Project: Semiautomatic video analysis of music therapy sessions with children

Final report, January 2017

Contributors: Alla Sapronova, Miles Granger, Karin Mössler, Klaus Johannsen, & Christian Gold

Contributions: AS wrote the draft report; MG conducted the analyses; KM provided critical comments; KJ acquired funding; CG acquired funding and helped to revise the report.

1.) Objectives

Primary objective: The primary objective of the project was to develop a semiautomatic video analysis tool for therapy sessions to support the assessment and evaluation of therapies.

Secondary objectives:

1. Development of computer vision methods to identify and track the motion of two individuals (one therapist and one client), with application to up to 180 clients (up to 10 30-minute segments per client)

2. Development of machine learning algorithms to correlate the motion with the participant's severity of illness. The outcome of the project will be a proof of concept for a fully automatic video based motion analysis system to support video assessment and evaluation of therapies.

2.) Methodology

2.1) Computer vision methods for identification and motion tracking of two individuals.

In 2016 the video analysis tool to automatically detect two person on the video was developed, tested and is ready to be employed.

The automatic video analysis tool consists of a person detection module and an algorithm for track motion of two individuals as described in project's secondary objective 1.

The person detection module is based on state-of-the-art "You Only Look Once" (YOLO) deep neural network. YOLO is a fast real-time object detection system that frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. Bounding boxes and class probabilities are predicted directly from full images in one evaluation by applying a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region. The bounding boxes are later weighted by the predicted probabilities.

Full description to the YOLO approach can be found at http://pjreddie.com/darknet/yolo/ or "You Only Look Once: Unified, Real-Time Object Detection" by Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi.

To avoid duplicate detection, a machine learning method (k-means) was used in addition to the YOLO system. k-means clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

In this work the k-means algorithm fits to the detected bounding boxes' content and predicts if there is the same person on it based on what the k-means algorithm has, based upon last 20 frames. It was found that last 20 or so frames should be similar enough to distinguish between two persons.

After k-means algorithms detects the person, the list of people is further processed to look for false positives indicators.

The person detection module receives the bounding box's (from YOLO) coordinates and size, confidence rate and number of frames processed. It returns True if two people were confirmed detected or False otherwise. This identifies only frames which both (two) people are detected. The detected frames with two people are stored for analysis and other frames are not recorded. An example of the module's return from the first 20 video segments is shown in Table 1.

#	Video ID	Total_Frames_for_analysis	Total_seconds_for_ analysis
0	oc038_sqD	243	10.13
1	oc038_sqF	308	12.83
2	oc053_sqG	5070	211.25
3	oc053_sqH	415	17.29
4	oc054_sqA	3163	131.79
5	oc054_sqB	2980	124.17
6	oc057_sqC	5222	217.58
7	oc057_sqH	4880	203.33
8	oc058_sqB	3258	135.75
9	oc058_sqI	5094	212.25
10	oc062_sqD	5276	219.83
11	oc062_sqK	5270	219.58
12	oc067_sqF	612	25.5
13	oc067_sqG	408	17
14	oc070_sqB	3733	155.54
15	oc070_sqK	926	38.58

Table 1. Example of the module's return from first 20 video segments.

2.2) Development of machine learning algorithms to correlate the motion with the participant's severity of illness

The development of machine learning algorithm to correlate the motion with the patient's severity of illness has started but was not possible to complete within this project. The model for PCA analysis of the patient's motion is developed and tested. The relationship between the severity of illness and the patient's motion has to be further explored with closer collaboration between therapists and data analysts.

The analysis of the patient's motion includes the following:

- Pre-processing the selected data
- Assigning machine learning process to sort detected individuals
- Apply principal components analysis (PCA) to find the moments of interaction between two persons that are likely to be important for the therapists to see.

2.2.1. Pre-processing, including data cleaning and normalization.

The image detection in the video results in dirty data (periods of no detection, duplicates, etc.) that have to be cleared prior to video analysis. The algorithm locates and corrects any remaining mis-categorizations that can be due to double detection, camera moves, long periods without any detection, etc.

To make sure that all the measurements are comparable across videos, the recorded data have to be normalized: the data are scaled down so that relative maxima are at most 1 and minima are 0. Finally, the duplicates are detected and ejected.

2.2.2. Distinguishing between detected individuals.

To detect the selected individual as person#1 or person#2, the "who's who" machine learning / logical process is applied. Each video is processed with an unsupervised method that clusters and scores the detected attributes (as described below).

At first the database for each video fragment is populated. The database contains three major groups of descriptions: video-related information, person 1 related information, person 2 related information; and contains the following attributes:

video ID, process_time, absolute_distance_from_the_reference_point, anchor_point_angle, frame_count, frames_past_no_detection;

person_1_anchor_distance, person_1_angle, person_1_area, person_1_center_x, person_1_center_y, person_1_coords, person_1_length, person_1_image, person_1_width;

person_2_anchor_distance, person_2_angle, person_2_area, person_2_center_x, person_2_center_y, person_2_coords, person_2_length, person_2_image, person_2_width.

