Automatic classification of emotion-related states in video datasets of human-human interaction

**Background**

The goal of the thesis project is to accurately and consistently classify emotion-related states inferred by analyzing video datasets of human-human interactions using various machine learning and image analysis techniques. The available dataset consists of a number of videos depicting infant-adult and infant-adult-object interactions annotated with regards to the emotional affect and engagement of the children towards the interaction target. The dataset has been provided by the Uppsala Barn & Baby Lab division of the Psychology Department. In turn, the result of the analysis and the classification algorithm could later on be used, for example, in helping artificial intelligent agents better understand and interact with humans.

As prerequisites, courses in Artificial Intelligence, Image Analysis and Intelligent Interactive Systems have been undertaken and passed by the student. The work will be done partially at Uppsala University at the Social Robotics Lab of the IT Department and partially at the student’s residence.

Regarding previous work in the same area, there have been other studies conducted in emotion classification in children-robot interaction [1] and in children-mother interaction [2].

**Description of the task**

With an increase in computing power and development of cutting-edge machine learning and image analysis techniques, the scientific study of human-human behavioral interaction is a highly researched topic in the recently emerging field of Social Robotics. The aim of this project is to provide a good approach on the inference of emotional affect from video data and to gather relevant results in analyzing infant behavior. The main benefits for this kind of emotional affect recognition would be in the algorithms used by educational robot assistants that interact frequently with children.

The steps in analyzing a video clip would consist in employing techniques from Computer Assisted Image Analysis such as image segmentation of the area of interest (in this case the agents of the interaction), motion features recognition and extraction (such as extraction of skeleton data), and ending with a classification of the emotional charge of the interaction using various Machine Learning algorithms. There are many machine learning algorithms that suit the task at hand, so a comparison between accuracy and efficiency of the chosen algorithm against various other implementations will also be made.

**Methods**

Classification

Feature Extraction

Image Segmentation

Neural Network

The main data manipulation method would consist in employing various Python scripts using different specialized modules for computer vision and machine learning. Some potential approaches for image segmentation are HSV-based segmentation or motion-based segmentation. Feature extraction could be used with techniques described in [3] and [4] such as part-based models or kinematic models. The machine learning part would be handled by a neural network architecture, including deep convolutional networks and potentially recurrent networks [5][6]. Finally, the classification would be complete which will in turn allow for a rigorous analysis of the results.

**Relevant courses**

1TD396 – Computer Assisted Image Analysis I

1MD032 – Intelligent Interactive Systems

1DT012 – IT and Society

**Delimitations**

The thesis does not aim to provide a state-of-the-art method of either feature extraction or emotional affect inference from video feeds, but rather study various methods of doing the above mentioned and comparing accuracies and feasibility. The project is not meant to go over the integration of the methods in real-life robots. The study of the time and resource efficiency of various algorithms could be considered if time allows it.

**Time plan**

February W7: Literature study; identification of methods and previous research done in the area along with an analysis and identification of the most appropriate image detection, feature extraction and machine learning algorithms to use for the task at hand.

February W8 – March W10: An implementation of image segmentation of the relevant areas in the video feed that depict the agents and a superficial approach on the body motion feature detection.

March W11 – W13: Refining and optimizing the body motion features identification and extraction until satisfactory.

April W14 – W17: Easter holiday planned, may interfere with the schedule. Start of the machine learning training using various algorithms, recording accuracy and other metrics relevant for comparison.

May W18 – W19: Finalizing the neural network training, compiling the best techniques used and results obtained.

May W20 – W22: An in-depth comparison of the algorithms using various graphical tools for a better visualization of the results.

June W23 – W24: Finishing retouches to the overall project state. Estimated end of the project.

The writing of the report thesis that would describe and record the methodology and the results of the project would be happening alongside the whole duration of the project.

**References**

[1] Sanghvi, Jyotirmay & Castellano, Ginevra & Leite, Iolanda & Pereira, André & Mcowan, Peter & Paiva, Ana. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction. 305-312. 10.1145/1957656.1957781.

[2] Egmose Ida, Varni Giovanna, Cordes Katharina, Smith-Nielsen Johanne, Væver Mette S., Køppe Simo, Cohen David, Chetouani Mohamed (2017). Relations between Automatically Extracted Motion Features and the Quality of Mother-Infant Interactions at 4 and 13 Months. Frontiers in Psychology Volume 8. DOI 10.3389/fpsyg.2017.02178, ISSN 1664-1078.

[3] Noroozi, Fatemeh and Corneanu, Ciprian Adrian and Kaminska, Dorota and Sapinski, Tomasz and Escalera, Sergio and Anbarjafari, Gholamreza. (2018). Survey on Emotional Body Gesture Recognition. arXiv e-prints. arXiv:1801.07481

[4] Filntisis, Panagiotis P. and Efthymiou, Niki and Koutras, Petros and Potamianos, Gerasimos and Maragos, Petros (2019). Fusing Body Posture with Facial Expressions for Joint Recognition of Affect in Child-Robot Interaction. arXiv e-prints. arXiv:1901.01805

[5] Yan, Sijie and Xiong, Yuanjun and Lin, Dahua (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. arXiv e-prints. arXiv:1801.07455

[6] Li, Chuankun and Hou, Yonghong and Wang, Pichao and Li, Wanqing (2017). Joint Distance Maps Based Action Recognition With Convolutional Neural Networks. IEEE Signal Processing Letters. 624-628. DOI 10.1109/LSP.2017.2678539. arXiv: 1704.07595