Welcoming Gesture Recognition into Autism Therapy

Carlos Duarte
University of Lisbon
FCUL, Campo Grande
1749-016 Lisboa, Portugal
cad@di.fc.ul.pt

Daniel Costa
University of Lisbon
FCUL, Campo Grande
1749-016 Lisboa, Portugal
dancosta@di.fc.ul.pt

Luis Carriço
University of Lisbon
FCUL, Campo Grande
1749-016 Lisboa, Portugal
lmc@di.fc.ul.pt

André Falcão
University of Lisbon
FCUL, Campo Grande
1749-016 Lisboa, Portugal
afalcao@di.fc.ul.pt

David Costa
University of Lisbon
FCUL, Campo Grande
1749-016 Lisboa, Portugal
dcosta@lasige.di.fc.ul.pt

Luís Tavares
University of Lisbon
FCUL, Campo Grande
1749-016 Lisboa, Portugal
luisferreiratavares@gmail.com

Abstract
Gesture imitation has recognized benefits as a therapy for children with Autism Spectrum Disorder. Even tough automatic gesture recognition has advanced greatly in the last years, its application in the field of autism therapy has been mostly irrelevant. In this paper we present a solution that: 1) integrates gesture imitation into storytelling therapies; 2) is capable to learn new gestures without an explicit learning stage; 3) provides automatic gesture recognition capable of assisting therapists during interventions, and might support social skills practice at home.

Author Keywords
Autism; Gesture recognition; Automatic gesture learning

ACM Classification Keywords
I.4.9 [Image processing and computer vision]: Applications; J.3 [Life and medical sciences]: Health.

General Terms
Algorithms, Design, Human Factors

Introduction
Autism Spectrum Disorder (ASD) covers a set of developmental disabilities that can cause significant social, communication, and behavioral challenges. ASD is
characterized by profound deficits in communication, social functioning, and the presence of repetitive and stereotyped behaviors or interests. Besides possibly experiencing difficulties developing and understanding language skills, children with ASD may have difficulty communicating non-verbally, such as through hand gestures, eye contact, and facial expressions.

Non-verbal communication is particularly relevant to early ASD diagnosis because children use hand gestures to express their desires before they can do so with verbal language. The range of gestures representing social skills includes pointing, showing (arm extending towards other person while holding an object), giving (placing an object in reach of another person), clapping, waving, among others.

Children with ASD exhibit significant impairments both in imitation of gestures as well as in their spontaneous use. Because gestural behavior plays an important role in the establishment and maintenance of social interactions, difficulties in this area might contribute to the difficulties individuals with autism have. For children with ASD, gesture imitation has been found to be related to the development of social interaction [6]. These findings suggest that teaching gesture imitation may improve the child’s social skills and even language development [7] as well as spontaneous gesture use [9]. Thus, from a therapeutic perspective gesture imitation should be an important focus of intervention programs [2].

In this paper we report on our initial efforts into using automatic gesture recognition mechanisms in the scope of autism interventions. This is part of a larger effort aiming at supporting autism therapy through an array of tools focusing on the therapist as well as the child. The work is being conducted together with a team of autism therapists following a participatory design approach. In the following we present previous efforts on teaching gestures to children with ASD; how gesture recognition can be integrated in a wider storytelling based framework; the challenges we faced when including a gesture recognition mechanism in a social stories player that is acceptable for intervention use; and conclude with results from preliminary evaluation and the planned future work.

Teaching gestures to children with ASD
According to Ingersoll et al. [2] there has been “little research focusing on increasing the imitation of meaningful gestures in children with autism”. Consequently, there is also a lack of intervention techniques designed to address this problem [2], even tough several approaches have been attempted: behavioral [3], structured [10] or naturalistic [2].

Children with ASD typically react well to the use of technology [9]. This has been already explored in the context of interventions, with different technologies having already been used: video and animations, virtual reality, mobile devices, robots, interactive tables, etc. We aim to use gesture recognition as a component of a technology assisted intervention with the main objective of teaching imitation of gestures.

Storytelling for children with autism
Storytelling, and social stories in particular, is one of the tools used by autism therapists to improve the social skills of children with ASD [1]. Through storytelling, children with ASD can improve their imaginative abilities, collaborative skills and social competences.

