Direct On-Line Imitation of Human Faces with Hierarchical ART Networks

Patrick Holthaus[1] and Sven Wachsmuth[2]

***Abstract*— This work-in-progress paper presents an on-line system for robotic heads capable of mimicking humans. The marker-less method solely depends on the interactant's face as an input and does not use a set of basic emotions and is thus capable of displaying a large variety of facial expressions. A preliminary evaluation assigns solid performance with potential for improvement.**

I. Mimicry in Social Robotics

Mimicry of facial expressions (c.f. [1] for a comprehensive overview) is an automatic process that has been reported to occur at around five times a minute in human beings [2]. It is believed to facilitate face-to-face interaction between interlocutors and has positive effects for both the mimicker and the one being imitated [3]. For example, the affiliation between two interaction partners is being increased with the help of mimicry [4].

To have robots engage in an interaction to a higher degree, and thus be socially more intelligent, it is desirable to give them the capability of mimicking human facial expressions. With such a system, a robot would be able to express a certain amount of empathy towards a human.

A system enabling the robot to mimic a human has to work in an on-line fashion, and be robust plus flexible. Most existing mimicry-systems for robots today classify a fixed set of facial expressions (e.g. six basic emotions, [5]) and then generate the corresponding pre-modeled posture or a combination of them which is then displayed on the robot. [6] however, use a regression to learn direct motor mappings from facial images. This way, arbitrary postures can be displayed on the robot without the need to classify facial expressions beforehand. Also, the method does not need markers in the interactant's face or a feature point extraction because the raw image is used as an input for the regression. Therefore, one does not need to prepare an interaction partner in any way for the imitation to work.

II. A Robotic Imitation System

In [6], various methods for learning the relation between input image and posture have been shown to have comparable performances. Their results however are solely based on an off-line evaluation with pre-recorded set of images from the same camera. Our goal in this work has been to improve the proposed method and establish a system that has the same flexibility and accuracy as before but being implemented as an on-line system on a robot that runs in real-time and can be continuously improved by further learning.

Fig. 1. A robotic head Flobi (simulated) imitates the facial expression of a human in front of a web cam. The green rectangle on the left marks the input image determined by a face detector. Motor values (perc.) for each actuator are given at the respective joints.

Our system is specifically trained for the robotic head "Flobi" [7] that has 15 degrees of freedom for facial expressions but can in principle be extended to other robots that have similar expressive capabilities. It relies on a standard face detection algorithm [8] which cuts out the face from the camera input image. The gray-scale image of the facial region is then histogram-normalized to floating point values between $[0..1]$ and scaled to a 64x64 image. For the regression, parts with no information about the facial expression, mostly the lower corners of the image and nose region, have been masked out, leaving a 2614-dimensional input vector remaining.

For the prediction of the robots posture, a regression on top of a topology-learning ART (Adaptive Resonance Theory) neuronal network [9] is used. It uses the image vector directly to predict a 15-dimensional posture where each of the dimensions describes the position of one actuator in the range of $[0..1]$. This method is well-suited because it runs in real-time, and can be trained on-line. Please refer to Fig. 1 for an example of the imitation. The input is depicted on the left side, as well as the resulting posture on the right side.

III. Training Procedure

Although the prediction can in principle be learned completely on-line, best results are achieved when training models beforehand. Existing models can then be adjusted during runtime. For learning the relation between a human facial expression and a posture, 27 people have been placed in front of the robot with the task to imitate the robot's face. Nine of the participants were female, some had facial hair or glasses. Altogether 30 facial expressions per person

[1] P. Holthaus is with the Applied Informatics at the Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany patrick.holthaus at uni-bielefeld.de

[2] S. Wachsmuth is with the Central Labs Facilities of the Center of Excellence Cognitive Interaction Technology, Bielefeld University, 33615 Bielefeld, Germany sven.wachsmuth at uni-bielefeld.de

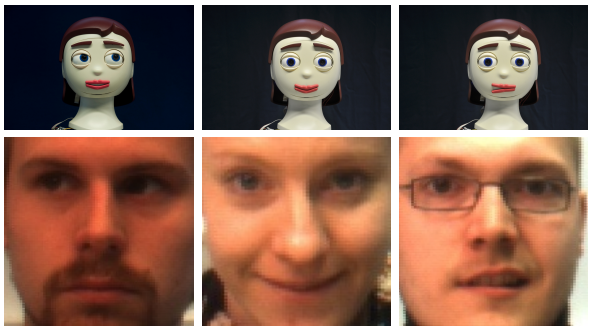

Fig. 2. Example training images for three of the participants. People try to approximate the robot's facial expression which is then recorded by the experimenter. Each image is transformed multiple times to match different lighting conditions.

have been recorded and stored in conjunction with the corresponding robot posture.

