Learning to Make Facial Expressions

Tingfan Wu, Nicholas J. Butko, Paul Ruvulo, Marian S. Bartlett, Javier R. Movellan
Machine Perception Laboratory, UC San Diego
La Jolla, CA 92093
{ting,nick,paul,marni,movellan}@mplab.ucsd.edu

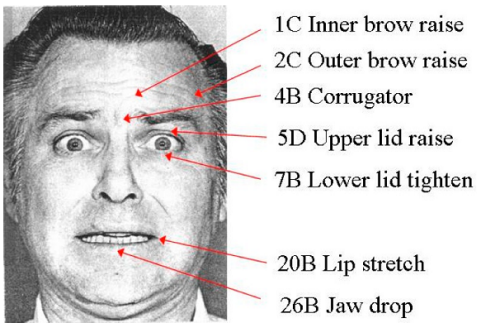
***Abstract*—This paper explores the process of self-guided learning of realistic facial expression production by a robotic head with 31 degrees of freedom. Facial motor parameters were learned using feedback from real-time facial expression recognition from video. The experiments show that these control properties can be learned through an active exploration and feedback loop. The mapping of servos to expressions was learned in under one-hour of training time. We discuss how our work may help illuminate the computational study of how infants learn to make facial expressions.**

I. Introduction

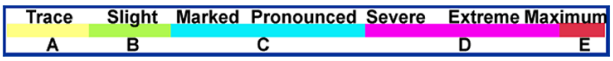
The human face is a very complex system, with more than 44 muscles whose activation can be combined in non trivial ways to produce thousands of different facial expressions. As android heads approximate the level of complexity of the human face, scientists and engineers face a difficult control problem, not unlike the problem faced by infants: how to send messages to the different muscles so as to produce interpretable expressions.

Others have explored the possibility of robots learning to control their bodies through exploration. Olsson, Nehaniv, and Polani [1] proposed a method to learn robot body configurations using vision and touch sensory feedback during random limbs movements. The algorithm worked well on the AIBO robots. However, the AIBO has only 20 degrees of freedom and is subject to well known rigid body physics. The Einstein robot employed here, on the other hand, has 31 degrees of freedom and the mapping between servo actuators and facial expressions is not trivial. In practice setting up the robot expressions requires many hours of trial-and error work from people with high level of expertise. In addition as time progresses some servos may fail or work differently thus requiring constant recalibration of the expressions.

In developmental psychology, it is believed that infants learn to control their body through systematic exploratory movements [2]. For example, they babble to learn to speak and wave their arms in what appear to be a random manner as they learn to control their body and reach for objects. This process may involve temporal contingency feedback during social interactions with the caretaker [3] or other individuals around. In our work, we apply this same idea to the problem of a robot learning to make realistic facial expressions. The robot engages in a series of body-babbling episodes in which the robot actuates its servos in a sequence of random configurations. Feedback as to which servo configurations give rise to which facial expressions is determined using video capture of Einstein's face coupled with an automated expression recognition system. While the feedback in this case is visual, this type of feedback can also be interpreted as a type of proprioception not unlike the "organ relations" hypothesized by Meltzoff and Moore [4].



(a) The FACS Action Units (AUs)



(b) The A-E facial intensities defined in FACS.

Fig. 1: A face can be FACS-coded into a set of numbered AUs (each number is a facial muscle group) along with letter-grades denoting intensity.

II. Methods

A. Robotic Head

The robot head "Einstein", was developed by Hanson Robotics. The face skin is made of a material called Frubber, that deforms in a skin-like manner contributing to the realism of the robot expressions. The head is actuated by 31 servo motors, 27 of them controlling the expressions of the face and 4 controlling the neck (See Figure **??**). While robot is able to simulate the actions of all major muscle groups in the face and neck there are some important difference in the way the human muscles and the robot servo motors actuate the face. In contrast to human muscles, these servos can both pull and push loads and thus each motor can potentially simulate the action of 2 individually controlled muscle groups. Orbicular muscles, like the *Orbicularis oculi* and the *Orbicularis oris* produce circular contractions whereas the robot servos produce linear contractions that are coupled via circular tendons.

