Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Project No. 611391

DREAM
Development of Robot-enhanced Therapy for Children with Autism Spectrum Disorders

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 611391

D5.4 Diagnostic Tools

Due date: **31/03/2017**
Submission Date: **11/04/2017**

Start date of project: **01/04/2014** Duration: **54 months**

Organisation name of lead contractor for this deliverable: **University of Skövde**

Responsible Person: **Serge Thill** Revision: **1**

Project co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Service)	
CO	Confidential, only for members of the consortium (including the Commission Service)	

Contents

Executive Summary

The purpose of this deliverable is to document the suitability of the algorithms documented in D5.1 and D5.2 for diagnostic purposes that could be of general use for therapists. The document addresses this in three steps: First, we provide a general overview of the suitability of current methods for the automation of the diagnostic process. Second, we discuss the degree to which the current runtime system can assist intervention session evaluations. Together, these two steps fulfil the purpose of the deliverable, and the therapist perspective of these aspects are documented in corresponding deliverables in WP2 (D2.2.1 and D2.3.1).

Third, we also return to ongoing efforts to automate the process of video annotation for the therapists. In particular, given results previously documented in D5.1, we now present a first deep learning approach. Current results in this respect are still not satisfactory, and efforts in this respect will therefore continue.

Principal Contributors

The main authors of this deliverable are as follows (in alphabetical order).

Paul Baxter, University of Plymouth
Erik Billing, University of Skövde
Cristina Costescu, Babeş-Bolyai University
Sergey Reydyuk, University of Skövde
Serge Thill, University of Skövde

Revision History

Version 1.0 (S.T. 11-04-2017)
Final version.

1 Introduction

The purpose of deliverable 5.4 is to document work carried out in the remainder of WP5 in a manner that is useful to professionals outside the DREAM project itself; in particular with respect towards automating parts of the diagnostic process.

It is important to note (as documented in D5.2) that the distinction between the functionality necessary to operate a running DREAM system has become rather distinct from the functionality necessary for more general diagnostic purposes, and that the algorithms developed for the runtime system are very dependent on the specific hardware used. Here, we therefore devote less space to a detailed discussion of the functioning of the run-time system (referring instead to D5.2 and – more importantly – the source code available in the project's repository for these aspects), and more space to a more general evaluation of the needs of a more general automation of diagnostic processes and the degree to which the current run-time system can address these.

Specifically, we will threfore address three aspects of diagnostic tools in this deliverable:

1. General requirements for automation of diagnostic processes for ASD
2. Use of the run-time system for diagnostic purposes
3. Automatic annotation of intervention videos

2 General requirements for automation of diagnostic processes for ASD

This part of the present document has a sister part in WP2, documented in D2.2.1. In particular, we refer to table 1 of D2.2.1 (reproduced in the present document), which documents behavioural cues used in diagnosis mapped onto the modalities required to adequately assess these cues. The present section complements D2.2.1 by addressing the present-day feasibility of automating the detection and adequate classification of these cues. Note that this is significantly different from the detection and classification required for the run-time system: in this section, we are primarily concerned with *diagnosis*, not assessing performance or engagement with respect to intervention scripts.

Table 1: The table mapping diagnostic criteria to modalities, reproduced from D2.2.1

	Required modalities									Classif.	
	Gaze tracking	Speech detection	Speech analysis	Posture tracking	Gesture tracking	Facial Expressions	Object tracking	Sound detection	Specific events	Covert behaviour	Interaction-centred
Category A **Persistent deficits in social communication and social interaction across contexts**											
A1 Deficits in social-emotional reciprocity											
1. One-sided conversations	·	✓	·	·	·	·	·	·	·	·	✓
2. Failure to offer comfort to others or to ask for it when needed	·	·	✓	·	✓	·	·	·	✓	✓	✓

Table 1: The table mapping diagnostic criteria to modalities, reproduced from D2.2.1

