Automatic Emotion Recognition in Robot-Children Interaction for ASD Treatment

Marco Leo, Marco Del Coco, Pierluigi Carcagnì, Cosimo Distante
ISASI UOS Lecce
Campus Universitario via Monteroni sn, 73100 Lecce Italy
marco.leo@cnr.it

Massimo Bernava, Giovanni Pioggia
ISASI UOS Messina
Marine Institute, via Torre Bianca, 98164 Messina Italy

Giuseppe Palestra
Univerisita' di Bari
Via Orabona 4, 70126 Bari, Italy

Abstract

Autism Spectrum Disorders (ASD) are a group of lifelong disabilities that affect people's communication and understanding social cues. The state of the art witnesses how technology, and in particular robotics, may offer promising tools to strengthen the research and therapy of ASD. This work represents the first attempt to use machine-learning strategies during robot-ASD children interactions, in terms of facial expression imitation, making possible an objective evaluation of children's behaviours and then giving the possibility to introduce a metric about the effectiveness of the therapy. In particular, the work focuses on the basic emotion recognition skills. In addition to the aforementioned applicative innovations this work contributes also to introduce a facial expression recognition (FER) engine that automatically detects and tracks the child's face and then recognize emotions on the basis of a machine learning pipeline based on HOG descriptor and Support Vector Machines. Two different experimental sessions were carried out: the first one tested the FER engine on publicly available datasets demonstrating that the proposed pipeline outperforms the existing strategies in terms of recognition accuracy. The second one involved ASD children and it was a preliminary exploration of how the introduction of the FER engine in the therapeutic protocol can be effectively used to monitor children's behaviours.

1. Introduction

Autism Spectrum Disorders (ASD) are a group of lifelong disabilities that affect people's communication and understanding social cues. The state of the art witnesses how technology may offer promising tools to strengthen the research and therapy of ASD. In particular, the area of robotics is introducing tremendous possibilities for innovation in treatment for individuals with ASD. Positive results in the use of robots as attractors or mediators, as well as measurement instruments were reported [25, 2].

Advances in recent years have enabled robots to fulfil a variety of human-like functions and different aesthetic and functional characteristics, i.e. non humanoid, animal like and humanoids, which influence on therapy is currently under investigation. In fact, considerable attention has been given to the robot characteristics, but not as much emphasis has been placed on the best ways to integrate the robot into therapy sessions. Anyhow, the robot is tailored to interact with individuals with ASD. Individual preferences, responses and reactions to the robot features and actions, as well as any possible discomfort due the nature of the disorder, are usually taken into account. Thus, even if the clinical use of interactive robots with individuals with ASD has received considerable media attention over the past decade, the efficacy and effectiveness on such an approach is in its infancy. Moreover, much of the published research is in journals that focus on robotics (e.g., Autonomous Robots, Robotics) rather than in prominent ASD journals or clinically-focused journals. Therefore, it is important to review existing research on the clinical applications, rather than focusing on the advances of the technology. For this purpose, it is crucial to outline a rationale for the clinical use of robots [7]. In general individuals with ASD: (a) exhibit strengths in understanding the physical (object-related) world and relative weaknesses in understanding the social world [12, 1], (b) are more responsive to feedback, even social feedback, when administered via technology rather than a human [17], and (c) are more intrinsically interested in treatment when it involves electronic or robotic components [19, 21, 24]. Scientific evidences show the interaction of individuals with ASD and robots is useful to elicit pro-social behaviours, to maintain attention, to induce spontaneous linguistic behaviour, as well as to decrease stereotyped and repetitive behaviours. In this work, our main effort aims to provide a clinical perspective by focusing attention on a clinical protocol tailored to improve the pre-requisite of theory of mind, i.e. eye contact, joint attention, symbolic play, and the basic emotion recognition skills. The robot serves as a social mediator, eliciting and enhancing interaction between autistic children and people in their surroundings, mainly their therapists and parents [20]. The protocol is organized in levels at mounting difficulties. Each level is dedicated to train a different skill. For each level, different triadic exercises, i.e. robot, subject and therapist/parent, at mounting difficulties are suggested. The therapist begins the subject-robot interaction at the lower level and at the easier exercise. Once the subject is able to comply with the selected exercise, the therapist starts with the next exercise. When all the exercises of a level were succeeded, the therapist starts with the next level. By our knowledge, this work represents the first attempt to use machine-learning strategies during robot-ASD children interactions in terms of facial expression imitation, making possible an objective evaluation of children's behaviours and then giving the possibility to introduce a metric about the effectiveness of this specific therapy. In particular, the work focuses on the basic emotion recognition skills. In addition to the aforementioned applicative innovations this work contributes also to introduce a facial expression recognition (FER) engine that automatically detects and tracks the child's face and then recognize emotions on the basis of a machine learning pipeline based on HOG descriptor [6] and Support Vector Machines [5]. The FER engine introduces itself a level of innovation since it exploits HOG descriptor in a more efficient way with respect to existing works in the state of the art. Besides, it allows real time recognition making children-robot interactions as natural and comfortable as possible. In order to give evidence of the above, two different experimental sessions were carried out: the first one tested the FER engine on publicly available datasets demonstrating that the proposed pipeline outperforms the existing strategies in terms of recognition accuracy. The second one involved ASD children and it was a preliminary exploration of how the introduction of the FER engine in the therapeutic protocol can be effectively used to monitor children's behaviours. The rest of the paper is organized as follows: Section 2 is aimed at an overview of the leading FER approaches presented in literature; Section 3 presents the whole system giving, step by step, a detailed description of each component and a theoretical presentation of the FER engine; Section 4 deals with the presentation and discussion of experimental results, from the preliminary optimization to the tests on field; Section 5 is finally devoted to the conclusions and future works discussion.

