Crowdsourced Facial Expression Mapping Using a 3D Avatar

Crystal Butler
New York University
New York, NY 10003, USA
crystal.butler@nyu.edu

Lakshmi Subramanian
New York University
New York, NY 10003, USA
lakshmi@cs.nyu.edu

Stephanie Michalowicz
New York University
New York, NY 10003, USA
sam676@nyu.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892535

Abstract
Facial expression mapping is the process of attributing signal values to a particular set of muscle activations in the face. This paper proposes the development of a broad lexicon of quantifiable, reproducible facial expressions with known signal values using an expressive 3D model and crowdsourced labeling data. Traditionally, coding muscle movements in the face is a time-consuming manual process performed by specialists. Identifying the communicative content of an expression generally requires generating large sets of posed photographs, with identifying labels chosen from a circumscribed list. Consequently, the widely accepted collection of configurations with known meanings is limited to six basic expressions of emotion. Our approach defines mappings from parameterized facial expressions displayed by a 3D avatar to their semantic representations. By collecting large, free-response label sets from naïve raters and using natural language processing techniques, we converge on a semantic centroid, or single label quickly and with low overhead.

Author Keywords
Facial expressions; avatars; crowdsourcing; 3D facial modeling; FACS; expression recognition

ACM Classification Keywords
H.1.2 User/Machine Systems: Human information processing; H.5.m User Interfaces: Miscellaneous; I.3.7 Three-Dimensional Graphics and Realism: Virtual reality

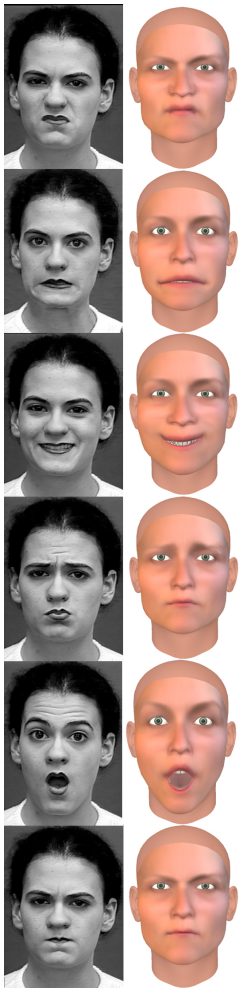

Figure 1: The six basic emotions from the Cohn-Kanade database [8], matched to our avatar: disgust, fear, happiness, sadness, surprise, and anger.

Introduction

When people interact, up to 65% of the communication that occurs is nonverbal [2]. The face is singularly expressive. Facial expressions can convey felt emotion, intent, attitudes, cognitive states, or social signals. Systems that provide automatic expression recognition over a wide range of inputs have several important real world applications. Computer-based agents capable of automatically recognizing pain, confusion, and alertness, for example, could provide first-line monitoring information for human caregivers. Facial behavior can also be used as an indicator of mental health, or in the case of physical trauma such as stroke, the degree of functional impairment.

Facial expression modeling could play a significant role in the diagnosis and treatment of a variety of psychiatric and physical disorders: facial emotion recognition (FER) variability has been found to correlate with psychiatric disorders such as schizophrenia, autism and depression [9, 13]. Virtual humans that accurately model a nuanced range of expressions can be applied to interactive systems for diagnostics and expression recognition training for humans. They can augment other service sectors as well, such as nursing, teaching and customer care, and may increase comfort levels of their human clients [7].

Toward a Facial Expression Lexicon

Progress in these domains has been hindered by the limitations of labor-intensive traditional approaches to researching the signal value of facial expressions. In current systems, automated expression recognition requires mapping a novel face to a particular set of topological changes. Features are learned from sets of thousands of training images painstakingly annotated by human experts to indicate which muscles are involved [8]. In conjunction, studies validate the perceived communicative value of expressions by having naive observers label the images, usually with words from an a priori set of emotion names.

To date, this approach has been largely limited to investigating six or seven basic expressions of emotion that are arguably considered universal (not culturally bound): happiness, sadness, anger, surprise, fear, disgust, and, optionally, contempt (figure 1). The generally accepted standard for mapping muscle activations to expressions is the Facial Action Coding System, or FACS [5]. In FACS, facial movements are 'coded' – that is, annotated by human experts – using discrete craniofacial muscle movements called Action Units (AUs). The mapping of AUs to perceived communication values is not the intent of FACS and is left to researchers.

