Using a **generative deep learning model** for automatic **facial gesture and head movement generation** for robots and virtual agents based on YouTube videos

Learning Non-verbal Behavior for a Social Robot from YouTube Videos

Patrik Jonell, Taras Kucherenko, Erik Ekstedt, Jonas Beskow

INTRODUCTION

- Non-verbal behavior is crucial for positive perception of humanoid robots.
- Most of existing work on modeling non-verbal behavior show limited variability due to the fact that the models employed are deterministic
- **We propose a novel method for generation of a limited set of facial expressions and head movements, based on a probabilistic generative deep learning architecture called Glow [1].**

METHOD

- We have implemented a workflow which takes videos directly from YouTube, extracts relevant features, and trains a model (based on Glow [1]) that generates gestures that can be realized in a robot without any post processing.
- Two user studies were conducted on Amazon mechanical turk, one using a Furhat robot and one using a virtual agent in order to evaulate the perceived appropriateness and coherency of the proposed system vs three other conditions.

GLOW

Glow [1] is a flow-based generative model introduced by Kingma and Dhariwal. The main idea of a flow-based model is to use normalizing flows [2] in order to learn a transformation from the standard Gaussian distribution to the distribution of the data. This is done by applying a sequence of invertible transformation to the gaussian and then learning parameters of these transformations from the dataset by optimizing the likelihood of the data according to the model.

MODEL

The model was based on Glow [1], but modified to produce sequences of facial parameter vectors conditioned on audio features. The model was trained to generate a fixed-length output of 160 frames conditioned on speech spectra.

RESULTS

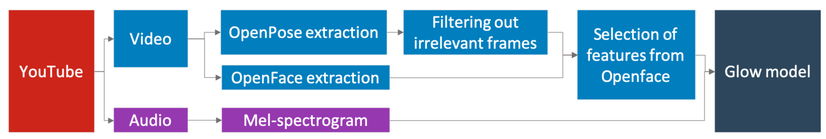
- Having no facial expressions or head movement is significantly worse than having some facial expression
- The proposed method is perceived as good as (for Study 1 and Study 2 in coherence) or even better (for Study 2 in appropriateness) compared to "*Ground Truth*" and "*Random Alignment*"
- That there was no significant difference between "*Random Alignment*" and "*Ground Truth*"

DISCUSSION

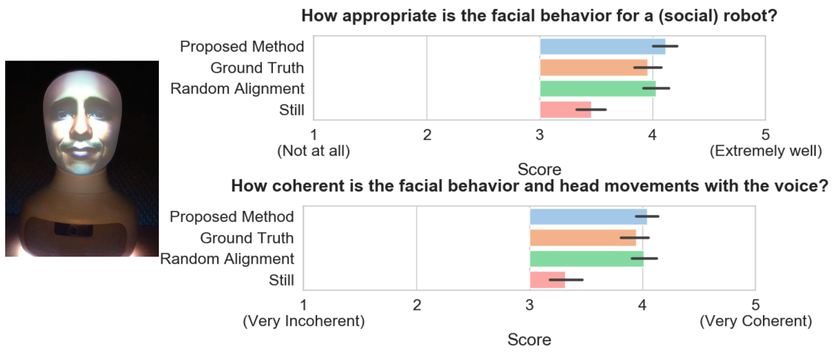
- That there was no significant difference between "Random Alignment" and "Ground Truth" seems to suggest that evaluators were not influenced by the timing of head motion and facial expression or that the realization of the facial gestures was poor (the robot did for example smooth out the motion).
- That the proposed method was perceived as slightly better than the "Ground Truth" and "Random alignment" can be due to noise introduced by not fitting the model entirely to the data, and this added noise being perceived in a positive way.

CODE

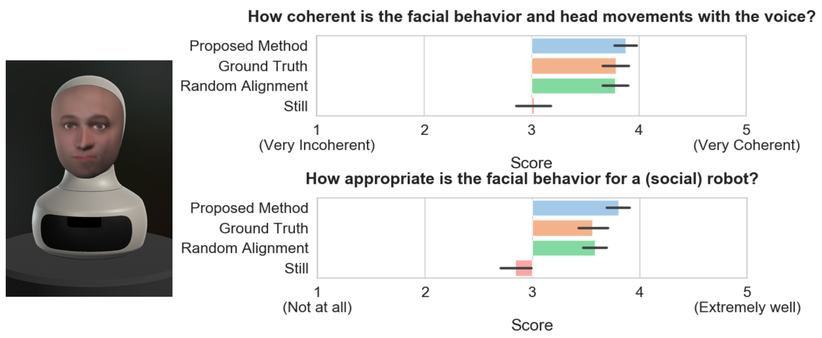
github.com/jonepatr/glow-non-verbal-robot-behavior



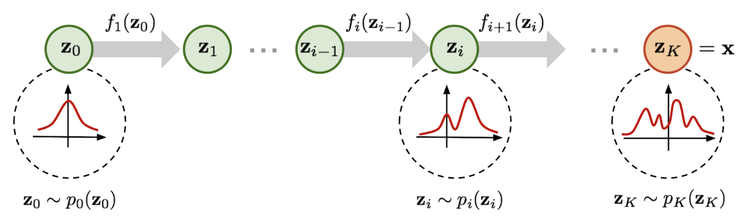
Implemented workflow for feature extraction



Results from Study 1 where the facial expressions were realized in a Furhat robot



Results from Study 2 where the facial expressions were realized in a virtual agent (the Furhat robot simulator)



The "steps of flow" where a standard gaussian dirstribution is being transformed in a sequence of steps to the distribution of the data

source: https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

References

[1] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in Advances in Neural Information Processing Systems, 2018, pp. 10 215–10 224.

[2] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," In Proceedings of The 32nd International Conference on Machine Learning, 2015, pp. 1530–1538.

Read the full paper

KTH VETENSKAP OCH KONST