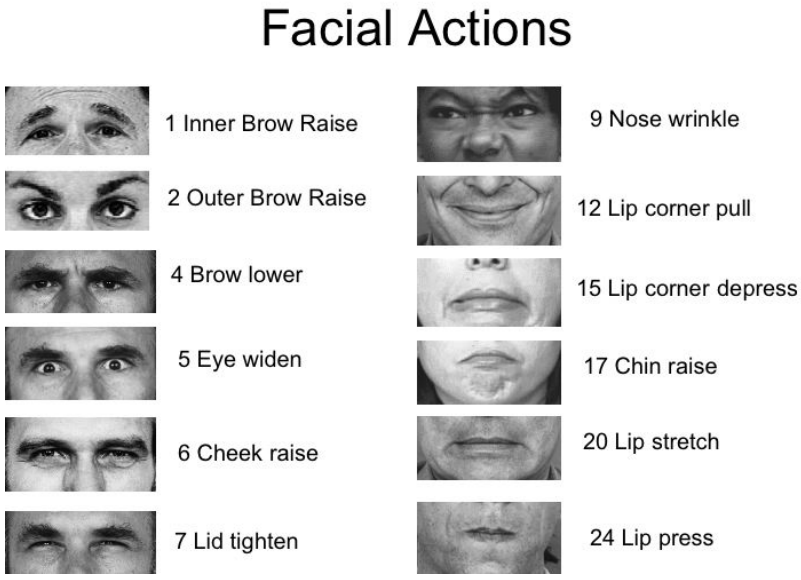


Fig. 2: A close-up of actions units defined in FACS.

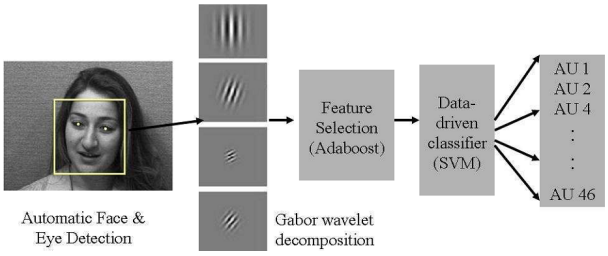


Fig. 3: Software framework of computer expression recognition toolkit

B. Facial Action Coding System

As the basic unit of facial expressions in our work we use the Facial Action Coding System (FACS) developed by Paul Ekman [5]. FACS is a comprehensive language to describe facial expressions in terms of atomic muscle movements, named facial action units (AUs). Figure 2 shows some major AUs. Given a face image along with the neutral face of the same person, a certified FACS coder can code the face (Figure 1a) with a set of activating AUs along with their intensity measured in 5 discrete levels (Figure 1b) base on the appearance change on the face.