	Required modalities									Classif.	
	Gaze tracking	Speech detection	Speech analysis	Posture tracking	Gesture tracking	Facial Expressions	Object tracking	Sound detection	Specific events	Covert behaviour	Interaction-centred
3. Does not initiate conversation with peers	·	✓	✓	✓	·	·	·	·	·	✓	✓
4. Lack of showing, bringing, or pointing out objects of interest to other people	·	·	·	✓	✓	·	✓	·	·	✓	✓
5. Use of others as tools	·	·	·	✓	✓	·	·	·	·	·	✓
6. Failure to engage in simple social games	·	·	·	✓	✓	·	·	·	·	✓	✓
A2 Deficits in nonverbal communicative behaviours used for social interaction											
1. Impairments in social use of eye contact	✓	·	·	·	·	·	·	·	·	·	✓
2. Limited communication of own affect	·	✓	✓	·	✓	✓	·	·	·	✓	·
3. Abnormalities in the use and understanding of emotion	·	·	·	✓	✓	✓	·	·	·	✓	✓
4. Impairment in the use of gestures	·	·	·	·	✓	·	·	·	·	·	·
5. Abnormal volume, pitch, intonation, rate, rhythm, stress, prosody or volume in speech	·	✓	·	·	·	·	·	·	·	·	·
6. Lack of coordinated verbal and nonverbal communication	✓	✓	✓	·	✓	·	·	·	·	✓	·
A3 Deficits in nonverbal communicative behaviours used for social interaction											
1. Lacks understanding of the conventions of social interaction	·	✓	·	·	✓	·	·	·	·	✓	✓
2. Limited interaction with others in discussions and play	✓	·	✓	✓	·	·	·	·	·	✓	✓
3. Limited interests in talking with others	·	·	✓	·	·	·	·	·	·	✓	·
4. Prefers solitary activities	·	·	·	✓	✓	·	·	·	·	✓	✓

Table 1: The table mapping diagnostic criteria to modalities, reproduced from D2.2.1

	Required modalities									Classif.	
	Gaze tracking	Speech detection	Speech analysis	Posture tracking	Gesture tracking	Facial Expressions	Object tracking	Sound detection	Specific events	Covert behaviour	Interaction-centred
5. Limited recognition of social emotions	✓	✓	·	·	·	✓	·	·		·	✓
Category B **Restricted, repetitive patterns of behaviour, interests, or activities as manifested**											
B1 Stereotyped or repetitive speech, motor movements, or use of objects											
1. Repetitive hand movements	·	·	·	·	✓	·	·	·	·	·	·
2. Stereotyped or complex whole body movements	·	·	·	✓	·	·	·	·	·	·	·
3. Repetitive vocalizations such as repetitive guttural sounds, intonational noise making, unusual squealing, repetitive humming	·	✓	·	·	·	·	·	·	·	·	·
4. Perseverative or repetitive action / play / behaviour	·	·	·	✓	✓	·	✓	·	·	·	·
5. Pedantic speech or unusually formal language	·	·	✓	·	·	·	·	·	·	✓	·
B2 Excessive adherence to routines, ritualized patterns of verbal or nonverbal behaviour, or excessive resistance to change											
1. Overreactions to changes	·	·	✓	·	✓	✓	·	·	·	✓	✓
2. Unusual routines	·	·	·	·	✓	·	✓	·	·	✓	·
3. Repetitive questioning about a particular topic	·	·	✓	·	·	·	·	·	·	✓	·
4. Compulsions	·	·	·	✓	✓	·	·	·	·	✓	·
B3 Highly restricted, fixated interests that are abnormal in intensity or focus											
1. Focused on the same few objects, topics or activities	✓	✓	·	✓	✓	·	✓	·	·	✓	·
2. Verbal rituals	·	✓	✓	·	·	·	·	·	·	✓	·
3. Excessive focus on irrelevant or non-functional parts of objects	✓	·	✓	·	✓	·	✓	·	·	✓	·
B4 Hyper- or hypo-reactivity to sensory input or unusual interest in sensory aspects of environment											