2. Related works

This section reports the most relevant works in the literature focusing on Facial Expression Recognition (FER). Proposed solutions can be divided into two main categories: the first category includes the solutions that classify human emotions by processing a set of consecutive images while, the second one, includes the approaches which perform FER on each single image. By working on image sequences much more information is available for the analysis. Usually, the neutral expression is used as a reference and some characteristics of facial traits are tracked over time in order to recognize the evolving expression [8]. To this end, the use of key points and texture information [28], a modified version of well known Local Binary Patterns (LBP) combined with moments [10], a pyramid of LBP [11], a combination of Independent Component Analysis (ICA), Fisher Local Discriminant Analysis (FLDA) and Hidden Markow Models (HMM) [29], optical flow and non-linear features [27], are some of the most effective approaches used to represent facial traits to be tracked over time. The major drawback of these approaches is the inherent assumption that the sequence content evolves from the neutral expression to another one that has to be recognized. This constraint strongly limits their use in real world applications where the evolution of facial expressions is completely unpredictable. For this reason, the most attractive solutions are those performing facial expression recognition on a single image. The approaches in literature that work on a single image can be conveniently categorized depending on the strategies they use to lead towards the recognition of emotions. This way, two main categories arise: *Component Based Approaches* and *Global Approaches*.

Component Based approaches preliminary extract some facial components and then try to classify expressions on the basis of the matching among corresponding components or comparing the geometrical configuration among different components. An example is given by Pantic and Rothkrantz in [18] where a recognition system for facial expression analysis, from static face images by exploiting ten profile-contour fiducial points and 19 fiducial points of the contours of the facial components, is presented. Poursaberi et al. investigate the use of Gauss-Laguerre wavelets in association with geometrical position of fiducial points in order to provide valuable information for the upper/lower face zone [22].

The work in [32] proposes the use of "salient" distance features, which are obtained by extracting patch-based 3D Gabor features, selecting the most discriminative patches and performing patch matching operations. More recently, Happy et al. [9] propose a novel framework for expression recognition by using appearance based features of selected facial patches depending on the position of facial landmarks that are active during emotion elicitation. Then patches are further processed to obtain the salient ones containing the most discriminative features for classification.

Unfortunately, although the idea of making use of a preliminary selection of salient facial components and a subsequent emotion recognition phase based on geometrical or textural matching has been widely investigated, the achieved classification performances do not fulfill the demanding requirements of the technologies that a FER system has to serve. The main unresolved issues concern the alignment of components in different facial images, especially in case of extreme expressions. Moreover, they experienced high computational time due to the load for the fine extraction of the facial components (especially when iterative strategies are used) and then they appear to be not suitable for real world applications especially if low-power systems are involved (e.g. assistive robot, consumer analysis devices).

The above mentioned problems can be overcome by using ”Global Approaches” i.e. approaches that directly try to extract a representation of the expressions from the appearance of the global face. This research area has been deeply investigated, but there is still much effort to do since it is very challenging to find a global set of descriptors able to robustly characterize human expression traits. Some of the most recent related works that have arisen as a consequence of the theoretical improvements in the definition of more reliable local descriptors are listed below. In [26] authors use Local Binary Pattern (LBP) even in low resolution and compressed input images whereas, in [33], the same descriptors are combined with a kernel-based manifold learning method called Kernel Discriminant Isometric map (KDIsomap). Rivera et al. propose in [23] a new descriptor, named local directional number pattern (LDN), that extracts directional information by the use of compass masks and encodes such information using the prominent direction indices. The Sparse Representation-based Classification (SRC) is used with a Local Phase Quantization (LPQ) in [34] and Gabor filters in [14]. An algorithm for facial expression recognition, by integrating curvelet transform and online sequential extreme learning machine (OS-ELM) with radial basis function (RBF) hidden node, is proposed in [30].

