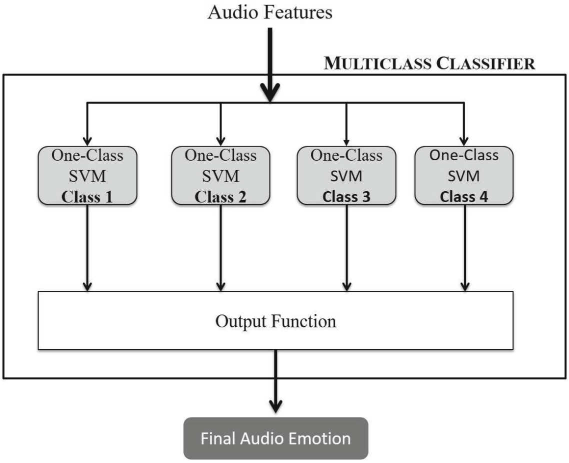


Fig. 5. General architecture of the emotion classifier using the voice commands

4.3 Combined Classifier

As described in previous sections, we have built a specific classifier for the video data and another for the voice data, each one, trained to automatically detect the possible user emotions during the interaction. Furthermore, each classifier also computes the classification confidence of a data sample and is characterized by the model's accuracy. Using this information we will assign the final label according to the following conditions:

- If both classifiers agree on the detected emotion, the general system's output will be the emotion detected by both classifiers. In this case, we calculate the agreement between both classifiers taken into account the confidence value for each classifier and their accuracies as shown in Eq. (3)

$$Agreement_{(cv,ca)} = \frac{Acc_{cv}q_{cv} + Acc_{ca}q_{ca}}{Acc_{cv} + Acc_{ca}} \tag{3}$$

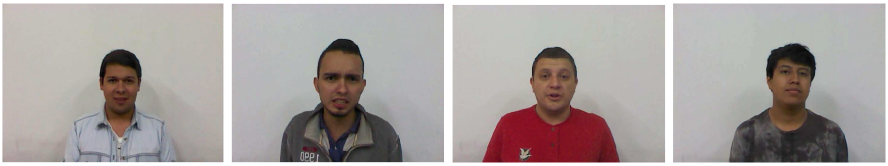
  where Acc_{cv} and Acc_{ca} are the classifier accuracies of the video and audio classifier respectively and q_{cv} and q_{ca} are the confidence values output by the video and audio classifiers respectively. The confidence values are assumed to be normalized between 0 and 1.
- If both classifiers output a different emotion, we must take into account the classification confidence of each classifier and also the model's accuracy. Therefore, the final emotion will be the one given by the classifier with highest weighted confidence value as shown in Eq. (4).

$$\max(Acc_{cv}q_{cv}, Acc_{ca}q_{ca}) \tag{4}$$

5 Experimental Selection and Tuning of Unimodal Classifiers

We have designed an experiment that allows us to test our proposed methodology. For this we have asked 11 users to perform 10 different repetitions, using voice and facial gestures, of the four emotions of interest. These signals are captured using a Kinect sensor finally obtaining a video with the facial gesture and an audio stream with the corresponding voice.

Using this method we have obtained a total of 110 videos and 110 audio samples. Each video is made of 640 × 480 pixel frames for a total of 30.994 images, for all emotions and users. Figure 6 shows one example of the collected image data for each emotion.



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

Fig. 6. Example of images captured per emotion: (a) Happiness (b) Anger (c) Surprise (d) Neutral

The collected data is split into training and testing set, leaving 77 video and audio data for training and 33 for testing. However, each video has a different number of frames. Table 1 shows the exact number of data reserved for training and testing in each emotion and for each classifier:

Table 1. Data distribution for training and testing per emotion

	Number of samples	Happiness	Anger	Surprise	Neutral
Training	Image	6465	5231	5040	5215
	Audio	77	77	77	77
Validation	Image	2415	2179	2244	2205
	Audio	33	33	33	33

5.1 Facial Expressions Classifier

The selection of the convolutional neural network architecture is achieved through an experimental approach by the variation of the number of receptive fields and their sizes. Initially, we have adopted the general network architecture shown in [15], where a similar problem is solved using CNNs, and implemented it using the Matlab toolbox **MatConvNet**, which contains many pre-trained CNNs for image classification and other tasks [16].

