More than a Feeling: The MiFace Framework for Defining Facial Communication Mappings

Crystal Butler
New York University
New York, USA
cb2610@nyu.edu

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

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

Winslow Burleson
New York University
New York, USA
wb50@nyu.edu

ABSTRACT

Facial expressions transmit a variety of social, grammatical, and affective signals. For technology to leverage this rich source of communication, tools that better model the breadth of information they convey are required. MiFace is a novel framework for creating expression lexicons that map signal values to parameterized facial muscle movements inferred by trained experts. The set of generally accepted expressions established in this way is limited to six basic displays of affect. In contrast, our approach generatively simulates muscle movements on a 3D avatar. By applying natural language processing techniques to crowdsourced free-response labels for the resulting images, we efficiently converge on an expression's value across signal categories. Two studies returned 218 discriminable facial expressions with 51 unique labels. The six basic emotions are included, but we additionally define such nuanced expressions as embarrassed, curious, and hopeful.

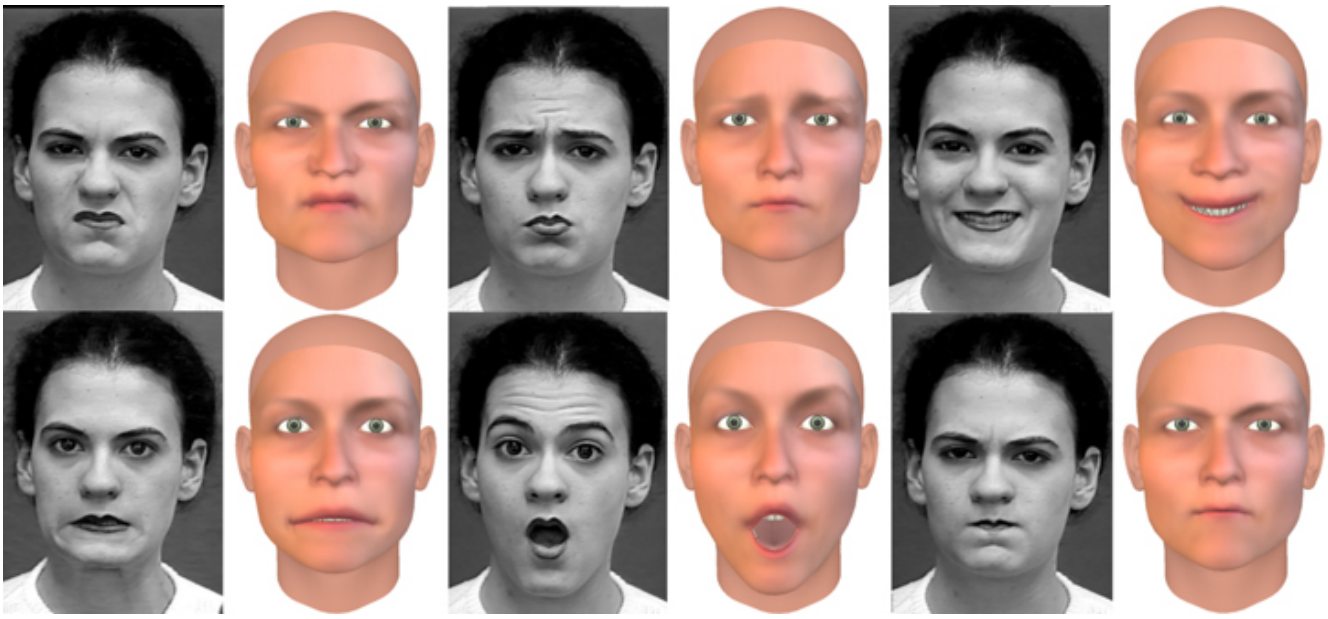


Figure 1. The six basic emotions from the Cohn-Kanade database [38], matched to our avatar, from top left: disgust, sadness, happiness, fear, surprise, and anger.

UIST 2017, *October 22–25, 2017, Quebec City, QC, Canada*

ACM 978-1-4503-4981-9/17/10…$15.00
https://doi.org/10.1145/3126594.3126640

Author Key Words

Facial expression recognition; virtual humans; 3D modeling; avatars; affective computing; natural language processing; social signal processing.

ACM Classification Keywords

Human-centered computing ~ Empirical studies in HCI • Applied computing ~ Psychology

INTRODUCTION

When people interact, up to 65% of the communication that occurs is non-verbal [21]. While non-verbal signals may be transmitted via gestures, posture, gaze, or paralinguistics

[32], the face is singularly expressive. Facial expressions (FEs) can convey felt or emblematic emotion [17], intent [22], cultural norms [53], cognitive states [51], or social signals [27]. Attempts have been made to create word relations to facial movements around emotion or hedonics [61], however, to date no compendium of labeled, quantified FEs that captures the full breadth of human facial behavior has been produced. The most accepted set of FEs is instead limited to six expressions of emotion (Figure 1).