3. System Overview

The proposed system is oriented to automatically manage a medical protocol aimed to improve the capacity of children affected by ASD to associate specific emotions to specific facial expressions. The protocol implements the idea to let the robot perform facial expressions and then ask the child to imitate the expression in order to evaluate his emotion imitation capability and to measure the elapsed time between the robot's request and the child's action. As already explained in the introductory section, the common implementation of this protocol delegates to human operators two main roles: on one hand the robot management and on the other hand the evaluation of the children's interaction level. In this way the protocol is affected by a degree of uncertainty and subjectivity. The introduction of machine learning techniques can release the protocol from the evaluation of human operators making information data about the therapy progress through the sessions objective. This paper implements this fundamental advancement by interposing between the two interacting subjects, represented by the child and the robot, a *processing unit* aimed to replace the role of the human operator. In particular a R25 robot, from the US Robokind [1] has been used as robotic component of the system. It is characterized by the capacity to reproduce facial expressions exploiting the huge number of micro actuators in ”his” face. The robot has been then equipped with a *processing unit* whose main elements are highlighted in Figure 1.

The *processing unit* implements the sequence of actions laid down by the protocol by using three different modules: the first one is devoted to the implementation of the robot actions according to the specifications in the protocol (*protocol management module*); the second one performs child's *facial expression recognition* and finally the third one, the *metadata storage module*, focuses on the storage, retrieval and analysis (through statistical and graphical tools) of the achieved meta-data in order to evaluate how the child's behaviors changes in time, along different therapeutic sessions.

The protocol management module deals with the temporal execution of the robot and child actions defined by the protocol. When the therapeutic session starts, the protocol management module waits until the FER engine detects a face in the video stream acquired by the camera on the R25 robot. The detection of a face is then considered as the ascertainment that the interaction between the robot and the child has started and then the *protocol management module* calls the routines supplying the commands to the robot (for performing expressions and synthesizing vocal messages) in order to implement the scheduled steps of the protocol. While the robot speaks and changes facial traits, the *FER engine* continues to process the face of the child in order to recognize if he's imitating the robot. In the meanwhile, the protocol management module waits that the expression recognized by the *FER engine* coincides with that performed by the Robot and, when this happens, the routines supplying the commands to the robot are again called by the *protocol management module* in order to give a positive vocal feedback to the child and to restore the neutral expression on the robot face. If the child fails to imitate the robot,

[1] http://www.robokindrobots.com/

after a predefined time interval, the system detects an imitation failure. Finally, all the information about the running child-robot interaction (success/failure, elapsed time, etc.) is stored by the *meta-data storage module* and then the procedure restarts according to the protocol specification.

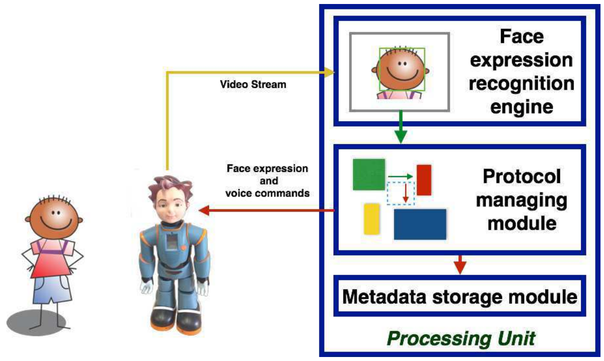


Figure 1. Child-Robot interaction System: the medical protocol is automatically managed by the Processing Unit. The *protocol managing module* gives to the robot the instructions aimed to stimulate the child response while *face expression recognition engine* exploits the robot video stream in order to analyze the child expression. Finally, the child expression and the response time are stored in the *meta-data storage module*

3.1. Robokind™ R25 Robot

As introduced in the previous section, the implemented system makes use of the R25 robot from Robokind ™. The R25 social robot has been designed specifically in order to teach children with autism critical social skills. Since the robot is intended to be interactive, it's equipped with a 5-megapixel autofocus camera in its right eye and 21 motors that provide it with 21 degrees of freedom (7 for the head, 10 for the arms, 1 for the waist and 3 for the legs). In particular, with regard to the head, 2 of the 7 degrees of freedom are related to pitch and yaw; the remaining ones are related to facial expressions.

In order to control all the devices, which the robot is provided with, the R25 is equipped with a OMAP 4460 dual core 1.5 GHz ARM Cortex A9 processor with 1GB of RAM and 16GB of SSD type memory. Regarding to the networking capabilities, the robot is provided of Wi-Fi and Ethernet connections. Finally, the operating system employed is based on Ubuntu Linux and the main software tools are open source as well, allowing easy and fully robot customizations.

3.2. Facial Expression Recognition

Facial expression recognition, from generic images, requires an algorithmic pipeline that involves different operating blocks. The scheme in Figure 2 has been used in this work: the first step detects human faces in the image under investigation and then detected faces are registered [3]. This preliminary operations allow the system to get the quite similar position for the eyes and in this way the subsequent HOG descriptor may be applied using a coherent spatial reference. The vector of features extracted by HOG is finally used for the classification of the facial emotions by means of SVM strategies. Finally, the managing of the temporal images stream is demanded to an ad-hoc decision rule. Each operating step is detailed in the following subsections.