The first step consisted on shrinking the images size on the input layer. This decision was taken based on the fact that the captured images contained large background regions. The images were cropped and resized to 44 × 44 for the first two experiments and 43 × 43 for the last two scenarios. The general network contains 9 layers as shown in Fig. 7, whose area and number of receptive fields (RF) were modified for each experimental scenario (Test 1, Test 2, Test 3 and Test 4) as shown in Table 2.

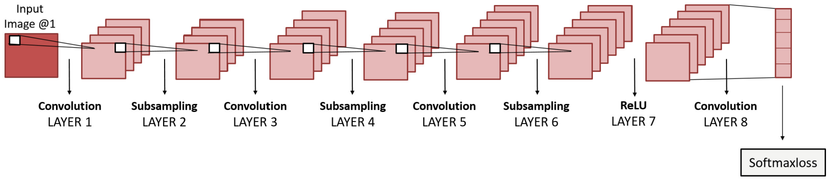


Fig. 7. Architecture of the convolutional neural network for emotion recognition in images

Table 2. Architecture parameters in the convolutional neural network for each experimental scenario

	Test 1		Test 2		Test 3		Test 4	
	Area RF	Number of RF	Area RF	Number of RF	Area RF	Number of RF	Area RF	Number of RF
Layer 1	2 × 2	50	3 × 3	50	4 × 4	50	5 × 5	50
Layer 2	2 × 2	50	2 × 2	50	2 × 2	50	2 × 2	50
Layer 3	2 × 2	100	2 × 2	100	3 × 3	100	3 × 3	100
Layer 4	2 × 2	100	2 × 2	100	2 × 2	100	2 × 2	100
Layer 5	5 × 5	200	3 × 3	200	2 × 2	300	2 × 2	200
Layer 6	2 × 2	200	2 × 2	200	2 × 2	300	2 × 2	200
Layer 8	3 × 3	4	4 × 4	4	4 × 4	4	4 × 4	4

The convolutional neural network for each experimental setup is trained and the training parameters and main results are shown in Table 3.

According to these results, the selected CNN architecture is the one of Test 3 since it attained the highest accuracy over the testing dataset $Acc_{cv} = 86.4\%$.

Table 3. Training parameters and classification accuracy for the different CNN architectures

CNN architecture	Training epochs	Training time	Accuracy (Acc_{cv})
Test 1	7000	49.6 h	60.26%
Test 2	10000	78.47 h	81.97%
Test 3	**1500**	**11.77 h**	**86.40%**
Test 4	1500	12.60 h	84.50%

5.2 Voice Classifier

The initial stage of the voice classifier consists on the feature extraction phase, where the MFCCs must be computed. To this end, we have used the C++ library Aquila DSP [17]. For each audio input we computed the first 13 coefficients since it has been shown that they represent good enough the signal features [14]. The audio data is split as shown in Table 4, leaving 70% for training and 30% for testing, each sample with a dimensionality of 13.

Table 4. Audio data distribution for each emotion

	Number of samples	Happiness	Anger	Surprise	Neutral
Dataset	Training	77	77	77	77
	Validation	33	33	33	33
	Total	110	110	110	110

Each One-class SVM of the multiclass audio classifier is trained for each class using a model selection methodology where a grid is built with the values of the generalization and kernel parameters and in each combination of the parameters, a model is trained and tested using both datasets and the chosen model is the one that presents higher generalization accuracy. To this end, we have used the open source C++ package libSVM, a widely and efficient SVM implementation [18].

Table 5. Training parameters and results for each voice classifier

Classifier	Selected kernel		Accuracy
	Sigmoid		
	γ	ν	
Anger	0.0082	0.581	**75%**
Happiness	0.02025	0.2147	**68.75%**
Surprise	0.0000189	0.6291	**81.25%**
Neutral	0.0000125	0.7181	**81.25%**

After performing the grid search technique using the three more commonly used kernel functions, polynomial, gaussian and sigmoid, the chosen kernel was the sigmoid. The results for the three One-class SVMs are shown in Table 5.