Table 1: The table mapping diagnostic criteria to modalities, reproduced from D2.2.1

	Required modalities									Classif.	
	Gaze tracking	Speech detection	Speech analysis	Posture tracking	Gesture tracking	Facial Expressions	Object tracking	Sound detection	Specific events	Covert behaviour	Interaction-centred
1. Abnormal responses to sensory input	·	·	·	✓	·	✓	·	·	✓	✓	·
2. Repetitively putting hands over ears	·	·	·	✓	·	·	·	✓	·	·	·
3. Extreme interest or fascination with watching movement of other things	·	·	·	✓	✓	·	✓	·	·	✓	·
4. Close visual inspection of objects	·	·	·	✓	✓	·	✓	·	·	·	·

2.1 The distinction between overt and covert aspects of behaviour

Being sensory based, automation must necessarily focus on observable behaviours of the children within a diagnosis interaction. We can further divide these behaviours into two classes: "overt" (Ov) behaviour is that which is directly observable (physical movement, gaze, posture, etc), whereas "covert" (Co) behaviour requires some degree of human interpretation of a range of overtly observable behaviours. Covert behaviours, for example, include emotional state (as opposed to emotional expression), and non-appropriate behaviours (such as speech intonation), as these are dependent on context and inference of either internal state and/or qualitative characteristics. Automated methods are primarily beneficial for overt behaviour but can nonetheless assist interpretation of covert behaviours.

2.2 Locus and interactivity

The locus and interactivity of such behaviours must also be considered. This leads to the consideration of criteria as being either "Child-Centred" (CC) or "Interaction-Centred" (IC). Child-centred criteria are those for which only the behaviour of the child needs to be considered, while interaction-centered criteria require the sensing of both interaction parties to provide an accurate assessment: these impose additional challenges for automated methods: at a minimum, both the child and the therapist need to be covered by the sensory apparatus to capture the information necessary to characterise interaction-centred criteria behaviours.

2.3 Necessary sensory modalities

Based on the classification of behavioural cues as being observable (Ov/Co) and interactive (CC/IC), a set of behavioural modalities involved in assessing each of the criteria can be derived (see Table 1). From this, a set of sensors and processing methods can be determined for each modality. Five

major modalities may be distinguished: posture/gesture, gaze, facial expression, speech processing, and object/sound detection. There are a number of computational methodologies applicable to each behavioural modality, each of which is the subject of ongoing research and development. A complete overview of each of these would be both excessive and quickly out of date; instead, we provide a discussion of the current trends and limitations. Further, given the rigorous protocols for diagnosis, and the types of sensitivities prevalent among children with ASD, we focus on technologies that do not interfere physically with the child. Wearable sensors or similar technologies that are physically attached to the person of the child are therefore not considered as part of this overview.

2.3.1 Posture

The characterisation of (aspects of) postures (focused on the gross skeletons) and gestures (typically focused on hand/wrist use) of both the child and therapist covers a significant proportion (23 out of 33) of the example criteria. Vision-based methods (using standard cameras/2D images) for human motion capture are well established (Moeslund et al., 2006) , with face tracking being particularly developed. The recent advent of depth-based tracking and processing of detected skeletons in the scene (primarily using RGB-D data) resulted in additional well-established tools to facilitate various types of pose and behaviour analysis Han et al. (2013) , though naturally sensitive to occlusions. Depth-based methods can also be applied to hand gesture characterisation (Suarez and Murphy, 2012), although sensory resolution constraints (hands and fingers being more difficult to detect) mean that image-based methods may currently remain more appropriate (Mitra and Acharya, 2007). In terms of information directly relevant to the diagnosis criteria, these methods are typically targeted at the characterisation of individuals, and so would be most appropriate to overt and child-centred behaviours, followed by overt and interaction-centred behaviours, provided that both parties of the interaction are tracked.