Ideally, well-designed FE lexicons will accurately map facial muscle movements to fine-grained labels, identifying signal values for each movement configuration. MiFace is a process for building such lexicons, with a novel approach that we hope will facilitate rapid progress in FE modeling and recognition. Lexicons built using MiFace could be used to enrich research, mental healthcare, entertainment, security, and commerce. Areas of application include psychological testing and treatment, the design of intelligent affective agents, procedural character animation for games, and automated facial expression recognition (FER).

The research methods documented in this paper grew out of an awareness that a shared, quantifiable understanding of nonverbal communication can serve as the foundation for broad growth in computer-based modeling and recognition of human behavior. While there are many communication modalities that require study and integration to paint a complete picture, fine-grained understanding may be best obtained by examining the parts individually. Our focus is on the face, which is capable of conveying particularly nuanced signals.

CONTRIBUTION: A NEW WAY OF MAPPING FEs

Darwin's *The Expression of the Emotions in Man and Animals,* which provides detailed descriptions of facial behaviors and posits their evolutionary origin, was first published 145 years ago [13]. Why, then, is the generally accepted set of facial expressions still limited to six basic emotions? Humans are capable of generating and interpreting a much larger set of FEs. We argue that the labor, expertise and expense required by the traditional approach to studying FEs has hindered the development of broader lexicons.

Facial expression mapping is a challenging problem, difficult to address using existing computational and psychological research techniques. Debates over the universality of FEs and their relation to emotion have dominated the literature in psychology [16, 25, 52, 51], with little attention given to expanding the accepted lexicon. The gold standard for mapping muscle activations to FEs is the Facial Action Coding System (FACS) [17]. Using FACS, human experts annotate discrete craniofacial muscle movements called Action Units (AUs).

Automated recognition systems have been limited in the number of FEs they identify because they are informed by psychology research, and require large sets of images annotated by FACS-certified coders [4, 10, 38] as input to a machine learning phase in which features for each FE are extracted. Creating an image set from scratch requires expert FACs coders to either pose each expression, coach volunteers, or manually annotate images captured in the wild. The minimal FACs competency needed to code each expression with the correct AUs and intensities, and achieve satisfactory inter-rater reliability, requires at least 100 hours of dedicated training [4, 10, 24, 38]. Typically, images depict a narrow range of subjects, resulting in sets that lack the diversity in AU intensity, race, age, and gender needed to obtain accuracy in real-world situations.

In contrast, the MiFace framework uses a 3D digital model (avatar) with known muscle activation parameters as the basis for generating FE images. Our avatar employs FACS-based deformations of 24 key AUs on a 0-1 scale, each of which can be manipulated independently. Any weighted combination of AUs can thus be generated programmatically. Crowdsourced naïve judges determine which combinations of AUs are considered recognizable expressions, and provide sets of single-word labels describing their perceptions of what an expression communicates. When set coherence is strong (as described in the Hierarchical Agglomerative Clustering sub-subsection of Design and Methods), a single, representative label is determined using natural language processing (NLP). While our initial tests were run using one avatar, an infinite number of appearances based on age, race, gender, and ethnicity morphs (alterations in the morphology of a digital model) is possible.

Traditional methods of building a FE database are detailed further in [38], which describes the creation of the Extended Cohn-Kanade Dataset. In our two proof-of-concept studies, detailed in the Labeling and Semantic Similarity Results section, we identified 218 reliably recognizable facial expressions mapped to known AU activation parameters and crowdsourced labels. These results indicate that our method can be used to expand the known FE space substantially with much lower overhead than traditional methods require.

APPLICATIONS: FE LEXICONS FOR SMART INTERACTIVE SYSTEMS

FER variability has been found to correlate with psychiatric disorders such as PTSD, schizophrenia, and depression [23, 41, 44, 47, 68], as well as syndromes of atypical development including Autism Spectrum Disorder and Attention-Deficit Hyperactivity Disorder [57]. An extensive FE lexicon could act as the foundation for more comprehensive computer-based diagnostics and social skills training [58] by modeling expressions on virtual humans. Reproducing FEs in this way would allow for the creation of test sets in which the model can vary by race, age, and gender, but remain consistent in the expressions displayed. Avatars for training could also be tailored to the user for an improved sense of affiliation.