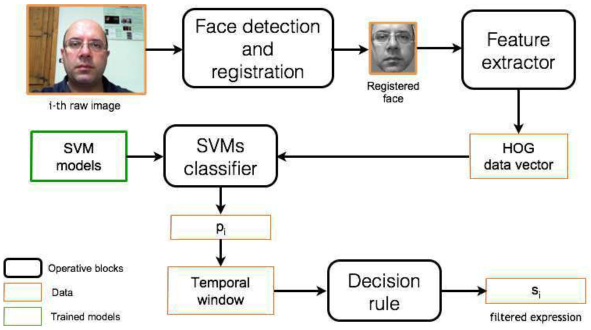


Figure 2. Proposed system pipeline: faces are cropped and registered and then HOG descriptor is applied to build a data vector that is provided as input to a SVM bank that gives the estimation of the observed facial expression. Finally, the prediction is queued in the temporal window exploited by the decision rule in order to filter possible misclassifications.

Face detection and registration

In this step, human faces are detected in the input images and then a registration operation is done. The registration is a fundamental preprocessing step since the subsequent algorithms work better if they can evaluate input faces with predefined size and pose. The face detection is performed by means of the general frontal face detector proposed by [31] which combines increasingly more complex classifiers in a cascade. Whenever a face is detected, the face registration is carried out as follows: the system, at first, fits an ellipse to the face blob (exploiting facial features color models) in order to rotate it to a vertical position and hence a Viola-Jones based eye detector searches the eyes. Finally, eye positions, if detected, provide a measure to crop and scale the frontal face candidate to a standard size of 65 × 59 pixels. The above face registration procedure is schematized in Figure 3. The registered face is then modeled using different features (average color using red-green normalized color space and considering just the center of the estimated face container; eyes patterns; whole face pattern) in order to re-detect it, for tracking purposes, in the subsequent frames [3]. Finally, it is given as input to the features extractor based on the HOG descriptor.

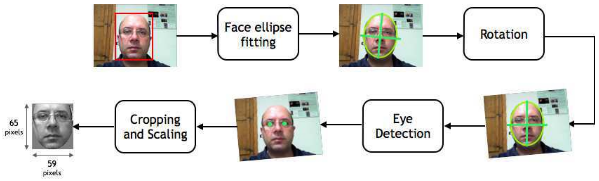


Figure 3. Face registration: the detected face is fitted in an ellipse used to rotate the face in a perfectly vertical position; successively eyes are detected and used to scale the image and to crop the area of interest.

HOG descriptor

Local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. This statement leads to the definition of the HOG technique that has been used in its mature form in Scale Invariant Features Transformation [15] and it has been widely exploited in human detection [6]. HOG descriptor is based on the accumulation of gradient directions over the pixel of a small spatial region referred as "cell" and in the subsequent construction of a 1D histogram whose concatenation supplies the features vector to be considered for further purposes. Let L be an intensity (grayscale) function describing the image to be analysed. The image is divided into cells of size $N \times N$ pixels (as in Figure 4 (a)) and the orientation $\theta_{x,y}$ of the gradient in each pixel is computed (Figure 4 (b-c)) by means of the following rule:

$$\theta_{x,y} = \tan^{-1} \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \qquad (1)$$

Successively, the orientations θ_i^j $i = 1...N^2$, i.e. belonging to the same cell j are quantized and accumulated into a M-bins histogram (Figure 4 (d-e)). Finally, all the achieved histograms are ordered and concatenated into a unique HOG histogram (Figure 4 (f)) that is the final outcome of this algorithmic step, i.e. the features vector to be considered for the subsequent processing.

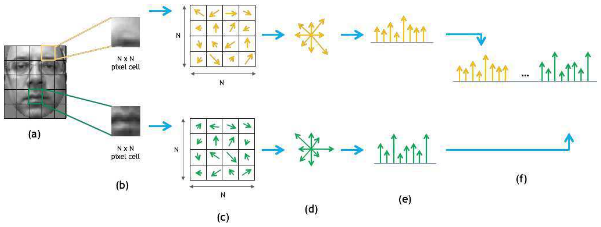


Figure 4. HOG features extraction process: image is divided in cells of size $N \times N$ pixels. The orientation of all pixels is computed and accumulated in an M-bins histogram of orientations. Finally, all cell histograms are concatenated in order to construct the final features vector. The example reports a cell size of 4 pixels and 8 orientation bins for the cell histograms.

SVM prediction

The HOG features vectors are then given as input to a group of Support Vector Machines (SVMs). SVM is a discriminative classifier defined by a separating hyperplane. Given a set of labelled training data (supervised learning), the algorithm computes an optimal hyperplane (the trained model) which categorizes new examples in the right class.