6 Results

In this section we show the most important results of our emotion recognition system. Initially, we show the performance of each separate classifier and finally the performance of the complete system. The accuracy of the voting procedure to detect the emotion from the video stream on the testing data was 100% in all cases, meaning that the classifier was capable of detecting all the emotions without mistake. For the voice classifier the 4 One-class SVM outputs are merged into one single multiclass output as described in Sect. 4.2 and the resulting multiclass voice classifier attained a 69.7% of accuracy in the testing set.

Table 6 shows the labels that were assigned by both, the convolutional neural network and the voice classifier during the testing process. The rows correspond to the real class and each column contains the number of data points that were classified as the emotion shown in each column. Recall that the total number of samples for each classifier is shown in Table 1.

Table 6. Labels assigned by each unimodal classifier for each emotion

	Recognized emotion							
	Happiness		Anger		Surprise		Neutral	
	Image	Voice	Image	Voice	Image	Voice	Image	Voice
Happiness	1987	22	104	6	307	2	17	3
Anger	110	4	1936	20	101	2	32	7
Surprise	205	2	45	3	1951	24	43	4
Neutral	149	2	37	4	80	1	1939	26

In the final step, both classifiers are combined as described in Sect. 4.3 taking into account their accuracies and confidence values. The results show that all the input data in the testing set were correctly classified by the multimodal classifier, even though the unimodal classifiers do not agree for all the samples. Table 7 shows the amount of input samples where the unimodal classifiers agree and disagree.

When both classifiers agree on their output, we measure the level of *agreement* (normalized from 0 to 1) between the classifiers as shown in Eq. (3). Figure 8 shows the distribution of the *agreement* coefficient. This shows that the agreement between the classifiers is mostly greater than 0.6.

The other scenario considers that the classifiers do not agree on the emotion that should be assigned to the input data. In this case, the chosen output is the one given by the classifier with highest weighted confidence. Figure 9 shows the

Table 7. Number of testing data samples with agreement and disagreement in the recognized emotions between the two unimodal classifiers

	Amount of data	
	Agreement between the classifiers	Disagreement between the classifiers
Happiness	22	11
Anger	20	13
Surprise	24	9
Neutral	26	7

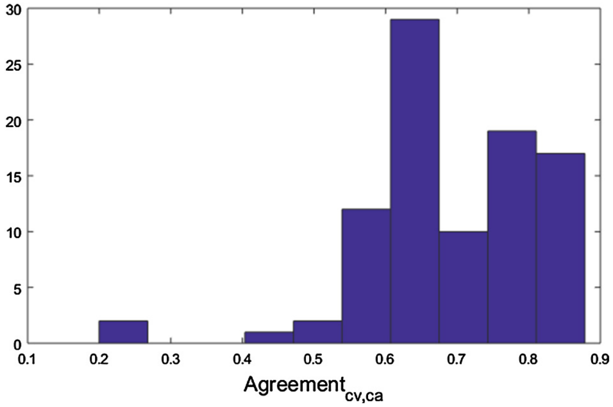
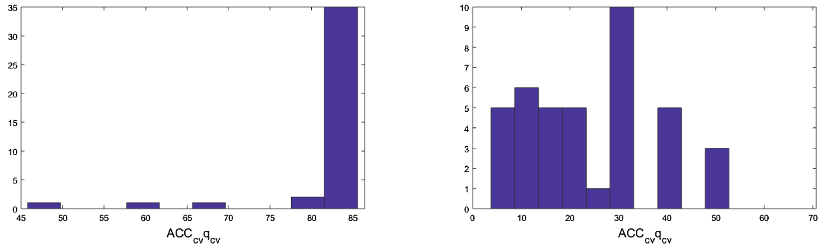


Fig. 8. Distribution of the *agreement* between the video classifier and voice classifier when both classifiers recognize the same emotion.



(a) (b)

Fig. 9. (a) Distribution of the weighted confidence value for the video classifier in the testing set when the classifiers recognize different emotions (b) Distribution of the weighted confidence value for the audio classifier in the testing set when the classifiers recognize different emotions distribution of the confidence values for both, the video (Fig. 9(a)) and audio (Fig. 9(b)). It is noteworthy that most of the confidence values are higher than 80 for the video classifier while the audio classifier shows confidence values lower than 30.