Virtual humans can also augment service sectors, acting as digital assistants [18], teaching coaches [6], or nurses' aides [71]. They may relieve the burden of "emotional labor" induced by performing client-facing work that requires maintaining a consistent attitudinal façade [28, 30]. By dependably providing an empathic response [37, 65, 30], they can increase satisfaction levels of their human clients.

Having a known lexicon of facial behaviors to draw from could significantly improve the quality of procedurally generated FEs displayed by ancillary non-player characters (NPCs) in video games. While primary characters are commonly animated using motion capture data generated by tracking professional actors [66], more cost–and time–efficient methods are required for the development of NPCs [55]. With a MiFace lexicon, context-based FEs could be generated for subtler NPC interactions using simple natural language triggers.

In addition to acting as a foundation for expressive digital agents, this lexicon could expand the range of FEs that can be identified by automated recognition. Automated FER has received a great deal of attention the past 15 years [40, 54, 70]. Systems that perform robust automatic FER can support a broad range of human activity, including: recognizing pain or confusion in a healthcare setting [30, 48]; providing tools to assess mental wellness [23, 44]; measuring the attention level of students or an audience [65, 67]; gauging a client's state of mind during a customer care session [60]; and enhancing security [34].

While as yet untested, recent improvements in motion capture retargeting—using the movements of a human actor to drive an animated character—may allow for automated FER based on movements calculated from the deformations of a digital target rather than machine learning over large image datasets. Because we use a digital model to emulate muscle movements, doing automated FER with MiFace should be possible by retargeting movements then matching the resulting model deformations to known expression mappings.

RELATED WORK

Facial Expressions as Signals

Social scientists have put forward a wide range of theories of emotion, arguing that they arise as part of the evolutionary process and are universal to all humans [16, 47], or are largely learned constructs that are culturally dependent [25, 39]. Some have avoided the issue of causality, preferring to focus on classification [42]. Of particular relevance to our method of processing label sets is the prototype approach to emotion, which views emotions as members of fuzzy sets grouped in a tree-like structure [19, 20, 56]. This perspective provides a much broader basis for the mapping of emotion to language than conceiving of emotion as discrete or basic entities. In fact, [56] identified 135 hierarchically clustered emotion names.

Some researchers have articulated an expanded expression space, notably [15] at Ohio State University, where a recent study found that observers could categorize compound expressions, e.g., fearfully surprised. Including the six basic emotions, their results define 21 discrete FEs. This experiment adds strength to our argument that many discriminable expressions remain to be identified, but 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 that represent the full diversity of human expression.

While such results support our assertion that an extensive set of FEs is definable, not all emotion words have corresponding facial signals, and not all expressions denote emotion. Physical or cognitive states such as pain, sleepiness, or confusion can be reflected in the face. Expressions can act as conversational "punctuation" [17]. As exemplified by the "not face," which is a facial representation of negation, they can also perform as non-verbal substitutes for linguistic communication [5]. Many spontaneously produced FEs naturally convey emotion. However, a more inclusive exploration of facial behavior can be found in the domain of social signal processing [64], which encompasses multiple modalities and characteristics of social information exchange.

Avatars for FE Research

Faceshift, the software used as the basis for creating our avatar, was developed for markerless facial motion capture and retargeting [66]. Retargeting is the process of transferring a human actor's movements to a digital character. In facial retargeting, surface deformations are captured by tracking markers on the actor's face or, as with faceshift, acquiring depth and texture information from a sensor. Retargeting is frequently used in modeling for animation and research [35].

In a sophisticated hybrid approach to using digital avatars for FE modeling and testing, [69] developed a highly realistic, FACS-based 3D morphable model capable of synthesizing arbitrary combinations and weightings of AUs, as does MiFace. Untrained observers reliably identified the six basic emotions in a forced-choice study. However, no research expanding the repertoire of recognizable facial behaviors was reported. FACSGen, a high-quality FACS-based modeling tool [50], does not rely on capturing human performance or geometry, and is natively digital. The D3DFACS Database comprises a set of 519 AU sequences, captured by scanning human actors with a configuration requiring six cameras [12, 11].

Ochs, Pelachaud, and McKeown tested polite, amused, and embarrassed smile variants using an embodied conversational agent (ECA) named Greta, a medium-fidelity digital human [43]. Greta's smiling behavior was designed and validated by crowdsourced study participants. An initial set of participants selected from a range of preset movement parameters to generate short animations of Greta performing smile variants. A validation study using a

separate group of participants demonstrated that viewer perceptions of Greta's smiling behavior matched the intents of the smile designers, and that her synthesized movements conformed to those expected based on existing FE mappings. The Greta studies show that a virtual human can elicit an appropriate affective response, and need not appear hyper-realistic to do so.