The method described here, in contrast, uses a 3D digital model (the avatar) with known muscle activation parameters as the basis for expression generation and labeling. The model employs FACS-based deformations on a 0-1 scale that can be manipulated independently. Any weighted combination of facial movements can be generated without the need for trained actors as in traditional FACS-based representations.

Our methodology employs crowdsourced, free-response expression labeling of images of our 3D avatar. By using free-response labeling, we avoid the bias that can be introduced by limiting labels to a predefined subset of descriptors. In addition, we are able to quickly gather large datasets and apply natural language processing (NLP) algorithms to the elements. Our NLP

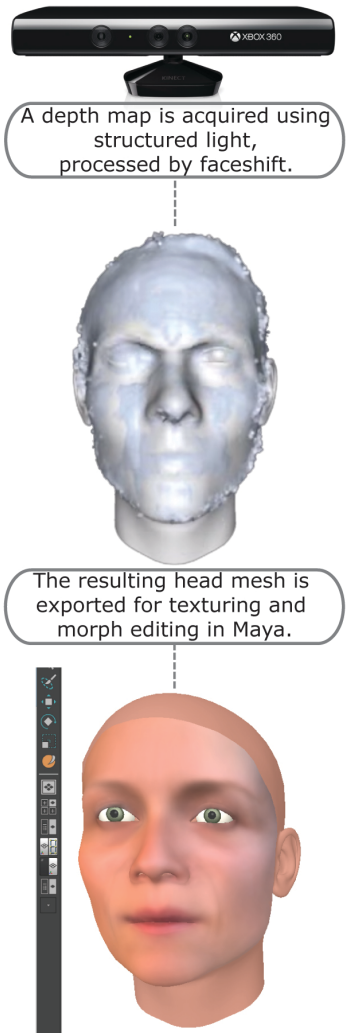

Figure 2: The avatar creation pipeline.

processing determines synonymy ratings between labels, and mathematical analysis of those results shows whether a "best" label exists if most of the elements of a label set share similar meanings. Images with positive results can be said to have high ecological validity, making for better real-world applications. This process can be used to build a broad lexicon of expressions with associated muscle activation vectors.

Related Work

Hybrid approaches using digital avatars for facial expression modeling and testing have been tried previously. Notably, [15] developed a highly realistic, FACS-based 3D morphable model capable of synthesizing arbitrary combinations and weightings of AUs, as does our system. Researchers generated the model by capturing high-resolution scans of several FACS-certified actors as they performed each AU, then amalgamated the results to produce a single linear combination per action. Tests showed that untrained observers reliably identified the six basic emotions in a forced-choice design. However, no research into expanding the repertoire of recognizable facial behaviors was done.

FACSGen, a high-quality FACS-based modeling tool [12], does not rely on capturing human performance or geometry, and is natively digital. Facial texture is photographed and applied to the avatar. As in the previously described system, FACSGen has been validated to show that avatar-driven expressions elicit the same responses as human faces displaying equivalent facial movements. No testing of expressions beyond the set of basic emotions has been performed.

The D3DFACS Database comprises a set of 519 AU sequences, captured by scanning human actors with a setup requiring six cameras [3]. The dataset can be used as the basis for building a 3D morphable model, but only a framework for doing so is outlined. A FACS professional coded the kinetic peaks of each recorded sequence, but the results were not validated against known expression configurations.

Researchers at Ohio State University recently published a study in which they categorize compound expressions of emotion, e.g. fearfully surprised [4]. Including the six basic emotions, their results define 21 discrete expressions. Three of the descriptors used are of the single-word label type in our study, namely appalled, hatred, and awe. While this experiment adds strength to our argument that many discriminable expressions remain to be identified, their method required the use of human participants to model the expressions, and FACS coders to identify the component movements. This technique does not scale for the production and testing of large image sets.

Avatar Design

The avatar was designed to produce believable, quantifiable FACS-based expressions. Initial production of the polygonal head mesh, or 3D geometry, and facial actions for the trial model were created using faceshift software [14] in conjunction with the Kinect, a depth-sensing camera. The resulting mesh and its associated predefined facial morph targets were exported as an fbx file for editing in Autodesk's Maya (figure 2).