C. Automated Methods for Facial Action Coding

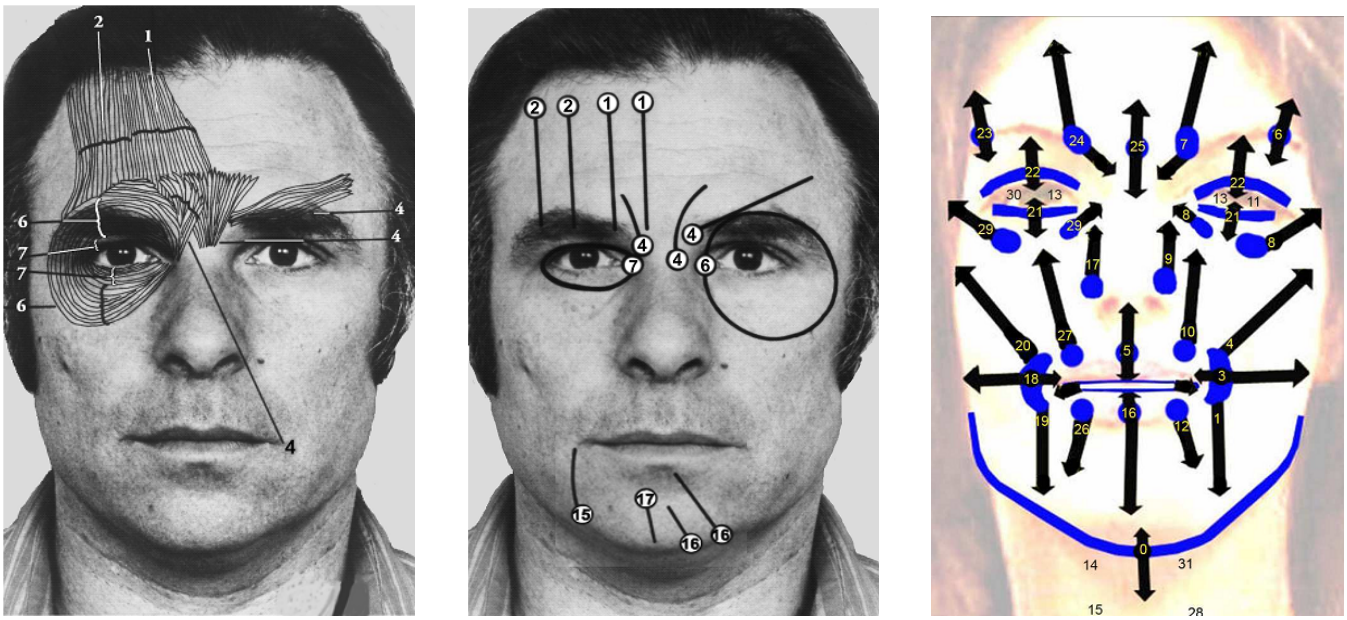
In recent years the computer vision community has made significant progress on the problem of automating FACS coding from video. Cohn's group at CMU [6] developed a system based on the use of active appearance model that tracks 65 fiducial points on the face. AUs are recognized based on the relative position of the tracked points. Our group at UCSD has been pursuing an alternative approach, called CERT (see Figure 3), to directly recognize expressions from appearance-based features rather than from relative locations of fiducial points [7]. First the region of the face is automatically segmented. The obtained image patch is then passed through a bank of Gabor filters that decompose it into different spatial frequencies and orientations. Feature selection methods, like Adaboost, are used to select the more relevant filters. Finally, support vector machines (SVM) are used to classify the existence of AUs given the extracted features.

In this paper, we use CERT as a way to simulate a proprioceptive system to a complex android head (Einstein). As the robot moved the head, CERT provided feedback about which AUs were active. AUs approximately correspond to individual face muscles, thus practically providing a proprioceptive (though visually guided) system to the robot.

D. Learning: Random Movements and Feedback

The expression recognition software, CERT, can be seen as a non-linear function F that takes a given image I and then outputs a vector $F(I) \in \mathcal{R}^m$ of detected intensities of m AUs. Let $\mathcal{S}$ be the number of servos used in the experiment. We denote j-th random configuration encountered during motor babbling as $\mathbf{s}_j \in \mathcal{R}^{|\mathcal{S}|}$, and the corresponding face images as $I_{\mathbf{s}_j}$. Further, let n denote the number of random movements collected.

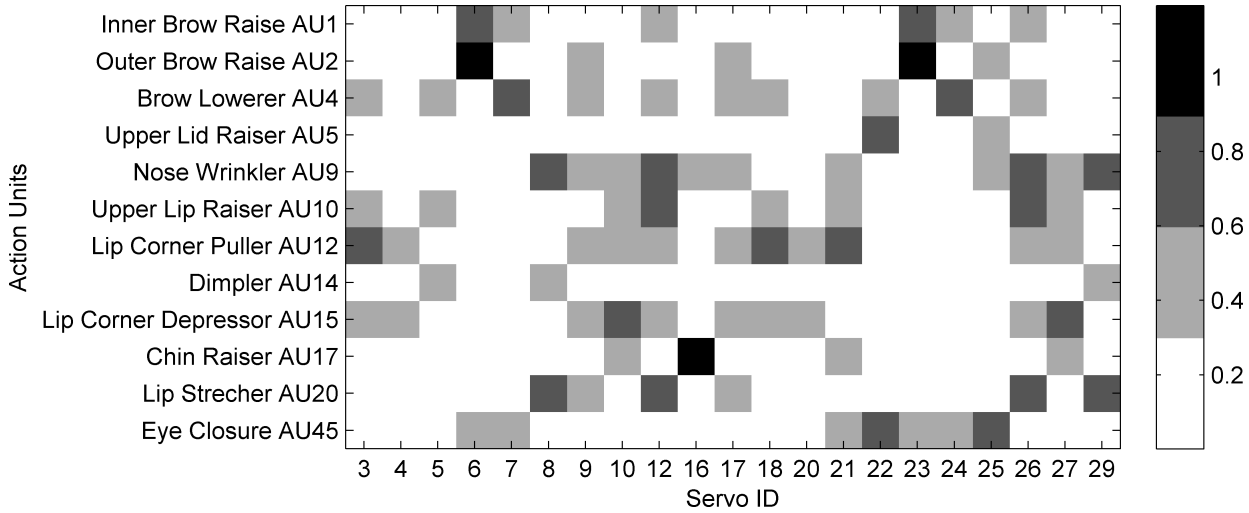
In order to produce a given expression, Einstein must learn have a mapping from a desired AU configuration of his face to activations of his servos. This mapping is known more generally as inverse-kinematics. In this document we choose to model this relationship linearly. For the i-th servo we train a linear regression model to minimize the following objective