Avatars have also been used to measure other kinds of social and emotional inferences made by observers. Morphs of an average male 3D face model were constructed and used as targets for study participants to judge attributes such as attractiveness, competency, dominance, extroversion, and trustworthiness based on facial appearance by [62]. Jack and colleagues produced a morphable, expressive avatar that incorporated temporal dynamics by using animated expressions. They tested for cultural universality of the six basic emotions [29]. Conversely, AniAvatar was used to evaluate the performance of an animated avatar for self-evaluation of affective state [59].

DESIGN AND METHODS

Creating the MiFace framework for defining facial communication mappings involved two proof-of-concept studies and the development of two critical components: a 3D model that is lifelike enough to accurately mimic human facial movements and a means for extracting a single word label from a set of responses that encapsulates the set semantics. To validate our results, we performed a human semantic similarity test. As outlined below, this section provides a detailed description of our model and experimental designs, with implementation details. Subsections are ordered stepwise as performed. Results follow in a separate section.

1. Avatar Design: Generating believable faces on a 3D model
2. Gathering Label Sets: Crowdsourced studies on Amazon's Mechanical Turk
   I. Phase One: Discerning expressions with communicative value
   II. Phase Two: Gathering label sets for each expression
3. Data Analysis: Similarity-moderated majority vote for best label
   I. Natural language processing to calculate a semantic centroid
   II. Hierarchical agglomerative clustering
4. Semantic Similarity Validation

Avatar Design: Generating Believable Faces on a 3D Model

Faceshift was chosen as the starting point for building our model based on the quality of its provided "average" head mesh (the model's polygonal surface geometry), capacity for creating and exporting custom facial movements, and ability to capture an actor's movement using a depth sensing camera without motion capture markers. More detail on faceshift can be found in the Related Work: Avatars for FE Research subsection, or in [66]. Initial production of the head mesh and facial actions for our trial avatar were created using the pipeline in Figure 2. An expert in FACS posed individual AUs at maximum activation for scanning to target motions defined in faceshift, some of which were provided with the software and some custom built. The resulting mesh and facial morph targets were edited in Maya, Autodesk's computer animation and modeling software.

Faceshift provides a fairly comprehensive set of morph targets, based loosely on FACS. Morphs are deformations of the model's base geometry, used to create smooth state transitions for animation. The model was skinned and a UV texture map–a two-dimensional image applied to the

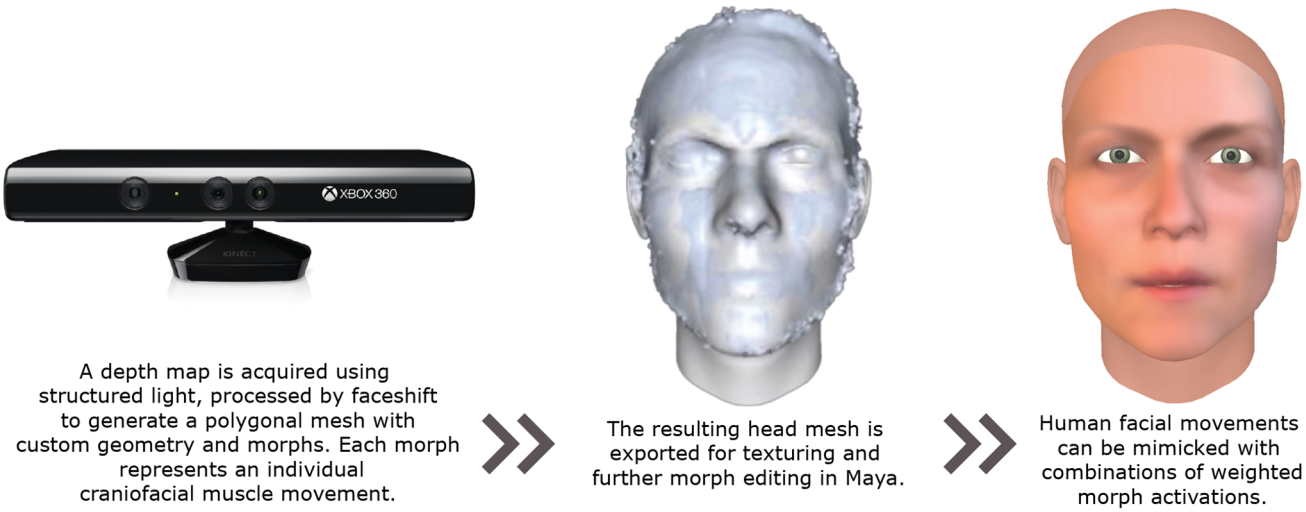


Figure 2. The avatar creation pipeline uses a depth sensor for head geometry capture, faceshift software trained with custom expressions, and Maya for final texturing and editing of facial movement morphs.