2.3.2 Gaze

Gaze is an essential cue in social interaction. Two aspects may be tracked: head direction (which overlaps with posture detection) and eye gaze. Head direction tracking is relatively robust, and with a number of readily available algorithms, e.g. (Zhu and Ramanan, 2012). Eye gaze tracking, however, provides a much better indication of the orientation of visual attention, but currently available methods are not as robust to noise (e.g. lighting conditions) or variations in head pose (Hansen and Ji, 2010). For both aspects, appropriate high-resolution camera placement is critical, with a typical requirement for a full-frontal view of the face. In a practical application setting this is frequently not possible, thus requiring multiple cameras instead. Recent methods that allow switching between multiple cameras to provide the best viewpoint facilitate this (Cai et al., 2015).

2.3.3 Facial expressions

The characterisation of facial expressions focuses on emotional expression. Emotion classification from faces is typically based on the basic emotions, which are often limited in real applications (Pantic and Rothkrantz, 2000). However, given that the therapist acts out the emotional expression (thus exaggerating the features), such methods may nevertheless be appropriate. Classification methods typically use Action Unit coding of facial expression features, with more recent attempts to incorporate other visual information, such as head behaviour (Zeng et al., 2007). Being a camera-based method, this characterisation of facial expression is subject to similar constraints as posture and gaze analysis.

2.3.4 Speech

The requirement for some form of speech analysis appears in 17 of the example criteria. Speech processing has received increasing attention in recent years as commercial applications have come to the public. Solutions therefore exist that could be applied to automated analysis of diagnosis interaction speech, although variability between speakers poses problems (Benzeghiba et al., 2007) that are particularly acute with child voices (Gerosa et al., 2007). There are two broad types of speech properties that may be distinguished in the context of the diagnosis criteria: (1) detection of the presence/absence of speech (10 criteria); and (2) the processing of the content of the speech (comprised on detection of reportativity, keyword recognition and understanding 11 criteria). The first level of these can be addressed through the application of statistically-based signal processing techniques, for which there are a range of established solutions (*e.g.* Rabiner and Schafer, 2007; Ramirez et al., 2007). Keyword recognition (which could also be used for repetition detection) lies in the area of speech recognition that is similarly well supported by a range of methods (Rabiner and Schafer, 2007), including more recently deep learning systems (Hinton et al., 2012), although there are limitations in real-world contexts. Speech understanding poses the most challenging level of analysis, with current technologies being limited to constrained settings until a greater level of context information can be incorporated (Moore, 2007). In all of these cases, maximising quality of the sound recordings using microphones (while minimising background noise, interference, etc), is clearly beneficial to maximise performance of automated methods. In application to the diagnosis interaction scenario this may necessitate the deployment of multiple microphones, which introduces further issues of signal integration and sound source localisation, particularly with multiple speakers (the child and the therapist) present (Athanasopoulos et al., 2015).

2.3.5 Object and sound detection

The tracking of objects and sounds appear in eight of the diagnosis criteria. This is considered separately from the modalities in the paragraphs above since it is not directed specifically at either of the humans; however, where manipulation of items is involved, there is an overlap with posture/gesture recognition. The same set of sensors may be deployed as for the other behavioural modalities, namely cameras (using 2D and depth images) and microphones. There are a range of well-established methods/algorithms in the literature that are effective for object tracking based on vision data, with recent advances using deep learning methods (*e.g.* Jia et al., 2014). However, if manipulation is involved (in items B1.4 and B3.1 for example), then object occlusions may be problematic.