Further theoretical notions about SVM, together with related implementation issues, can be found in [5]. The classical SVM approach is suitable only for a two classes problem but, unfortunately, FER involves multi-class handling. Multi-class problems can be addressed by the "one-against-one" method proposed in [13], an approach based on the construction of an SVM classifier for each pairwise of classes and a voting system aided to elect the predicted class when an unseen item is tested. More specifically, the multi C-support vector classification (multi C-SVC) learning task implemented in the LIBSVM library[4] was used in the experiments reported in Sections 4-6. Radial Basis Function (RBF) was used as kernel for non-linearly separable problems with penalty parameter $C = 1000$ and $\gamma = 0.05$ [4].

Temporal analysis

To make the system suitable for video sequence analysis, a decision making strategy based on the temporal consistency of FER outcomes has been introduced. The decision, about the expression in a video, is taken by analyzing a temporal window of size m and verifying if at least n ($n < m$) frames in the window are classified as containing the same facial expression. More specifically, the system performs a frame by frame analysis in the time window w of size m and for each frame an expression classification outcome is given as p_i where p is the predicted expression and i is the current frame index.

The expression in the window is classified as the expression s if

$$\sum_{j=i-m+1}^{i} (\Lambda(p_j, s)) \geq n \qquad (2)$$

where $\Lambda(p_j, s) = 1$ if $p_j = s$ and 0 otherwise.

This procedure allows the system to manage a temporal stream for a subject avoiding wrong expression predictions due to sporadic miscalssifications. In the tests on field, values of $n = 4$ and $m = 5$ have been used.

4. Experimental Results

As a preliminary step, a facial expressions dataset has been set-up in order to have a benchmark for all tests and, first of all, used in HOG parameters optimization (Subsection 4.1). The recognition performances have been then analyzed by means of confusion tables (Subsection 4.2) and the results compared with those of the leading methods in the literature (subsection 4.3). Finally, in Subsection 4.4, a set of ”on-field” tests have been carried out in order to prove the suitability of the proposed system in a Robot-children interaction task.

4.1. Experimental data setup

The definition of the facial expression dataset is a key factor due to its fundamental role for both training and test purposes. In this work, all the experimental sessions have been carried out on the Cohn-Kanade (CK+) [16]. CK+ is made up by image sequences of people performing 6 facial expressions. Each sequence starts with a neutral face expression and ends with the expressive face. The variety of subjects in terms of gender, as well as ethnicity and age, makes the dataset one of the most used to test the performances of FER solutions.

In order to extract, from the available sequences, a balanced (i.e. quite the same number of instances for each considered expression) subset of images containing expressive faces, the following images were selected: the last image for the sequences related to the expression of anger, disgust and happiness; the last image for the first 68 sequences related to expression of surprise; the last and the fourth from the last images for the sequences related to the expressions of fear and sadness. At the end, a subset of 347 images was obtained with the following distribution among the considered classes of expressions: anger (45), disgust (59), fear (50), happiness (69), sadness (56) and surprise (68). An additional configuration of the previous subset was also introduced in order to test the performance with 7 classes and in this case 60 facial images, with neutral expression, were added to the aforementioned ones.

A first use of the so built subset has been the optimization of HOG parameters (cell size and number of orientation bins). More specifically, FER average recall (referred to the CK+ subset with 6 expression) for different numbers of orientation bins, have been computed and graphically reported onto the y-axis in Figure 5 where the x-axis reports the cell size. From Figure 5 it is possible to infer that a cell size of 7 pixels led to the best FER performance. Concerning the choice of the number of orientations, the best results were obtained with value set to 7 even if also with 9 or 12 orientations the FER performance did not change significantly.

Choosing the optimal parameters configuration (cell size of 7 pixels and 7 orientation bins), the proposed pipeline is able to correctly classify the images, supplied as input during the 10-fold cross validation process, with average performance that can been numerically expressed with a recall of 95.9%, a precision of 95.8%, an accuracy of 98.9% and a F-score of 95.8%. It is worth noting as the configuration (cell size of 7 pixels and 7 orientation bins) has resulted the most performing one also with the dataset with 7 expressions highlighting as, once the preprocessing is given, the parameter configuration is general for the specific problem.

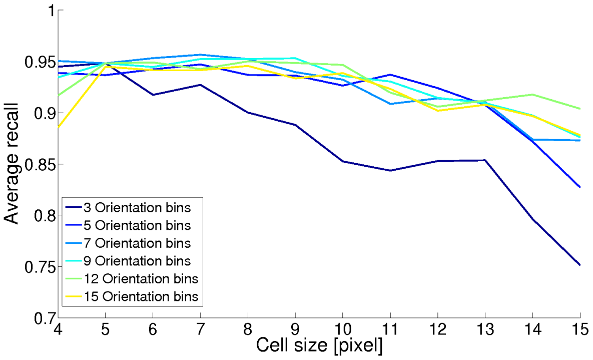


Figure 5. FER results using different cell sizes and number of orientation bins for the HOG descriptor: the x-axis reports the cell size in pixel and the y-axis refers to the average recall percentage.