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

Table 1. Action Units used in the study, with levels of activation from 0-1. AU4 was over weighted to make it more distinct. AUs 25 and 26 are combined to portray an open mouth. An "L" indicates unilateral activation on the left side of the face.

geometry's surface–was acquired using a commodity webcam to photograph the FACS actor. Some modifications to the morph set were required to better approximate the surface changes described in FACS. These adjustments and the addition of missing FACS morph targets were manually modeled in Maya.

After final modeling, the avatar could perform 28 FACS movements, both unilaterally and bilaterally where applicable. Mouth, jaw, eye, and head movements were also integrated, but in order to focus on the movements typically studied in core FE research only mouth and jaw kinematics were used in the initial studies. Morph transforms were applied linearly to the base head mesh on a scale from 0-1, with 0 indicating no modification and 1 representing the maximum change from baseline. Sets of still images depicting AU combinations were generated programmatically in Maya using scripts written in the Maya Embedded Language.

To make our study manageable, we constrained the number of AUs per generated FE and displayed them at constant weights (Table 1). Using the FACS Investigator's Guide and the FACS certification test, we identified 22 key movements [17]. Based on expert knowledge, weights were fixed at levels that rendered the AUs plausible and clearly visible.

For Study One, three AUs were combined per FE, which is the most frequently occurring count found for FEs in the FACS Guide and test (excluding head and eye movements). Study Two used the same weightings applied to two AUs per FE. Because our proof-of-concept avatar is a relatively simple prototype, additional limitations were placed on using images in which highly additive or easily confused AUs would have been combined, i.e., AUs 6+7, 9+10, and 23+24. For all combinations of 22 choose 3 in Study One, that narrowed our list of candidates to 1380 movement combinations. Study Two began with 229 candidates.

Gathering Label Sets: Crowdsourced Studies on Amazon's Mechanical Turk

To determine which images were considered recognizable, realistic human FEs, and gather sets of candidate single word labels denoting the communicative content of a FE, we designed a two-phase process implemented on Amazon's Mechanical Turk (AMT), a site for crowdsourcing remote workers. Amazon defines a Human Intelligence Task (HIT) as a single assignment completed by an individual. For both phases, workers with acceptance rates of 95% or greater could participate. We did not collect demographic data about any of the participants.

Phase One: Discerning Expressions with Communicative Value

Workers on AMT were presented with 39 3x2 image grids, along with a neutral face (no AUs activated) for reference, with three workers assigned per grid. They were instructed to select the radio button next to any image that they believed represented a naturalistic human FE and paid $.10 per grid. A FE was classed as recognizable when two or more out of the three workers agreed it had communicative value. Of the 1380 candidate combinations from Study One, 341 were passed to phase two for labeling. Of the 229 images from Study Two, 114 went on to phase two.

Phase Two: Gathering Label Sets for Each Expression

Images rated as expressive from phase one of Studies One and Two were assigned to phase two for labeling on AMT. In phase two, 40 individual workers were assigned to each image, and tasked with freely choosing a single word label that best described the facial expression shown. Workers were payed at a rate of $.10 per response. Some images completed with as few as 37 responses, but most received a full set of 40 labels. Workers on AMT generally make the best income when they can do large numbers of familiar tasks. Because phase two of our studies required 40 unique workers per image, batches of HITS could become fragmented and unappealing as they drew close to finishing, leaving some HITS undone. In addition, a few workers who did a small number of HITS had their work thrown out entirely because they demonstrated a pattern of responses that did not follow our guidelines.

The AMT interface presented a single image per HIT, along with the same neutral face shown in phase one and instructions for evaluating the facial expression. Suggestions for producing a thoughtful evaluation include visualizing the expression dynamically, as it would be displayed in the course of conversation, and actively mimicking the movements modeled to produce a resonant internal response.