2.3.6 Other types of events

While only appearing twice in the example criteria, the detection of specific types of events presents variations on, and conjunctions of, the behaviour modalities discussed above: firstly, the timing and precise content of these events is unpredictable and not tied to a particular part of the protocol, and secondly, the location of these events may be away from the locus of the diagnosis interaction. Both of these aspects present additional challenges to automated detection and characterisation. The second aspect, however, gives rise to the possible necessity for an entire room be augmented with sensors to capture all potential poses and positions, of both humans and objects. This possibility would also address issues with (partial) occlusions that would reduce the efficacy of the single modality methods described above. While such environment augmentation is a potential solution, it brings additional cost (in terms of equipment and human involvement required) and data management (in terms of

sensor coordination, volume, and validation) problems that may be difficult to surmount for real-world deployment.

2.4 Summary

It is apparent from the above that while there is definite scope for such automated quantification, there are a number of limitations of sensory technologies and their associated methods. A number of these are due to practical sensory constraints (e.g. the positioning and coverage of individual sensors), but the more problematic issues are typically related to those criteria involving a covert behavioural component, i.e. those behaviours that require some degree of interpretation in addition to the observation of the overt phenomena. Human therapists/assessors naturally bring their prior experience and extensive training into the diagnostic assessment process; for automated methods, this prior knowledge and experience must be codified for it to be applied. The problematic qualitative nature of such developed experience is an area in which the sensory interpretation methods discussed are currently lacking, for which deeper/more complex (perhaps even cognitive) models are required if they are to be sufficient to adequately augment human characterisation efforts.

Twelve behavioural cues have been classed as being characterisable solely using overt behaviours. As described above, the available technology is particularly suitable for such behaviours, since interpretation based on qualitative human experience is not required. A further complicating factor in the automated assessment process is the presence of Interaction Centred (IC) behavioural cues (13 of the 33 behavioural cues). These entail the tracking and characterisation of multiple individuals (minimally the child and the therapist) and their coordination, which is feasible, though posing additional challenges. Taking these two aspects together, it is noticeable that some of the modalities lend themselves more readily to immediate application than others (see supplementary material), gesture tracking being the clearest example of this. Conversely however, speech analysis clearly remains a challenge, even assuming high performing speech recognition.

3 Use of the run-time system for diagnostic purposes

The run-time system permits the logging of all relevant variables over time; for the present purpose these include in particular the performance and engagement assessments (see D5.2), as well as the various behaviour classifications, position in the intervention script, and so forth. In other words, these logs permit the analysis of interventions, such as documented in deliverable D2.3.1. As also noted in the previous section, this is distinct from the requirements imposed upon the automation of diagnosis.

As such, the diagnostic use of the run-time system data rather lies in tracking statistic pertaining to the performance in intervention tasks (such as in D2.3.1). Since the system's primary focus lies on the specific interventions used in the DREAM context, this functionality does not automatically translate onto other contexts (other than as a demonstration of how one might achieve that given suitable performance assessment algorithms for different types of interventions). The assessment of child engagement (see D5.2) is an exception to this since this is not used directly in the run-time system, but only in later evaluations. In addition, the definition of engagement used for these purposes (see also D5.2) is kept at a general level.

In terms of the general requirements for automating diagnosis as such – discussed in the previous section – the hardware design of the DREAM setup means that it is most suitable for tracking overt, child-centered indicators that do not require speech processing (since the hardware is designed with

the needs of tracking performance during interventions in mind). This is essentially in line with the general technological feasibilities and limitations discussed in the previous section. To fully address automation of diagnosis, as in the previous section would require hardware setups that go significantly beyond the scope considered in the DREAM DoW.

4 Automatic annotation of intervention videos

As already discussed in deliverable D5.1, assisting therapists in annotating videos of interventions is a highly desirable endeavour. A well-functioning system could significantly reduce the burden on therapists while simultaneously ensuring higher accuracy, which in turn could be beneficial for the training of future assessment or diagnostic tools that require some training data in their development. In brief, possible applications include:

- **therapist's assistant** – a system which preprocesses recorded interventions and suggests annotations for the child's positive and negative emotions;
- **therapist's cross-validation tool** – a system which analyses therapists' annotations and makes annotations' timestamps more accurate;
- **annotation generator** – a system which analyses recorded interventions and produces annotations automatically;

The results obtained in D5.1 with respect to this aim were somewhat modest. This continues to be the case for the new efforts documented here; it is clear that this is a hard problem and although progress is made, there is plenty of room for improvement.