(a) (Upper face) muscular anatomy

(b) Facial action units associated with facial muscles

(c) Servo layout on the robotic face



(d) The relation between servos and AUs

Fig. 4: A comparison between human (a) facial muscles, (b) FACS AUs , and (c) robotic servo layouts on Einstein. and (d) the learned connections between the AUs and servos learned in our experiment.

function:

$$\min_{\mathbf{c}_i, b_i} \sum_{j=1}^{n} ||(F(I_{\mathbf{s}_j})^T \mathbf{c}_i + b_i) - (\mathbf{s}_j)_i||^2, \tag{1}$$

where b_i is the bias term in regression. The learned model parameter $\mathbf{c}_i \in \mathcal{R}^{m+1}$ provides a linear mapping from desired AU to corresponding configuration of i-th servo.

Linear models are simple, fast, and require no extra parameter tuning. The obvious disadvantage is that if underlying mapping between servo actuations and expressions is not linear, the model will not work well. It was thus unclear to us before performing the experiment how well the proposed approach would work.

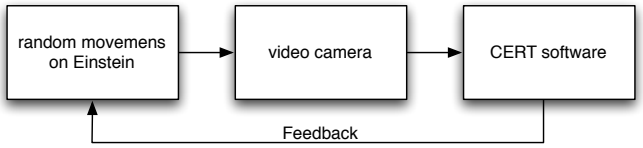
E. Demonstration: Facial Expression Synthesis

As soon as the model parameters are learned, we can use the model to generate servo movements $\{\mathbf{s}_i\}, i \in \mathcal{S}$ for target AU configurations $\mathbf{a}$ with following servo position.

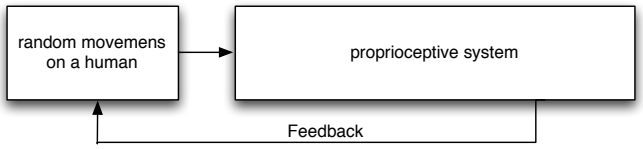
$$\mathbf{s}_i = \mathbf{a}^T \mathbf{c}_i + b_i. \tag{2}$$

III. Experiments

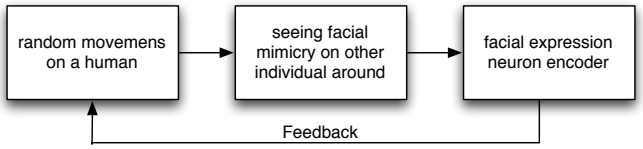
The real-time expression recognition was done using CERT version 4.4.1 running on a Dual Core Intel Based Mac Mini. The CERT software recognizes 12 AUs (see Figure 4d for a list). The output of CERT is a real-valued vector for each video frame indicating the estimated magnitude of each facial



(a) Einstein



(b) human



(c) human

Fig. 5: The proposed framwork of learning to demonstrate facial expression on Einstein(a) and human(b)(c).

Fig. 6: Asymmetric random facial movements.

action. The output is roughly base-lined at zero, with outputs above zero indicating the AU was present. However, the actual baseline of neutral expression is subject dependent. Therefore, we collect the baseline for Einstein $\mathbf{a}_N$, which will be used in expression synthesis stage.

Communication with the ROBOT hardware was handled using RUBIOS2.0 a Java based open source communications API for Social Robots [8]. RUBIOS 2.0 is built on top of QuickServer, an open source Java library for multi-threaded, multi-client TCP server applications.