4.2. Confusion matrices for all the datasets

Once established the unique best configuration of the HOG parameters, the performance of the proposed approach were better analyzed. In particular, in a multi-class recognition problem, as the FER one, the use of an average performance value among all the classes could be not exhaustive since there is no possibility to inspect what is the separation level, in terms of correct classifications, among classes. To overcome this limitation, for each dataset the confusion matrices (expressed in terms of recall in order to keep coherence with other works [9]) are then reported in Tables 1 and 2.

This makes possible a more detailed analysis of the results that can point out the missclassification cases and the interpretation of their possible causes. First of all, from the confusion tables it is possible to observe that the proposed pipeline achieved an average performance value rate over 90% for all the tested datasets and that, as expected, its FER performances decreased when the number of classes, and consequently the problem complexity, increased. In fact, in the case of the CK+ dataset with 6 expressions, the recall was of 95.9% whereas after the addition of the neutral expression it decreased to 94.1%.

Going into a more detailed analysis, Tables 1 and 2 highlight an ambiguity between anger, disgusted and sad expressions. For all the aforementioned expressions, strict lips and low position of eyebrows are in fact very similar, in both location and appearance.

Similarly, the sad expression experimented some erroneous classification in the anger face expression due to the strict lips and low position of eyebrows that are very similar for the two expressions. Finally, the happy expression is the most insensitive to ambiguities and reached the 100% of classification in all the tests.

Table 1. Performance of proposed approach (CK+ 6 expressions). Average performances: recall = 95.9%, precision = 95.8%, accuracy = 98.8%, F-score = 95.8%. (orientation bins = 7, cell size = 7). An=Anger, Di=Disgusted, Fe=Fearful, Ha=Happy, Sa=Sad, Su=Surprised.

	An	Di	Fe	Ha	Sa	Su
An	**88.6**	4.5	2.4	0	4.5	0
Di	5.6	**89.0**	1.8	1.8	0	1.8
Fe	0	0	**100**	0	0	0
Ha	0	0	0	**100**	0	0
Sa	0	0	0	0	**100**	0
Su	1.3	0	1.3	0	0	**97.4**

Table 2. Performance of proposed approach (CK+ 7 expressions). Average performances: recall = 94.1%, precision = 94.3%, accuracy = 98.5%, F-score = 94.1%. (orientation bins = 7, cell size = 7). Ne=Neutral, An=Anger, Di=Disgusted, Fe=Fearful, Ha=Happy, Sa=Sad, Su=Surprised.

	Ne	An	Di	Fe	Ha	Sa	Su
Ne	**89.6**	1.8	0	0	0	8.6	0
An	4.4	**86.8**	4.4	0	0	4.4	0
Di	0	5.4	**92.9**	1.7	0	0	0
Fe	0	0	0	**93.9**	4.1	0	2.0
Ha	0	0	0	0	**100**	0	0
Sa	0	0	1.8	0	0	**98.2**	0
Su	1.3	0	0	1.3	0	0	**97.4**

4.3. Comparison with the state of the art

In this subsection, achieved results are compared with those of the leading State-of-the-Art FER solutions. Differently from other research fields, in the FER one there is not a shared dataset to be used as benchmark for a fair evaluation of different algorithms. In order to avoid errors introduced by the re-implementation of each method, most of the works in the literature refer to CK+ dataset that, unfortunately, is dramatically unbalanced and, for this reason, it requires a selection of a subset of available expression occurrences before to be used. How this selection has to be performed is not well stated and then the reported comparisons are always biased by this important drawback. In addition, there is not a standardised evaluation procedure even if most of the works make use of a cross-validation procedure and then represent the recognition performances by means of confusion matrices. In order to accomplish this crucial task, in this paper, the approach implemented in the most up-to-date works for FER recognition is then used [9], i.e. a CK+ subset of observations was selected and then a k-fold cross validation was used to fill the confusion matrix. In this way, the performance of comparing approaches can be extracted from the relative papers avoiding to affect them by implementation issues. To be fairest as possible, for most of the sequences in the dataset, we retained only one image (we experimentally proved that choosing more than one image the overall performance increase). For balancing reasons, for a few sequences also the fourth image from the last one was retained and, in this way, the largest possible subset of images (in respecting a reasonable balancing among categories) was build. Choosing the fourth from the last image introduces less correlation then the third from the last or the second from the last.

Table 3 reports the comparison results demonstrating that the proposed approach gave the best average recognition rate. In particular, it is worth noting that the performance achieved by the approach under investigation exceed also those of the recent work in [9] that represents the reference point for the FER problem. A deeper analysis of the Table 3 evidences that the proposed method suffers more than competitors to recognize the expression of disgust. This drawback could be due to the fact that, while performing this expression, the facial muscles shape is quite similar to that of the expression of anger hence the edge analysis performed by HOG, sometimes, cannot be able to bring to light differences as other approaches based on texture analysis can instead highlight. However, this is a limitation only for the recognition of the expression of disgust since for all the remaining expressions the FER performances of the proposed method largely exceed those of the comparing methods highlighting that the analysis of the edges is the best method to recognize facial expressions as it throws away all possible ambiguity introduced by non-edge based features.