Participants were also given guidelines for word selection, including instructions to use a thesaurus. We provided a link to an online thesaurus (Figure 3). They were instructed to enter the single-word label that best described the emotion, mindset, or internal state being signaled by the avatar's FE. Label sets were cleaned prior to the analysis phase. Cleaning included spelling correction if the choice of word was clear; in ambiguous cases the response was

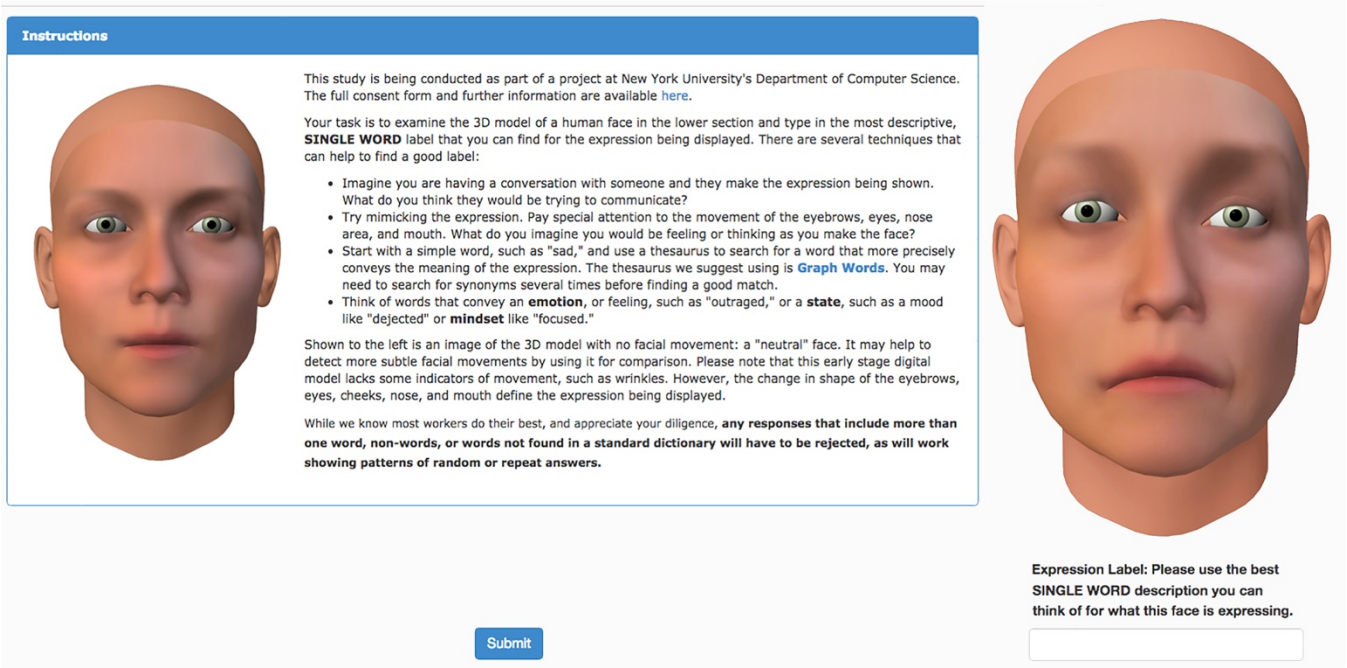


Figure 3. Phase two of the AMT labeling studies asked participants to pick the single-word label that best describes an expression. Guidelines were given on "reading" the face and word selection. A neutral face, with no facial muscle activation, was always shown for comparison to the expressive image.

removed. We also removed a small number of nonsense terms and multi-word phrases.

Data Analysis: Similarity-Moderated Majority Vote for Best Label

While using microtask platforms to quickly recruit a large, diverse participant pool is by now a well-established research practice [1, 8, 9, 31, 33], free-response labeling designs have been the exception in FE research. Lack of a rigorously defined, quantitative method for determining term relatedness within a set may be a limiting factor for prior free-response studies [26]. We address this deficit by applying a novel two-step process. First, the constituent responses of a label set are analyzed using a NLP algorithm to find the top-ranked, representative label for each set. Step two applies hierarchical agglomerative clustering to check whether most of the elements of a label set share meanings similar to the representative for that set.

Given a label set corresponding to a single FE image, we first performed simple data cleaning operations to correct for spelling errors and remove non-words. Then, we computed similarity scores for all word pairs using a modified version of the GloVe cosine distance program run with the 300-dimensional Wikipedia 2014 + Gigaword 5 pre-trained word vector set [45]. Although this corpus includes a vocabulary of 400,000 unique words, some of the labels supplied by our study participants are not referenced. Word pairs in which either or both labels were missing from the vocabulary were discarded. When multiple respondents provided the same label, it was allowed to repeat in the set as many times as given. Same word pairs were assigned a similarity score of one, which is the maximum on a scale that ranges from -1 (complete opposites) to 1. As derived from the weighted sum function of [14], we compute the overall weight $S(i)$ of a label:

$$S(i) = \sum_{\substack{i=1 \\ j=i+1}}^{n} (sim(l_i, l_j))$$

where $(sim(l_i, l_j))$ is the word vector cosine similarity between two labels and n is the total number of labels per expression. Instances in which a label is repeated are collapsed, and the algorithm outputs a list of unique labels with their sums within each set. If tests for set coherence are positive, the label with the maximum summed weight is considered the best label and assigned as the signal value of the associated FE. Example output of the weight summing across a single set is shown in Table 2.