To be more robust against potential changes in lighting or recording properties when changing location, face detector, or camera, the training mass has been enlarged by scaling and translating the image (+/-1 pixel), and altering its contrast, saturation, gamma, brightness, and hue by approximately +/- 1-2 percent. A total of 8748 images per person and posture has been used in the training procedure. See Fig. 2 for an example of training pairs.

IV. Preliminary Evaluation

First results show that, although our predicted posture has more degrees of freedom, the prediction functions at least as good as reported in [6]. In the best performing configuration, the network has been trained with parameters $\phi = 0$, $\beta = 0.0$, $\tau = 200$, and $\rho = 0.9$. The approximate reconstruction error lies at 0.04. However, if the error function disregards actuators that should not differ from the default value, i.e. only moving joints are considered, the average error lies at around 0.18. Contrary, if only actuators are measured that are supposed to predict the default value, the error is only at 0.02. Fig. 3 illustrates the results.

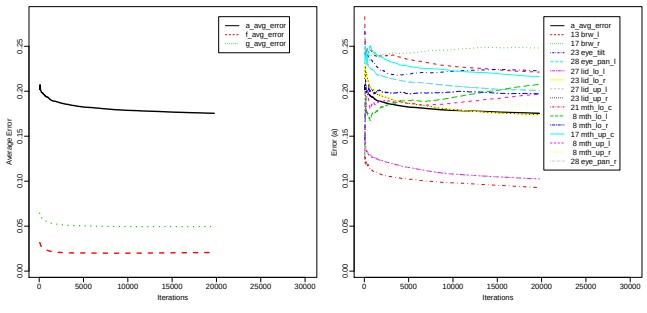

Fig. 3. Evolution of prediction errors by train iteration. On the left, the average error is splitted by overall error (dotted), and actuators that should predict the default value (dashed) and those that are supposed to deviate from default (solid). The latter is broken down by individual actuator on the right side.

The results indicate that the chosen training and prediction methods perform well enough to be established as an on-line imitation system for a robotic head. They are confirmed by first tests with other types of cameras and lighting conditions, where the prediction did not lose much of its precision. The relatively high error for alternating actuators and fairly low error for default predictions suggest that a posture close to default is predicted in a high number of cases. Training thus could still be improved, for example by acquiring a greater number of training images from different people or incorporating other input sources.

V. Conclusion

In this work, we have presented a system that is able to mimic facial expressions of a human interlocutor. It uses a regression on a topological ART network to map the input image of a human face to the posture of a robotic head. Our method runs in real time and can be improved by on-line learning. The system's performance is comparable to other off-line methods which is approved by a preliminary evaluation. Our approach also has the potential to be improved, for example by including additional training material.

VI. Acknowledgments

This work has been supported by the German Research Foundation (DFG) within the Collaborative Research Center 673, Alignment in Communication. We also greatly acknowledge the support of student assistant Marian Pohling in the technical realization of this work.

References

[1] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction." *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.76.6.893

[2] N. Chovil, "Social determinants of facial displays," *Journal of Nonverbal Behavior*, vol. 15, no. 3, 1991. [Online]. Available: http://link.springer.com/article/10.1007/BF01672216

[3] M. Stel and R. Vonk, "Mimicry in social interaction: benefits for mimickers, mimickees, and their interaction." *British Journal of Psychology*, vol. 101, no. Pt 2, pp. 311–23, May 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19646328

[4] J. Lakin and V. Jefferis, "The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 145–162, 2003. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1025389814290

[5] P. Ekman, "Basic emotions," *Handbook of Cognition and Emotion*, vol. 98, pp. 45–60, 1999.

[6] M. Tscherepanow, M. Hillebrand, F. Hegel, B. Wrede, and F. Kummert, "Direct Imitation of Human Facial Expressions by a User-Interface Robot," in *International Conference on Humanoid Robots*, Paris, France, 2009, pp. 154–160.

[7] I. Lütkebohle, F. Hegel, S. Schulz, M. Hackel, B. Wrede, S. Wachsmuth, and G. Sagerer, "The Bielefeld Anthropomorphic Robot Head "Flobi"," in *International Conference on Robotics and Automation*, IEEE. Anchorage, Alaska: IEEE, 2010, pp. 3384–3391.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[9] M. Tscherepanow, "An Extended TopoART Network for the Stable On-Line Learning of Regression Functions," in *International Conference on Neural Information Processing (ICONIP)*, B.-L. Lu, L. Zhang, and J. Kwok, Eds., vol. 7063, Springer. Shanghai, China: Springer, 2011, pp. 562–571. [Online]. Available: http://www.springerlink.com/content/g46154308108150t/