We are currently exploring digital storytelling as a means to train social skills of children with ASD in general. One
intervention that can be applied through storytelling is gesture imitation. Our implementation follows the standard linear structure of a story. The story begins with an introduction, where a video or animation presents a scene where a social skill will be needed (e.g., two friends approaching each other). This is followed by the midpoint, where the gesture that should be made is shown to the child (e.g., a video of a child waving), and the child imitates the gesture. Finally, the conclusion, where, based on the performance of the child, a reinforcement is given (e.g., the friend waving back if the child performed well).

A therapist wishing to teach one particular skill, selects an appropriate story, together with the gesture that needs to be performed during the intervention. During the session, the story will be played back to the child. The therapist controls the transitions between the parts of the story. This allows the therapist to replay any part of the story whenever there is a need for it – for instance, the child was distracted by some other stimulus, or the therapist feels the need to clarify some aspect of the story to the child.

**Gesture recognition and storytelling**

Following the story structure presented above, gesture recognition mechanisms can be incorporated into the second part of the story, to assess how well the child performs the requested social gesture. From the perspective of the gesture recognition mechanism, the task is straightforward. When the therapist selects the story to use, the gesture that the child will execute, and thus the gesture to be recognized, is identified. This means the recognizer will not need to distinguish between gestures. Its task will be to classify how well the gesture was performed (if it is performed at all). While this is a straightforward task with current technology, other challenges had to be tackled in order to make this system useful for autism therapists. These challenges can be split into two classes: distribution of control between system and therapists; and automatic gesture gathering.

**Therapist control of gesture recognition**

While an automatic gesture recognition system could assist therapists by taking some of their workload during an intervention (e.g., the system can compute the time required for the child to perform the correct gesture without the therapist having to tell the system the exact moment the gesture was performed), there are some constraints that prevent therapists from giving up control to the gesture recognition system.

Children with ASD can perform the gestures with certain nuances that make it challenging for the recognition system. For instance, it is not uncommon that the waving goodbye gesture is performed with the child’s palm of the hand facing himself instead of the person he is waving to. If the gesture recognition system is not able to determine which way the palm of the hand is facing, it can incorrectly classify the gesture. This was happening with our first Kinect-based implementation. It has been addressed in a posterior implementation, thus solving this problem. However, we cannot guarantee that similar subtle distinctions will not occur in future gestures.

Furthermore, it was clear from the therapists that how well a gesture was performed is not the single factor in assessing the child’s performance. Consider a child that previously would not raise his arm when requested to wave goodbye. If that child now raises the arm, even without waving it, his performance is considered very positive, even if the gesture was not correctly executed.

Taking the two previous points into consideration, we have designed the system in a way that: 1) keeps track of
gesture performance as classified by the therapist and the
gesture recognition mechanism; 2) the therapist’s
decisions override the gesture recognition decisions,
allowing the therapist the possibility to continue the story
even when the child did not perform a gesture that has
been correctly classified by the recognition mechanism.

Collecting gestures automatically
The biggest challenge when deploying a system with
automatic gesture recognition as part of a storytelling
front-end is to build and make available a library of
gestures to use in social stories. It is clearly not possible
to record in advance all the gestures that could be used.
The therapists do not have the skills or the availability to
record the gestures by themselves. That means that we
need a mechanism to create a database of gestures,
requiring the least intervention from therapists, in order to
have a usable solution.

Considering our usage scenarios, the data we have
available is twofold. First, we can collect streams of
children performing gestures. These streams can be
collected during the interventions. We know for each
stream what is the gesture the child is expected to
perform (because each stream is associated with a story
that exercises a specific gesture). Second, for each stream
we have a therapist provided assessment of how well the
child executed the gesture (here we ask for the gesture’s
classification, not the child’s performance classification, as
discussed in the previous section).

Consequently, we have available several streams of
different children performing the same gesture. The
streams have different durations and gestures performed
at different quality levels. While we know each stream has
the gesture in it, we do not know when it occurs inside
the stream. We have devised a solution allowing us to
compare the streams in order to find the segments that
are common and to be able to identify which segments
conform to the gesture being executed. Afterwards we
extract these segments in order to build a library of
gestures that can be used for recognition purposes.

This stream matching solution is composed of three steps:
stream normalization, comparison and segmentation.

Stream normalization
The collected streams are represented as arrays of frames.
Each frame contains the data relative to all skeleton
points tracked by a Microsoft Kinect. Because each
stream can potentially be from a different child, and
different children will have different physical
characteristics (height, limb length) and might be
standing at different distances from the capturing device,
we need to start by normalizing the skeleton coordinates.
This will ensure our solution to be robust to variances in
translation, scale and human proportions.