A. *Learning*

In order to collect data for learning a mapping between facial expressions and servo movements, Einstein initially generated a series of random servo movements (See Figure 5a) uniformly and independently for each servo within the range

TABLE I: correlation coefficient how well AU input predicts servo movements.

face region	training	testing
upper	0.7868	0.7237
lower	0.5657	0.4968

of safe operation for each servo. While the robot is capable of moving each of its degrees of freedom independently, we chose to impose our prior knowledge that the upper and lower eyelids should move in a coupled fashion. We enforced this by sampling a random activation for the upper eyelid and then setting the activation for the lower eye-lid accordingly.

We excluded the servos for directing the eye gaze (3 servos), the jaw (1 servo), and the neck (4 servos) since they were not related to the elementary facial muscle movements currently recognized by CERT. Two additional servos, 1 and 19, were also disabled after discovering that, when random motor babbling caused pulling in opposition to servos 4 and 20, servo burnout resulted (Figure 4c). We are currently developing a mechanism for the robot to automatically sense the energy spent by the servos and therefore to automatically avoid harmful servo configurations, possibly by adding a fatigue term that simulates the limited capacity of human facial muscles to contract for long periods of time. Such a change might also lead to more realistic learned strategies for facial expression synthesis.

We collected 500 instances of perception-production pairs. Each instance consists of the configuration of the servos and the outputs of the 12 facial action unit detectors produced by CERT. Since CERT estimates activations of individual facial muscles, here CERT could be seen as playing the role of a human proprioceptive system, informing which facial muscles are activated at every point in time. The 500 instances were then used to train a linear regression model. The results are shown in Table I. We observed a very good performance for expressions in the upper face region and moderate performance for the lower face. We suspect that this may be due to the facial hair in the robot (mustache) reducing the accuracy of the feedback provided by CERT. However the problem may have been due to the underlying mapping between servos and expressions, which may have been more non-linear for expressions in the lower face. We are currently investigating which of these two explanations is more consistent with the data.

Figure 4d displayed the mapping between AU and servo control signals learned by the model. The values are normalized by the dynamic range of AU intensity and servo movements. In each row, the figure shows the set of servos related to the generation the AU, with dark shading indicating strong involvement. For example, servo 6 and 23 plays the major roles in demonstrating AU2, while servos 9, 17 and 25 also provides minor contribution. On the other hand, each column shows which AUs predict or explain the servo movement the best. For example, the movement of servo 6 is mainly explained by AU17 (chin raise, Figure 2).

B. Action Unit Synthesis

Coding of human facial action units best done in relation to a neutral face. Here we face a similar issue in that we have to account for the Einstein's neutral expression and use it as baseline to synthesize other action unit configurations. Let the baseline AU intensities of Einstein's neutral face be denoted by $\mathbf{a}_N = F(I_N)$ where I_N is the neutral expression face of Einstein when all the servos are relaxed. Then, the synthesized AU intensities were set to $\mathbf{a}' = \mathbf{a}_N + \mathbf{e}_i$, where $\mathbf{e}_i$ is a vector of zeros with the exception that the i-th element to be one. Finally, we generated the corresponding servo movements by $\mathbf{s}'_i = \mathbf{a}'^T \mathbf{c}_i + b_i$.

Figure 7 shows examples of some of the synthesized AUs. We put the neutral expression in (a) for reference. (b) is the synthesized AU1 expression (inner eyebrow raise). For comparison, we also put the neutral and AU1 expression demonstrated by a human in (c) and (d). Figure (e)-(h) gives more examples on AU2, AU4, AU5 and AU9.

IV. Discussion

While Hanson Robotics made an effort to explicitly map each servo to individual action units in the Facial Action Coding System, we observed that the model learned to activate multiple servos to produce each AU. Subjectively the AUs learned by the model, which synthesized multiple servos, appeared more realistic than the equivalent AUs originally shipped with the robot that had been set by hand. For example, AU4 (eyebrow narrowing) is recognizable by changes in appearance that occur mainly at the midpoint between the two eyebrows. In humans the muscles that contribute to the appearance of AU4 are the left corrugator and right corrugator. If servos are tuned by hand, a heuristic assignment will be moving inner eyebrow servos 7 and 24. Our model learned this obvious connection clearly. However, as stated in FACS manual [9], the appearance change of AU4 "push the eye cover fold downwards and may narrow the eye aperture." Our model also learned to close the upper eyelid a bit to narrow the eye aperture. Similar phenomena were also found in the lower face. AU 17 "chin raise" is recognizable by the bulging around the chin region (see Figure 2). While the robot does not have any servos in that region of the face, the model learned to produce the appearance of bulging using 3 lip servos (servos 10,16 27).