Faceshift provides a reasonably comprehensive set of morph targets, based loosely on FACS. Morphs are a set of deformations of the model's base polygonal

AU	Weight
1	0.7
L2	0.6
2	0.6
4	1.5
5	0.7
6	0.7
7	1.0
9	0.8
10	1.0
11	1.0
12	0.7
L14	1.0
14	1.0
15	0.7
17	0.8
18	0.8
20	0.7
23	1.0
24	1.0
25/26	0.6
28	0.7
43	0.3

Table 1: Action Units used in the trial study, with their levels of activation from 0-1. AU4 was overweighted to make it more distinguishable. AUs 25 and 26 are combined to portray an open mouth. The "L" preceding an AU indicates a unilateral activation on the left side of the face.

geometry, used to create smooth state transitions for animation. Modifications and additions to the morph set were required to approximate the physiognomic changes described for core FACS actions. The model was also skinned and textured using a low-resolution RGB image captured with a common webcam.

After modification, the avatar could perform 28 FACS movements, both unilaterally and bilaterally where applicable. Mouth, jaw, eye, and head movements were also integrated, but only mouth and jaw kinematics were used in this study. Morph transforms are applied linearly to the neutral base head shape on a scale from 0-1, with 0 indicating no modification and 1 representing the maximum change from baseline.

Pilot Study

In order to make our initial study manageable, we put some constraints on the number of AUs per generated expression and displayed them at constant weights (table 1). We identified 22 key movements using the FACS Investigator's Guide and the FACS certification test. Weights were fixed at levels that rendered the AUs plausible and clearly visible. Three AUs were combined per expression, which is the most frequently occurring count found for expressions in the FACS Guide and test (excluding head and eye movements, which we did not incorporate). For all combinations of 22 choose 3, that narrowed our list of expression candidates to 1540. A program written in Maya's native scripting language, MEL, generated the expression images. Manual elimination of images where self-collisions occurred in the model geometry brought the count down to 1384.

Phase One

Twenty-nine workers on Amazon's Mechanical Turk (AMT) were presented with a unique subset of 231 3x2 image grids, along with a neutral face for reference. They were instructed to select the radio button next to any image that they believed represented a naturalistic human facial expression and paid $.10 per grid. Three workers responded per grid, providing three yes-no votes per expression.

A Fleiss' kappa analysis of inter-rater reliability was calculated:

$$k = \frac{P_a - P}{P_{max} - P_e}$$

and results from unreliable workers were discarded, leaving 21 participants. Kappa reliability still fell in the poor to fair range, with a mean of 0.15. Workers may have answered randomly, and too few workers assigned per task. Because kappas were not strong, classifying an expression as recognizable was predicated on agreement between all three workers, leaving 183 expressions for further testing.

Phase Two

The remaining images were assigned to AMT for labeling by 40 individual workers per image at a pay rate of $.10 per response. Participants were instructed to use single word labels that they felt best described the emotion, intent, or internal state being signaled by the avatar's expression. With a few exceptions, the entire batch of 7320 Human Intelligence Tasks (HITs) completed within 48 hours. Most tasks ended with the full set of 40 labels per expression, but some had as few as 37 responses. We use the terms "label" and "word" interchangeably within this paper.

Label	
afraid	fearful
angry	furious
annoyed	glad
anticipating	happy
anxious	intense
astonished	interested
awestruck	mad
bored	miffed
calm	neutral
cheerful	peeved
concerned	pleased
confident	plotting
confused	questioning
contemplative	relaxed
content	resentful
contented	resigned
curious	sad
dejected	sarcastic
depressed	secretive
determined	shocked
disappointed	skeptical
disbelieving	sly
disdainful	smug
disgusted	stunned
displeased	surprised
distressed	suspicious
distrustful	tired
disturbed	uneasy
doubtful	unsure
eager	upset
enraged	worried
excited	

Table 2: The 63 unique expression labels gathered from our pilot study.

Deriving the Semantic Centroid Label

To determine the semantic centroid for a particular crowdsourced label set, we apply the Lesk algorithm from Similarity for WordNet [1], which computes a synonymy score given a word pair. The Lesk algorithm uses WordNet [11] as the lexical basis for calculating the amount of shared language between the definitions, or glosses, of two terms. Because the relations in WordNet generally do not cross part of speech boundaries, our process requires transforming all labels (words), where possible without changing their senses, to adjective form so consistent scoring can be obtained.

Given a label set corresponding to a single avatar, we first perform simple data cleaning operations to correct for spelling errors. Then, we compute synonymy scores for all word pairs. We assign a MAX = 400 as our Lesk score ceiling and normalize all scores in a 0-1 scale using the MAX value. Lesk defines no explicit upper bound for synonymy scores, so we determined a maximum cutoff by running the algorithm on all pairs of adjectives in WordNet, including same-word pairs. An analysis of the results (N = 23,109,275, M = 6.26, SD = 9.45) yielded a statistical baseline for significance. Between same-word pairs, the mean score was 372.53, rounded to 400 for the MAX value.