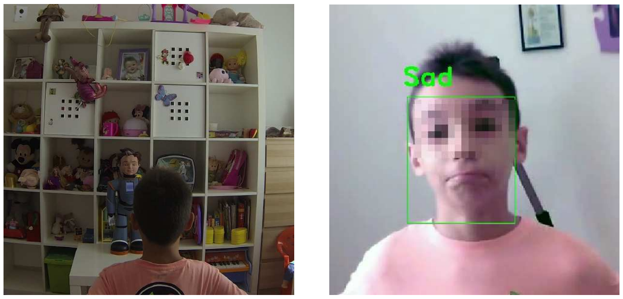
Table 3. Performance comparison of our approach versus different State-of-the-Art approaches (CK+ 6 expressions). An=Anger, Di=Disgusted, Fe=Fearful, Ha=Happy, Sa=Sad, Su=Surprised.

	[29]	[22]	[35]	[32]	[9]	PROPOSED
An	82.5	87.1	71.4	87.1	87.8	88.6
Di	97.5	91.6	95.3	90.2	93.3	89.0
Fe	95.0	91.0	81.1	92.0	94.3	100
Ha	100	96.9	95.4	98.1	94.2	100
Sa	92.5	84.6	88.0	91.5	96.4	100
Su	92.5	91.2	98.3	100	98.5	97.4
AV	93.3	90.4	88.3	93.1	94.1	**95.8**

4.4. Tests on field

This subsection reports a set of preliminary experiments carried out involving 3 children with ASD (high-functioning autism or Asperger's syndrome). This experimental phase had a twofold goal: on the one side it was addressed to verify if the system implements a procedure that makes the children able to interact in a comfortable and natural way and, on the other side, to test the system's components (in particular the FER engine) in a real environment.

As indicated by the medical protocol, 4 different expressions were investigated (happiness, sadness, anger and fear).Each child was placed in front of the robot and he/her was asked to imitate the expression performed by the robot for 5 times (then the total number of interactions monitored was 60, 20 for each child). The recognition model trained using the CK+ subset with 7 expression was used. The *processing unit* was running on an external hardware (e.g. a MacBook Pro with i7 class processor) wireless connected: this choice was made in order to get a higher number of frames processed per second (25fps), assuring this way a natural interaction. The use of the processing resources available on board of the robot may be afterwards taken into consideration only following an algorithmic optimization of the procedures that must be optimize to exploit at best the limited computational resources of the on board processing architecture. In Figure 6 the experimental environment is shown. The system was able to successfully complete the protocol for all the children. In particular 31 interactions were successfully completed with the imitation by the children (9 happiness, 6 sadness, 8 anger and 8 fear) and correctly recognized by the system. On the other hand, the imitation completely failed for 26 interactions. More precisely in 19 interactions children did not put in place the imitation since they were attracted by the robot components or they were tired/annoyed whereas, in the remaining 7 ones, the imitation was wrong performed and consequently not matched by the system. The imitation results were confirmed by the personnel who attended the experiments. The remaining 3 interactions have to be deeply analyzed since in those cases the FER engine did not match the child's expression with the expected one (1 anger and 2 fear) even if the child tried the imitation (as pointed out by the attending personnel). An off-line analysis of the images acquired by the camera on board of the robot (notice that during the experiments all the images were stored for debug purposes) revealed that in 1 case the face was not detected since the child strongly rotated it while imitating the expression of fear whereas, in the last two cases, the expressions of anger were misclassified as fear.



(a) External view

(b) Robot view

Figure 6. An example of interaction: the child is in front of the robot. The FER engine exploits the robot view in order to process the child expression

5. Conclusion and Future Works

This work introduced machine-learning strategies during robot-ASD children interactions in order to make possible an objective evaluation of children's behaviours and then to give the possibility to introduce a metric about the effectiveness of the therapy. In particular, the work focused on the basic emotion recognition skills and it contributed to introduce a facial expression recognition (FER) engine that automatically detects and tracks the child's face and then recognize emotions on the basis of a machine learning pipeline based on HOG descriptor and Support Vector Machines. Two different experimental sessions were carried out: the first one tested the FER engine on publicly available datasets demonstrating that the proposed pipeline outperforms the existing strategies in terms of recognition accuracy; the second one involved ASD children and it was a preliminary exploration of how the introduction of the FER engine in the therapeutic protocol can be effectively used to monitor children's behaviours. Future works will deal with: 1) the optimization of the algorithms involved in the FER engine in order to exploit the processing resources available on board of the R25 robot 2) evaluate the systems along multiple therapeutic sessions involving the same children in order to take advantage of the analysis tools implemented by the *meta-data handling module*.