The attributes are used to perform PCA and assign the person's ID (1 or 2 here) to the image. To do that, the algorithm determines if two images in the same row are likely to be an image of the same person by checking whether their similarity score is higher than the threshold. The detections is done using Jaccard similarity score.

At the final step, the position of the detected individual is recalculated as necessary (in case of camera move, etc): to do this, a fuzzy hashing is used and then a difference between hashes of images on line N against images on line N+1 is calculated. If hashing distinction falls below the threshold of certainty, the instance is sent back to unsupervised classification using PCA.

2.2.3. Detecting moments of interactions between individuals.

The core task of this work was to detect the potentially “interesting” interactions between the individuals. As there is no strict definition of what kind of interaction can be considered as “interesting” from the therapist point of view, the following was suggested:

- Cast the video(s) into time-segments of length N and look for 'interesting moments' by
    - a.) applying time-series Convolutional Neural Networks (CNN) with Long Short Term Memory Neurons on raw images
    - b.) applying various Recurrent Neural Network models without CNN
- Cast video(s) into time-segments of length N and use ensemble model to predict regression based scores. For that, video should be split into N segments and each segment 'voted' and then all votes counted/averaged for final score

All the options would require denoted videos, identifying points/times in the video of interest and therefore will require more time and closer collaboration between data analysis and therapists.

As at the moment no denoted video is available, the test case was performed on the entire video with no target. An Unsupervised Neural Network Restricted Boltzmann Machine was used to identify segments. The sampled video was processed and the following information was returned: all possible interactions detection, the time lengths of the interaction, and the likelihood of the interaction to be “interesting” from the therapist point of view (based on the assumption that the “interesting” interaction is resulting in the movement frequency of both people to be identical for some period of time). From a sample video, a total of nine interactions with relatively high scores were detected, and the timestamps showing these interactions on the video were denoted.

To develop an algorithm that would correlate the motion with the participant’s severity of illness, further work will be required to:

- validate the denoted timestamps by the therapists

- denote at least one video of a good quality and a few “interesting” interactions between individuals to teach the model for automatic interactions detection.

3.) Results and discussion

The development of computer vision methods to identify and track the motion of two individuals (one therapist and one client), with application to up to 180 clients (up to 10 30-minute segments per client) as defined in secondary objective P1 was implemented. The video processing is done on the world's second most powerful GPU with 2500+ CUDA cores. From that, one can process hundreds of videos in parallel with our Hadoop computing cluster with the specialized model. Upon benchmark testing, the system is capable of responding to ~172k requests per second, more than enough to respond to requests specified in P1.

The development of machine learning algorithms to correlate the motion with the participant’s severity of illness as defined in secondary objective P2 was not implemented fully:

- The automatic detection of interaction of two persons was fully developed and implemented. The detected interactions (referred also as «moments») were marked as references to the video timestamps.

- The identification of the severity of illness to support the assessment and evaluation of therapies was not implemented.

Initially, it was proposed that the assessment of therapies can be done by detecting and counting the «interesting» moments of the interaction between a patient and a therapist on the video. By «interesting» moments here we refer to therapeutically significant moments of interaction between the child and the therapist. Unfortunately, there is no strict or mathematical definition of what the «interesting» moment could look like, and the identification of such moments shall be done either through unsupervised learning (if a sufficiently big data set is available) or though an explicit «moments» evaluation (labeling).

The video data-set provided with only excerpts of sessions (as illustrated in Table 1) could not be used for unsupervised classification because of the video quality. After pre-processing, it occurred that one video segment contains only a few seconds of usable video material for the analysis. Obviously, such length of a video often comes with no significant information: 0 or 1 interaction. The employment of unsupervised machine learning methods is only possible on a statistically significant amount of data, that is much more than 1, of course. Since the unsupervised learning technique cannot be directly employed here, it is necessary to label at least 70% of all detected interactions (either as «positive» (i.e «interesting interaction») or «negative») manually. Those labels shall be used to teach the model in a supervised mode to differentiate between detected interactions. The labeling job can only be done with a help of a therapist(s).

Employment of additional variables provided (e.g. age, sex, site, etc) did not help to differentiate between the interactions in terms of «interesting» or «not interesting». However, these variables could be useful later to predict session successes and to evaluate the therapies in an overall perspective.

4) Conclusions The system we developed allows fast video processing. The analytical part of the system includes tracing of people and detection of interactions between two people. The developed model can be employed in two ways:

1. the distance between the child and the therapist can be calculated at (almost) any time during the video and returned to a therapist as a table (distance vs time). Further work needs to be done to see if the variation in distance has a correlation with the successes of a therapy or severity of illness. Additional variables such as those provided (sex, age of a patient) can also be used here.

2. The system can provide the therapist with an array of timestamps (marks) of the detected interactions. The system cannot detect whether those interactions are «interesting» for the therapist or not, neither is the system capable of evaluating a therapy session, but the therapist can save time by watching only few minutes of a video at the marked

moments.

If more funding for a new project becomes available, with the help of a therapist it will be possible to create a fully automated system for the detection of «interesting» moments.

The question whether it is possible to create a predictive model that can forecast the effectiveness of a therapy for one particular patient would require more time, as more experiments with various types of machine learning algorithms would need to be done.