Performing per set all-pairs similarity calculations results in a weight vector of size n-1 for each word, n being the number of words in a set. As derived from the weighted sum function of [10], we compute the overall weight of a word label using the formula:

$$S(i) = \sum_{j=1}^{j} sim(w_i, w_j) / MAX$$

in which i is a given label, j is a comparison label, and sim() is the Lesk score between the two. We output the label with the maximum summed weight per set as the semantic centroid: the label that meaningfully captures the essence of the avatar's expression. We additionally check to ensure that the word with the second highest semantic score is closely associated with the top ranked word or has a significantly lower score.

Preliminary Results

Over all 183 expression label sets analyzed, mean synonymy was significantly higher than chance (Z = 2.58, $p < .001$), indicating that the crowd curated images were both communicative and likely to converge on a central signal value. Of the 183 response sets, 156 were found to have a centroid, 63 of which were unique words (table 2). See figure 4 for examples.

In several instances, different expressions had the same centroid. Overlap in label assignments may be due to word recall being weaker than recognition, and is also an indication that facial expression mapping is not one-to-one.

Future Work and Discussion

Initial testing offers a strong indication that it is possible to create a broad lexicon of nuanced expressions with associated signal value labels. However, limitations in our label processing methodology and the constraint of using single word descriptors, while an improvement over forced-choice designs, could be modified to obtain better results between parts of speech and accept compound labels. A means by which respondents can see words related to their initial choice and further refine their final answer could elicit more fine-grained labeling.

Label	Weight
worried	71.94
fearful	42.63
distressed	9.48
troubled	8.49
surprised	7.88
apprehensive	7.65
concerned	5.86
nervous	5.26
petrified	3.70
stunned	3.54
shocked	3.46
amazed	3.44
alarmed	3.18
startled	2.63
terrified	2.58
frightened	2.58
alert	2.11
dismayed	1.73
unnerved	1.65
fretful	1.28
powerless	1.25
intrigued	0.92
crushed	0.92
perturbed	0.66
bothered	0.32

Table 3: An example label set and the associated synonymy weightings calculated by comparing the semantic relatedness of each word to every other word in the list. The centroid is "worried," and corresponds to the far left image in figure 4.