During the experiment, one of the servos burned out due to misconfiguration. We therefore ran the experiment without that servo. We discovered that the model learned to automatically compensate for the missing servo by activating a combination of nearby servos.

Another interesting observation is that the robot learned to produce symmetric servo movements. This is likely due to the fact that the database of images of facial expressions that was used to develop the CERT software had predominately symmetric expressions.

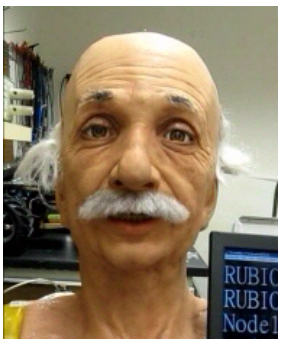
A. Developmental Implications

The primary goal of this work was to solve an engineering problem: How to approximate the appearance of human facial muscle movements with the available motors. Nevertheless this work also speaks to learning and development of facial expressions in humans. It is not fully understood how humans develop control of their facial muscles to to produce the complex repertoire of facial expressions used in daily social interaction. Some aspects of facial behavior appears to be learned, and other aspects appear to be innate. For example, cross-cultural data [10] suggests that some basic expressions, such as smiles, are shared universally among all the peoples in the world, leading scientists to hypothesize that they are innate. Moreover, congenitally blind individuals show similar expressions of basic emotions in the appropriate contexts, despite never having seem them [**?**], and even show brow raises to emphasize speech [**?**]. In contrast, aspects of facial expression that appear to be learned include symbolic gestures and display rules such as reducing the shown expression by contracting other muscles. The frequency of certain facial movements was higher in congenitally blind than sighted children, and only the sighted children masked their negative emotions [11]. Facial expressions appear to be refined in development through feedback from others' responses to one's facial expressions.

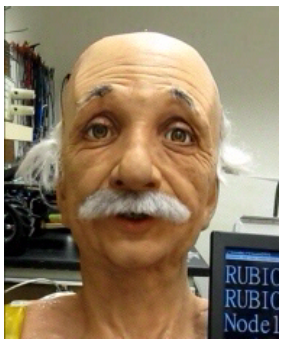
Our experiment demonstrates that complex facial expressions may be learned through feedback of the type made available by CERT. Two possibilities are shown in Figure 5b and c. One possibility is that CERT was basically serving the role of a proprioceptive system. As such the fact that CERT happens to use visual input is incidental. Similar feedback to that produced by CERT could have been obtained using proprioceptive sensors rather than visual sensors. Another possibility is that people can actually encode the expressions observed by others in a manner that mimics the function of CERT. There is empirical evidence that during social interaction people tend to mimic the facial expressions of their interlocutors [12], which implies that humans have the capability to visually encode facial expressions and map them onto their own muscle movements. This behavior could effectively serve as a mirror that would provide information about the effects of one's own muscle movements onto the external appearance of facial expressions. The fact that one could learn the mapping between muscle activations and expressions with less than 500 examples, provides some computational plausibility to the social mirroring hypothesis.

We are currently experimenting with an active learning mechanism to allow the robot to actively choose muscle movements, "facial babbling," so as to optimize learning efficiency. Instead of making random movements, the brain may move the face in more efficient ways to quickly reduce the uncertainly of the internal expression-to-muscle model. Such active exploration may employ information maximization similar to models of human exploratory behavior in eye-movements [13].

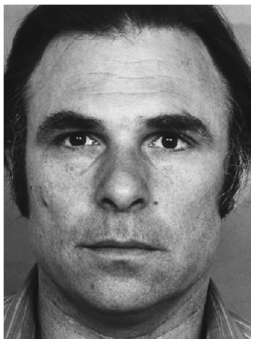
In additional to learning atomic expressions as defined in FACS, we are also currently investigating the mechanisms for learning holistic expressions of emotion, such as expressions of happiness, sadness, anger, surprise, and disgust.