D5.1 documented proof-of-concept results using a number of comparatively simple classifiers, and using Echo State Networks. We identified deep learning approaches as a viable alternative - here we present work using such an approach. In order to keep the task manageable, we essentially cast it as a facial expression classification task[1].

It is worth mentioning, as noted in D5.1, that the initial annotations provided by therapists are sparse, somewhat imprecise, and do not cover every frame of the video record in most cases (approximately 90% of the frames) the data is not labelled which decreases a training set and makes it harder to use the data for supervised machine learning algorithms. Efforts to improve the quality of these annotations themselves are also ongoing (see D5.1 for a longer discussion), but the work documented here primarily targets the exploration of deep learning algorithms in the context of somewhat unreliable training data.

4.1 Available Data

102 recorded interventions (over 10 hours of video records) with four autistic children were used as a test set. Annotations for these video files provided by therapists served as a ground truth (with the caveats previously mentioned and discussed in more detail in D5.1) for the solution evaluation stage of classifying the observed behaviour for each point in time. To facilitate the analysis stage, all the files were renamed according to the template and all nested catalogues were restructured to simplify data parsing.

[1] note that this is entirely different from the facial expression classifications available in the runtime system: here, the purpose is to take the full image data to extract as much information as possible that could be predictive of therapist annotations. In the run-time system, the purpose is to detect specific expressions relevant to the intervention scripts.

Layer	Output Size
Data	224×224
Convolution 1	$64 \times 112 \times 112$
Pooling 1	$64 \times 56 \times 56$
LRN 1	$64 \times 56 \times 56$
Convolution 2a	$96 \times 56 \times 56$
Convolution 2b	$208 \times 56 \times 56$
Pooling 2a	$64 \times 56 \times 56$
Convolution 2c	$64 \times 56 \times 56$
Concat 2	$272 \times 56 \times 56$
Pooling 2b	$272 \times 28 \times 28$
Convolution 3a	$96 \times 28 \times 28$
Convolution 3b	$208 \times 28 \times 28$
Pooling 3a	$272 \times 28 \times 28$
Convolution 3c	$64 \times 28 \times 28$
Concat 3	$272 \times 28 \times 28$
Pooling 3b	$282 \times 14 \times 14$
Classifier	$8 \times 1 \times 1$

Table 2: This table shows the number of outputs for each particular layer of the network as well as the layer's type. "Convolution" denotes a convolutional layer, "LRN" a local response normalization, "Concat" a merging layer and "Pooling" a pooling layer (Burkert et al., 2015).

4.2 Facial expression classification

First, children's faces were captured from the raw video data by using the Haar cascade classifier (OpenCV library, Viola-Jones object detection algorithm applied for face detection). As there are 3 sources of video records, one source was chosen for each frame. All extracted faces were stored and mapped to each particular video frame (each image name is encoded with the name of the child, name of the task, name of the intervention and the frame id).

A facial expression classifier was built by implementing a neural network architecture proposed by Burkert et al. (2015, see Fig. 1) in Keras[2] – a high-level neural networks API, written in Python and capable of running on top of either TensorFlow[3] or Theano[4]. Table 2 shows output sizes for each particular network's layer.

The Cohn-Kanade AU-Coded Facial Expression Database (Version 2, `CK+`, see Fig. 2 was used as a training set (Kanade et al., 2000; Lucey et al., 2010) for this neural network. This database includes both posed and non-posed (spontaneous) expressions and additional types of metadata[5].

This dataset was preprocessed by cropping and resizing each imager to remove noisy background and fit the model's input data format. Then, it was split into training and validation sets (85% and 15% respectively). After training the model, categorical accuracy reached 91%. However, applying this model to the frames extracted from the recorded interventions did not achieve high accuracy (see below).