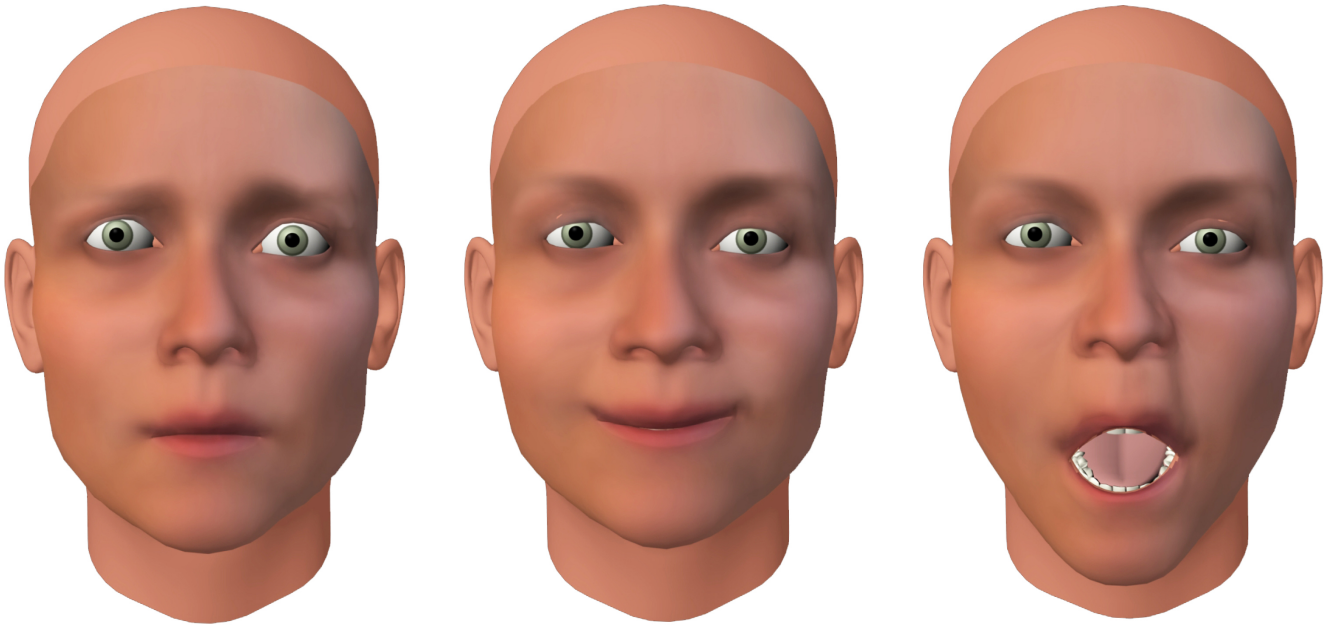

Figure 3: A sample of expressions with strong semantic centroids. From left to right: worried (table 3), pleased, and astonished.

Larger image sets need to be run over all AUs, at varying weights and counts per combination. Adding head and eye movements would expand the potential expression space. In addition, improvements to the model itself need to be made so the deformations that occur appear more realistic and prevent collisions between morphs. Validation of the model should be performed against known, coded expression sets. Because recognition is easier in a dynamic context, short animations illustrating the onset, peak, and offset of expressions would be ideal for future testing.

A comprehensive mapping of communicative content to known facial movement parameters will be a powerful tool for both animation and recognition. Virtual humans can be driven semantically rather than by performance capture or manual animation. By borrowing the technique of retargeting from motion capture [6], which transfers human movements to an animated character, automated expression recognition could be realized without a machine learning basis. Rather than training software on datasets of thousands of manually annotated photographs, matching an AU vector resulting from retargeting a novel face to our model would return the matching facial expression.

References

1. Satanjeev Banerjee and Ted Pederson. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. *Lecture Notes In Computer Science*, 136-145.
2. Judee K. Burgoon, Laura K. Guerrero, Kory Floyd. 2009. *Nonverbal Communication.* Allyn & Bacon, Boston, MA.
3. Darren Cosker Eva Krumhuber, and Adrian Hilton. 2010. Perception of linear and nonlinear motion properties using a FACS validated 3D facial model. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization* (APGV '10), 101-108. http://doi.acm.org/10.1145/1836248.1836268
4. Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings Of The National Academy Of Sciences Of The United States Of America,111(15), E1454-E1462.* doi:10.1073/pnas.1322355111
5. Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. Facial Action Coding System: The manual on CD ROM. A Human Face, Salt Lake City, 2002.
6. Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime facial animation with on-the-fly correctives. *Acm Transactions On Graphics, 32(4).* http://dx.doi.org/10.1145/2461912.2462019
7. Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. Research Report: It's only a computer: Virtual humans increase willingness to disclose. *Computers In Human Behavior, 3794-100*. doi:10.1016/j.chb.2014.04.043
8. Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, and Zara Ambadar. 2010. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference On Computer Vision And Pattern Recognition - Workshops, CVPRW 2010, (*2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*)*, 94-101. doi:10.1109/CVPRW.2010.5543262
9. Seiko Minoshita, Nobuaki Morita, Toshiyuki Yamashita, Maiko Yoshikawa, Tadashi Kikuchi, and Shinji Satoh. 2005. Recognition of affect in facial expression using the Noh Mask Test: comparison of individuals with schizophrenia and normal controls. *Psychiatry And Clinical Neurosciences*, *59*(1), 4-10.
10. Deepak P, Prasad M. Deshpande. 2015. *Operators for Similarity Search: Semantics, Techniques and Usage Scenarios.* Springer International.
11. Princeton University. 2014. About WordNet. Retrieved January 13, 2016 from http://www.wordnet.princeton.edu/
12. Etienne B. Roesch, Lucas Tamarit, Lionel Reveret, Didier Grandjean, David Sander, and Klaus R. Scherer. 2011. FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal Of Nonverbal Behavior*, *35*(1), 1-16.
13. Siman A. Surguladze, Andrew W. Young, Carl Senior, Gildas Brébion, Michael J. Travis, and Mary L. Phillips. 2004. Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. *Neuropsychology*, *18*(2), 212-218.
14. Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM Transactions On Graphics* (SIGGRAPH '11), *30*(4), http://dx.doi.org/10.1145/2010324.1964972
15. Hui Yu, Oliver G.B. Garrod, and Phillipe G. Schyns. 2012. Technical Section: Perception-driven facial expression synthesis. *Computers & Graphics*, *36*(Novel Applications of VR), 152-162. doi:10.1016/j.cag.2011.12.002