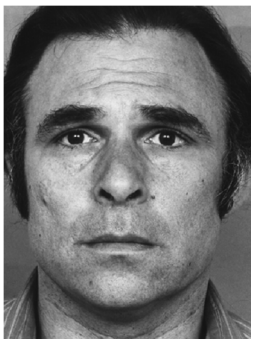
(a) Neutral



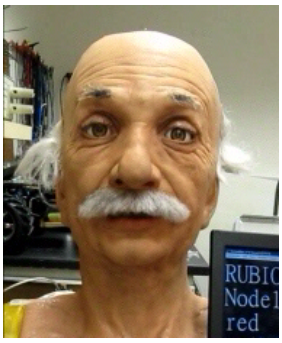
(b) AU1: Inner Brow Raise



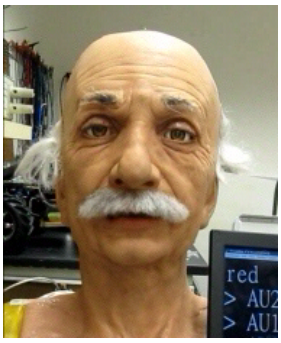
(c) Human Neutral



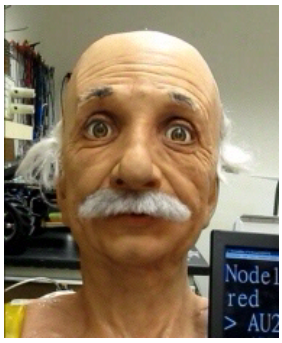
(d) Human AU1: Inner Brow Raise



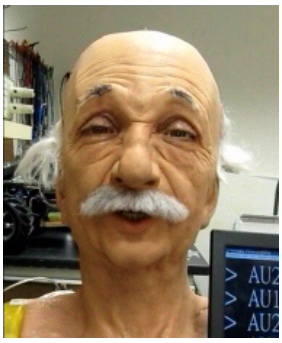
(e) AU2: Outer Brow Raise



(f) AU4: Brow Lower



(g) AU5: Eye Widen



(h) AU9: Nose Wrinkle

Fig. 7: Action units learned by Einstein

V. Acknowledgments

Support for this work was provided by NSF grants SBE-0542013 and NSF IIS INT2-Large 0808767. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] L. Olsson, C. Nehaniv, and D. Polani, "From unknown sensors and actuators to actions grounded in sensorimotor perceptions," *Connection Science*, vol. 18, no. 2, pp. 121–144, 2006.

[2] P. Rochat, "Self-perception and action in infancy," *Experimental Brain Research*, vol. 123, no. 1, pp. 102–109, 1998.

[3] D. Messinger, M. Mahoor, S. Cadavid, S. Chow, and J. Cohn, "Early Interactive Emotional Development," in *7th IEEE International Conference on Development and Learning, 2008.*, 2008, pp. 232–237.

[4] A. N. Meltzoff and M. K. Moore, "Explaining facial imitation: a theoretical model," *Early Development and Parenting*, vol. 6, pp. 179–192, 1997.

[5] P. Ekman and W. Friesen, "Facial Action Coding System (FACS): A technique for the measurement of facial action," *Palo Alto, CA: Consulting*, 1978.

[6] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, pp. 97–115, 2001.

[7] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[8] M. P. Laboratory, "RUBIOS," http://mplab.ucsd.edu/?page_id=392, 2009.

[9] P. Ekman, W. Friesen, and J. Hager, "Facial Action Coding System (FACS): Manual and Investigator's Guide," *A Human Face, Salt Lake City, UT*, 2002.

[10] P. Ekman, "The argument and evidence about universals in facial expressions of emotion," *Handbook of Social Psychophysiology*, vol. 58, pp. 342–353, 1989.

[11] D. Galati, B. Sini, S. Schmidt, and C. Tinti, "Spontaneous facial expressions in congenitally blind and sighted children aged 8-11," *Journal of Visual Impairment and Blindness*, vol. 97, no. 7, pp. 418–28, 2003.

[12] U. Dimberg and M. Thunberg, "Rapid facial reactions to emotional facial expressions," *Scandinavian Journal of Psychology*, vol. 39, no. 1, pp. 39–45, 1998.

[13] N. Butko and J. Movellan, "I-POMDP: An Infomax Model of Eye Movement," in *7th IEEE International Conference on Development and Learning, 2008.*, 2008, pp. 139–144.