[2] `https://keras.io`
[3] `https://www.tensorflow.org/`
[4] `http://www.deeplearning.net/software/theano/`
[5] `http://www.pitt.edu/~emotion/ck-spread.htm`

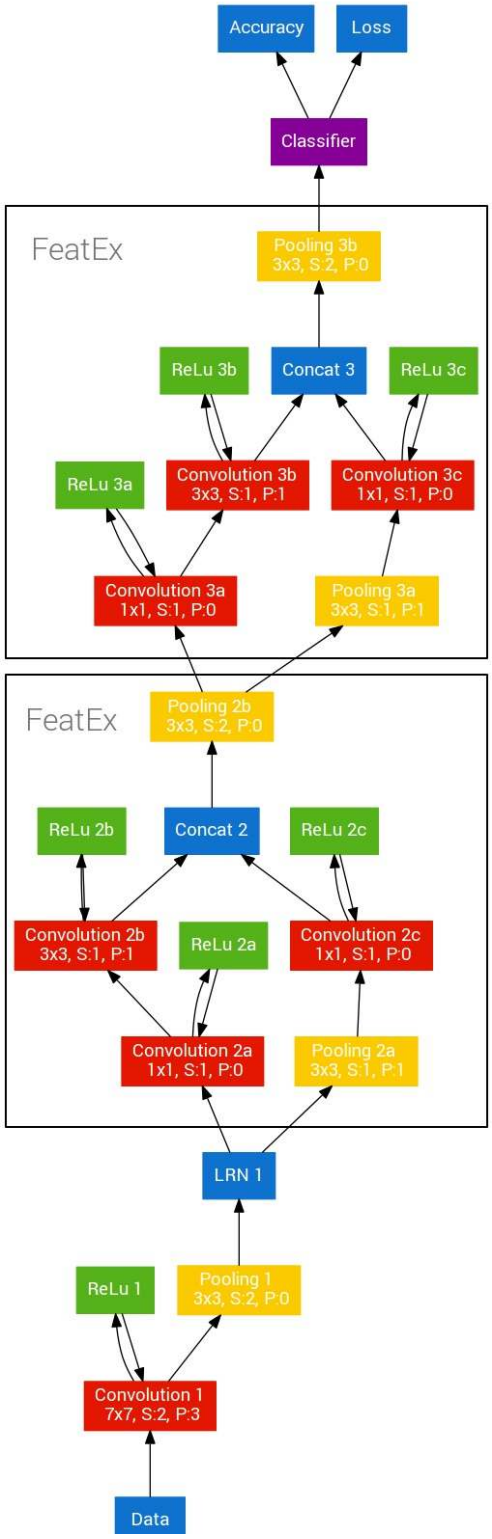

Figure 1: A convolutional neural network architecture proposed by Burkert et al. (2015). It is used for facial expression recognition, adjusted for the project's needs.

Figure 2: Example images taken from the CK+ database that represent *Fear*, *Happiness*, *Surprise* and *Sadness* respectively from left to right. ©Jeffrey Cohn

4.3 Label cleaning

In order to analyse annotations provided by therapists, they should be standardised with appropriate names, tiers' titles and content. In particular, annotations produced earlier in the project but no longer relevant can be deleted.

To achieve this, all tiers were extracted for each particular intervention. A Python script was provided for automatic tiers mining. Then, the list of unique tiers' titles was created. This revealed a need to clean up variations and typos in the labels, the use of both Romanian and English in labelling, and so on. For the present purposes, these problems were resolved manually, but a script is being developed for a more complex annotation algorithm which cleans, renames and reorders tiers and aggregates them into several ELAN files based on the templates given for all the annotations produced during the previous stages of the project.

4.4 Annotation generation

Using the trained facial expression classifier, "positive emotions" annotations were generated from scratch for all of the video records which had corresponding annotations generated manually by a therapist.