Acknowledgments

The authors would like to thank the non-profit organization ”Amici di Nico” in Matino (LE) - Italy and, in particular, Dr. Annachiara Rosato.

References

[1] S. M. Anzalone, E. Tilmont, S. Boucenna, J. Xavier, A.-L. Jouen, N. Bodeau, K. Maharatna, M. Chetouani, and D. Cohen. How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3d + time) environment during a joint attention induction task with a robot. *Research in Autism Spectrum Disorders*, 8(7):814 – 826, 2014.

[2] J.-J. Cabibihan, H. Javed, J. Ang, Marcelo, and S. Aljunied. Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. *International Journal of Social Robotics*, 5(4):593–618, 2013.

[3] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández. Encara2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, 18(2):130–140, 2007.

[4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[7] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell. The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in autism spectrum disorders*, 6(1):249–262, 2012.

[8] F. Dornaika, E. Lazkano, and B. Sierra. Improving dynamic facial expression recognition with feature subset selection. *Pattern Recognition Letters*, 32(5):740 – 748, 2011.

[9] S. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *Affective Computing, IEEE Transactions on*, PP(99):1–1, 2015.

[10] Y. Ji and K. Idrissi. Automatic facial expression recognition based on spatiotemporal descriptors. *Pattern Recognition Letters*, 33(10):1373 – 1380, 2012.

[11] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz. Framework for reliable, real-time facial expression recognition for low resolution images. *Pattern Recognition Letters*, 34(10):1159 – 1168, 2013.

[12] A. Klin, D. J. Lin, P. Gorrindo, G. Ramsay, and W. Jones. Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, 459(7244):257–261, 05 2009.

[13] S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing*, volume 68 of *NATO ASI Series*, pages 41–50. Springer, 1990.

[14] W. Liu, C. Song, Y. Wang, and L. Jia. Facial expression recognition based on gabor features and sparse representation. In *Control Automation Robotics Vision (ICARCV), 12th International Conference on*, pages 1402–1406, Dec 2012.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, June 2010.

[17] S. Ozonoff. Reliability and validity of the wisconsin card sorting test in studies of autism. *Neuropsychology*, 9(4):491, 1995.

[18] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(3):1449–1461, June 2004.

[19] A. Peca, R. Simut, S. Pintea, C. Costescu, and B. Vanderborght. How do typically developing children and children with autism perceive different social robots? *Computers in Human Behavior*, 41:268–277, 2014.

[20] F. Petric. Robotic autism spectrum disorder diagnostic protocol: Basis for cognitive and interactive robotic systems.

[21] G. Pioggia, R. Igliozzi, M. Ferro, A. Ahluwalia, F. Muratori, and D. De Rossi. An android for enhancing social skills and emotion recognition in people with autism. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 13(4):507–515, 2005.

[22] A. Poursaberi, H. Noubari, M. Gavrilova, and S. Yanushkevich. Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification. *EURASIP Journal on Image and Video Processing*, 2012(1), 2012.

[23] A. Ramirez Rivera, R. Castillo, and O. Chae. Local directional number pattern for face analysis: Face and expression recognition. *Image Processing, IEEE Transactions on*, 22(5):1740–1752, May 2013.

[24] D. J. Ricks and M. B. Colton. Trends and considerations in robot-assisted autism therapy. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 4354–4359. IEEE, 2010.

[25] B. Scassellati, H. Admoni, and M. Mataric. Robots for use in autism research. *Annual review of biomedical engineering*, 14:275–294, 2012.

[26] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803 – 816, 2009.

[27] M. Siddiqi, R. Ali, A. Khan, E. Kim, G. Kim, and S. Lee. Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. *Multimedia Systems*, pages 1–15, 2014.

[28] M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou. Image ratio features for facial expression recognition application. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(3):779–788, June 2010.

[29] M. Uddin, J. Lee, and T.-S. Kim. An enhanced independent component-based human facial expression recognition from video. *Consumer Electronics, IEEE Transactions on*, 55(4):2216–2224, November 2009.

[30] A. Uçar, Y. Demir, and C. Güzeliş. A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering. *Neural Computing and Applications*, pages 1–12, 2014.

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

[32] L. Zhang and D. Tjondronegoro. Facial expression recognition using facial movement features. *Affective Computing, IEEE Transactions on*, 2(4):219–229, Oct 2011.

[33] X. Zhao and S. Zhang. Facial expression recognition based on local binary patterns and kernel discriminant isomap. *Sensors*, 11(10):9573–9588, 2011.

[34] W. Zhen and Y. Zilu. Facial expression recognition based on local phase quantization and sparse representation. In *Natural Computation (ICNC), Eighth International Conference on*, pages 222–225, May 2012.

[35] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2562–2569, June 2012.