Natural Language Processing to Calculate a Semantic Centroid

In two studies, we considered a total of 1609 facial muscle activation configurations. Of these, 218 passed all stages of human evaluation and computational analysis to be deemed recognizable expressions. Images with positive results can be said to have high ecological validity, making them more suitable for real-world applications. In addition, we performed a test of two NLP algorithms: Lesk from Similarity for WordNet, and GloVe (Global Vectors for Word Representations). As outlined in the Semantic Similarity Validation subsection, we scored 156 expression

Label	Summed Weight	Label	Summed Weight	Label	Summed Weight
hopeful	31.7059	grateful	9.6560	distressed	6.4927
bemused	31.0547	fearful	9.5687	uncertain	5.8384
happy	23.8122	timid	9.2012	fawning	4.7292
cheerful	21.9200	sorry	9.0905	aroused	4.3909
glad	20.1252	relieved	8.3232	ruffled	4.2597
friendly	18.8266	cheery	8.3161	abashed	4.1450
sympathetic	18.3660	sheepish	7.6618	questioning	3.5559
optimistic	10.6255	encouraging	7.0095	tender	2.5739
elated	10.4330	exhilarated	6.9536	grimacing	2.5404
overjoyed	9.7151	childlike	6.6897	aspiring	1.3912

Table 3. The list of unique input words for the expression "hopeful," with summed word vector cosine scores of semantic similarity for each. Semantic similarity considers features of relatedness beyond strict synonymy.

word pairs using both algorithms and calculated the Spearman's correlation coefficients against human ratings of synonymy. GloVe performed best, $r_s = .30$, $p < .001$.

GloVe is an algorithm that populates the x and y axes of a multidimensional matrix (Euclidean vector space) with a vocabulary derived from a given corpus. The nonzero number of times each word occurs together with another word in the corpus is calculated and recorded in the matrix. The number of words in the vocabulary determines the number of dimensions (frequency vectors). To find semantic relatedness GloVe uses a cosine similarity measure (shown below) to compute the distance of cosine angles between two words in a multidimensional context vector space [46]. Here X and Y represent frequency vectors, with n elements [63]:

$$similarity = \cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\ \sqrt{\sum_{i=1}^{n} Y_i^2}}$$

GloVe is based on the *distributional hypothesis*, which states that words with similar meanings tend to occur in similar contexts. If we observe two words that frequently occur together, we can assume they mean similar things, if they frequently co-occur with a third word. That word creates context, which provides the foundation for determining semantic relatedness using this algorithm [63].

Often while discussing semantic distance in word space vectors, *semantic similarity* and *semantic relatedness* are used interchangeably. Here we would like to emphasize the differences between the two terms in relation to this study. Semantic similarity is a type of semantic relatedness. Any two words that occur together in a text can have a meaningful semantic relationship if they share attributes. For example, two semantically related words might be synonyms, meronyms, hyponyms, or even antonyms and still display similarity. They can also be words that functionally relate or are frequently associated (e.g., "pen" and "paper"). For our purposes, semantic similarity is used to designate synonyms [63].

Based on results from [46], the default vocabulary and vector files were replaced with those of the chosen Wikipedia 2014 + Gigaword 5 corpus, a 300-dimensional pre-trained word vector and vocabulary set. The standard GloVe-1.2 cosine similarity distance python algorithm was altered to: accept two words as distinct input instead of a single word or multi-word phrase, print only the similarity score for the indicated word pair instead of all related words and scores, print an error line for any words not found in the vocabulary, and print a score of one instead of negative infinity for a word paired with itself.

Hierarchical Agglomerative Clustering

In developing our method, we argue that a good label set should group well around the top label, which is the word with the highest summed cosine similarity score. To check the coherence of our label sets, we executed a series of steps based on hierarchical agglomerative clustering. This clustering method performs a sequence of binary joins starting at the leaf level and continuing until all the base elements have been connected in a tree, or dendrogram. A visualization of the dendrogram allows the viewer to see natural groupings that might exist in the data, and explore clustering at various levels of granularity.