Our solution is based on the one proposed in [8]. Firstly
we scale the captured skeleton frames to a predefined
human height. Thus, we always compare skeletons with
the same size. Because the sensor has a wide visual span,
the skeletons can be on very different positions. To solve
the global translation issue we take the center hip as a
reference point and set it as the origin.

Stream comparison
In this step we compare two streams in order to trim the
frames that do not contain the gesture. We start by
checking for inactivity in the beginning and ending of the
sequence, erasing the segments of the stream where only
residual movement is detected for all points of the
skeleton. The resulting streams include the gesture we
want to compare, probably together with other random
movements. To identify the segments of both streams that contain identical movements (which should represent the desired gesture) we used a solution based on the Needleman Wunsch algorithm [4].

The Needleman Wunsch algorithm, also referred as the optimal-match algorithm, is commonly used in bioinformatics to align protein or nucleotide sequences, but the general goal is similar to what we are looking for. While the original algorithm compares nitrogenous bases (four discrete values), we needed to compare (continuous) skeleton points. Our comparison is based on the Euclidean distance between the joints of the normalized skeletons in the frames from the two streams. Two joints are said to match when their distance is below a threshold, and differ otherwise.

Stream segmentation
The traceback step of the original algorithm led to poor results, failing in aligning sequences that were far from each other in their streams (e.g. in one stream the gesture was executed near the beginning and in the other, near the end). We replaced the traceback step with a different approach. We look for the maximum value in the substitution matrix, and consider that to be the end point of the gesture. To find the starting point we diagonally trace back (examine the cell located one row and one column before the current cell) until one of two conditions is met: the value of the cell with smaller index is larger than the value of the cell with bigger index, or we reach the first column or row of the matrix. Using this algorithm we are able to identify the first and last frames corresponding to the gesture we are looking for in the stream.

To use the gesture for recognition purposes we need to store these frames. If we haven’t stored a set of frames for this gesture before, we store the frames from the stream that has a better classification from the therapist. If we have one set already, we replace it if the new stream has a better classification from the therapist and the number of frames that resulted from the comparison is larger than 30 frames. We have this minimum number of frames to make sure we don’t store sets to small that might represent irrelevant gestures.

Preliminary evaluation
We have conducted an initial evaluation of the gesture collecting and recognition mechanisms. We collected 4 different gestures from 6 different people. The participants had different physical characteristics, with height ranging from 158 cm to 183 cm and age from 22 years old to 27 years old. All gestures were performed while standing up, but the participants could move any body part and perform whatever movements they wished besides the requested gesture. We then compared and segmented streams containing the same gesture. In this evaluation, we considered all gestures to have equal classification. On average, the segmented stream reaches the minimum number of frames after 3 comparisons. We stopped comparing streams for a gesture whenever they reached this number of frames. We then confirmed visually (by playing back the set of frames stored) that the frames represent the correct gesture. This was true for all captured gestures. We then used the stored frames to feed the recognizer (we used the KinectDTW recognizer) and asked 12 different participants to perform the gesture to validate it would be recognized. Each participant performed 48 gestures (24 correct gestures and 24 inexistent gestures). The system achieved a precision of 86.9% and a recall of 83.3%. When using the recognizer’s own gesture recorder the values were: 74.2% for precision and 95.8% for recall.
Conclusions and future work

In this paper we showed how gesture recognition can be integrated within a storytelling framework that aims at improving the social skills of children with ASD through imitation of gestures. We presented a mechanism for learning gestures from streams of skeleton points collected during interventions. Our mechanism improved the recognizer’s precision but at a cost of a lower recall. We argue that this effect is beneficial for therapy sessions, because it lowers the possibility of an incorrect gesture being recognized as a correct one, even if the child needs to be more precise when performing the gesture. This will foster adoption by therapists and allow them to use the system to teach new gestures. Still, we will continue working on improving the recall rate.

We are now planning on deploying the working system in an autism therapy institution. Following, we plan to deploy it as part of a therapeutic framework where social skills can be exercised at home under the guidance of an automated gesture recognition system that controls stories created by a therapist.

Acknowledgements

This work was supported by FCT through funding of SiTECOSE project, ref. PTDC/EIA-EIA/117058/2010, and LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014

References