7 Conclusions

We have proposed an automatic classification model to recognize emotions in the context of human-robot interaction using multimodal inputs, i.e., audio and video. Specifically, the proposed case study is a social robot that acts as a sales agent and uses the information captured by the emotion recognition system to drive its speech during her interaction with the human. In this paper we show the behavior of the detection system.

The proposed architecture shows an accuracy of 100% when the classifier uses both data types (audio and video). On one hand, the video classifier is created using a convolutional neural network that is trained with images containing the facial gestures of the human. The individual outputs are merged using a voting procedure to finally decide the total video label. This voting procedure has shown to have a very high accuracy in data that were not used during the training process. On the other hand, the voice classifier uses an audio signal from the processed voice of the human during the interaction. The accuracy of this classifier is of 69.7% for the testing dataset showing that it is possible to use this information to accurately discriminate between the 4 human emotions. Finally, we have proposed a system that provides a confidence value for the classification task.

Currently, we are performing experiments in more open environments, such as with higher levels of noise for both data sources. Preliminary results show that the classifier is more robust to noise due to its use of the multimodal inputs. For future works, the presence of the two different classifiers could be used not only to detect the class, but complimentary classes that could be used by the robot to improve the interaction experience.

References

1. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E., Matsubara, H.: Robocup: a challenge problem for AI. AI Mag. **18**(1), 73 (1997)
2. Christensen, H.I., Batzinger, T., Bekris, K., Bohringer, K., Bordogna, J., Bradski, G., Brock, O., Burnstein, J., Fuhlbrigge, T., Eastman, R., et al.: A roadmap for us robotics: from internet to robotics. Computing Community Consortium (2009)
3. Multi-Annual Roadmap. For horizon 2020. SPARC Robotics, eu-Robotics AISBL, Brussels, Belgium (2017)
4. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426. ACM (2015)

5. Goodrich, M.A., Schultz, A.C.: Human-robot interaction: a survey. Found. Trends Hum. Comput. Interact. **1**(3), 203–275 (2007)
6. van Beek, L., Chen, K., Holz, D., Matamoros, M., Rascon, C., Rudinac, M., des Solar, J.R., Wachsmuth, S.: Robocup@ home 2015: Rule and regulations (2015)
7. Akgun, B., Cakmak, M., Jiang, K., Thomaz, A.L.: Keyframe-based learning from demonstration. Int. J. Soc. Robot. **4**(4), 343–355 (2012)
8. Luo, R.C., Wu, Y.C.: Hand gesture recognition for human-robot interaction for service robot. In: 2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 318–323. IEEE (2012)
9. Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J.F., Salichs, M.A.: A multimodal emotion detection system during human-robot interaction. Sensors **13**(11), 15549–15581 (2013)
10. Subashini, K., Palanivel, S., Ramalingam, V.: Audio-video based classification using SVM and AANN. Int. J. Comput. Appl. **53**(18), 43–49 (2012)
11. Agrawal, U., Giripunje, S., Bajaj, P.: Emotion and gesture recognition with soft computing tool for drivers assistance system in human centered transportation. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4612–4616. IEEE (2013)
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
13. Deng, L., Dong, Y.: Deep learning: methods and applications. Found. Trends Signal Process. **7**(3–4), 197–387 (2014)
14. Rodriguez, S., Pérez, K., Quintero, C., López, J., Rojas, E., Calderón, J.: Identification of multimodal human-robot interaction using combined kernels. In: Snášel, V., Abraham, A., Krömer, P., Pant, M., Muda, A.K. (eds.) Innovations in Bio-Inspired Computing and Applications. AISC, vol. 424, pp. 263–273. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28031-8_23
15. Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al.: Emonets: multimodal deep learning approaches for emotion recognition in video. J. Multimodal User Interfaces **10**(2), 99–111 (2016)
16. Vedaldi, A., Lenc, K.: Matconvnet – convolutional neural networks for MATLAB. In: Proceeding of the ACM International Conference on Multimedia (2015)
17. Django: Aquila digital signal processing C++ library (2014). https://aquila-dsp.org/
18. Libsvm – a library for support vector machines (2015). https://www.csie.ntu.edu.tw/~cjlin/libsvm/