To achieve this, images were passed as the input data to the classifier that returned the probabilities for the child's face of expressing one of the following emotions – *Anger*, *Sadness*, *Neutral*, *Happiness*, *Disgust*, *Surprise* and *Fear*. For the "positive emotions" tier, *Happiness* and *Surprise* were taken as markers. If the probability of choosing one of these two classes was high, the frame was marked as containing a positive emotion. The timestamps were then saved to the ELAN file for each particular time period displayed a high density of positive markers.

These annotations were then compared against the ground truth (here defined as the therapist annotations) on typical metrics including precision and recall were applied. Figure 3 shows an example of annotations comparison. The algorithm takes a therapist's annotation and the corresponding automatically generated time line and then computes a confusion matrix for performance measuring. Based on this matrix, the model's precision and recall are calculated. However, current results remain unsatisfactory, highlighting the need for additional work in this respect. Further efforts are ongoing.

5 Summary

Overall, we have tackled diagnostic tools from three perspectives in this deliverable: First, we have provided a discussion of what is currently feasible from a technical perspective in terms of automating the diagnostic process itself. Second, we have briefly described to what degree the current run-time

Figure 3: Example performance for the designed model. The graph shows the time line for an intervention, with black markings for annotated positive emotions, for automatically generated annotation ("Auto") and for one that was created manually by a therapist ("Therapist"). For this particular example, model's precision is equal to 40% (the amount of automatically marked frames which were also annotated by a therapist) and model's recall is 32% (the amount of frames annotated by a therapist which were also marked by the model).

system can be used for diagnostic and evaluation matters. Finally, we have reported on the ongoing efforts to produce tools that can automate the video annotation process.

In terms of the actual scope of this deliverable, the first two aspects (together with accompanying documention in WP2 deliverables as noted) cover the necessary. Automating therapist annotations goes beyond the scope, but is also a rather desirable technology to achieve. Efforts to improve performance in this respect will therefore continue and can be reported again in D5.3 (due M48), whose focus is the documentation of improvements to the assessment methods.

References

Athanasopoulos, G., Verhelst, W., and Sahli, H. (2015). Robust speaker localization for real-world robots. *Computer Speech & Language*, 34(1):129 – 153.

Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(1011):763 – 786. Intrinsic Speech Variations.

Burkert, P., Trier, F., Afzal, M. Z., Dengel, A., and Liwicki, M. (2015). Dexpression: Deep convolutional neural network for expression recognition. CoRR, abs/1509.05371.

Cai, H., Zhou, X., Yu, H., and Liu, H. (2015). Gaze estimation driven solution for interacting children with asd. In *2015 International Symposium on Micro-NanoMechatronics and Human Science (MHS)*, pages 1–6.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Acoustic variability and automatic recognition of childrens speech. *Speech Communication*, 49(1011):847 – 860. Intrinsic Speech Variations.

Han, J., Shao, L., Xu, D., and Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334.

Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA. ACM.

Kanade, T., Cohn, J. F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53.

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

Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324.

Moeslund, T. B., Hilton, A., and Krger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(23):90 – 126. Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour.

Moore, R. K. (2007). Spoken language processing: Piecing together the puzzle. *Speech Communication*, 49(5):418 – 435. Bridging the Gap between Human and Automatic Speech Recognition.

Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1424–1445.

Rabiner, L. R. and Schafer, R. W. (2007). Introduction to digital speech processing. *Foundations and Trends in Signal Processing*, 1(12):1–194.

Ramirez, J., Gorriz, J. M., and Segura, J. C. (2007). Voice activity detection. fundamentals and speech recognition system robustness. In Grimm, M. and Kroschel, K., editors, *Robust Speech Recognition and Understanding*. InTech.

Suarez, J. and Murphy, R. R. (2012). Hand gesture recognition with depth images: A review. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 411–417.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2007). A survey of affect recognition methods: Audio, visual and spontaneous expressions. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, pages 126–133, New York, NY, USA. ACM.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886.