Hierarchical agglomerative clustering algorithms for the analysis of documents and other texts nicely depict semantic relatedness. However, we have found that the task of finding the "best" label out of a set of individual words is not addressed in the NLP literature. We chose hierarchical agglomerative clustering as a secondary step to validate the coherence of our label sets because the method accommodates varying degrees of heterogeneity that exist between sets. Specifically, the number of clusters created per set can differ based on how much intra-set dissimilarity exists. Alternative clustering methods, such as k-means, require a static number of classes to be determined.

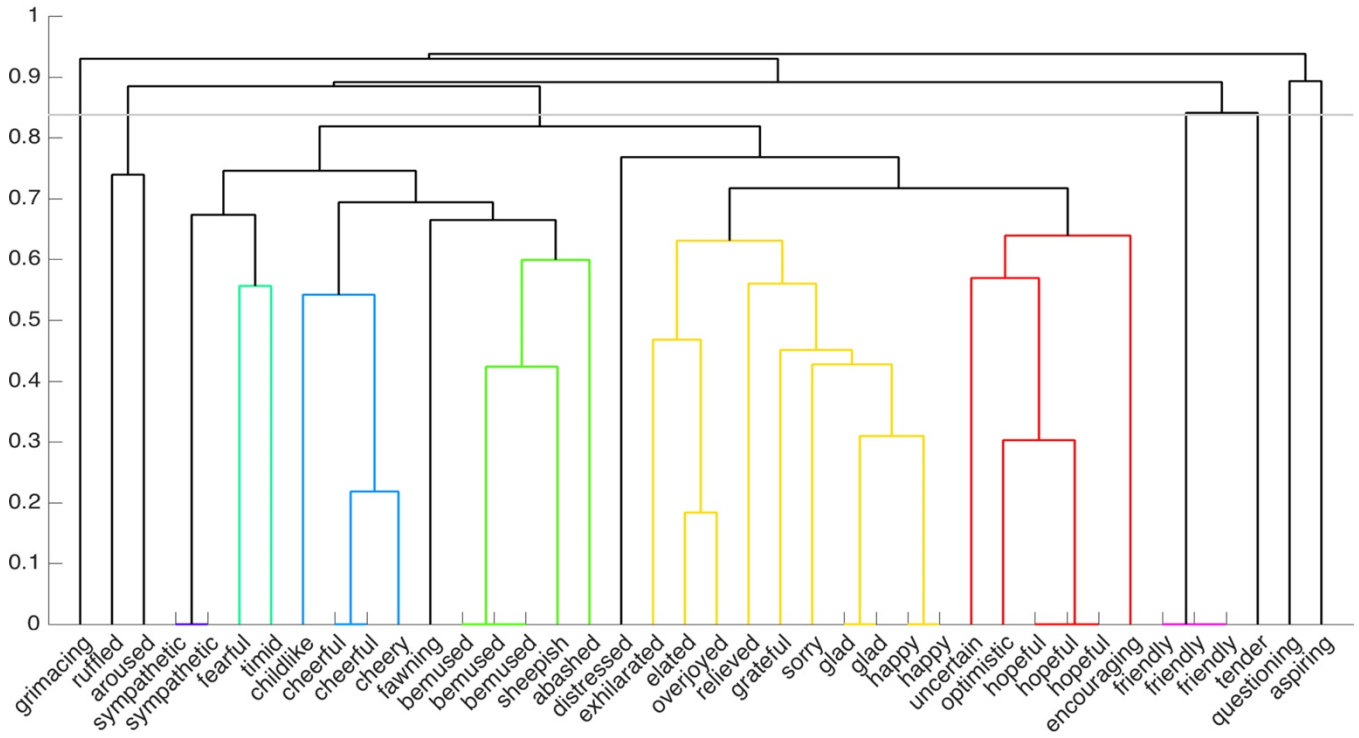


Figure 4. The dendrogram for the image depicting "hopeful", which breaks into seven clusters. The line through the y-axis at .8375 shows where the clusters are "cut" from the binary tree. Of the 40 leaf elements 31 fall into the primary cluster. At 77.5% membership in the primary cluster, this label set is near the low end of our floor for reliable recognition, which is 75%.

Analysis with this technique requires a distance matrix as input, which is easily generated by subtracting each of our cosine similarity scores from one. This transformation results in distances that range from zero, indicating complete similarity (i.e. same word), to an upper limit of two, which would represent an exact opposite. In practice, the maximum distance for any word pair was 1.3318 and the mean was 0.7665 for Study One. Study Two had a maximum distance of 1.1767 and a mean of 0.6497. The low upper bounds are likely due to labels having a degree of relatedness as facial communication descriptors**.**

To determine how best to perform clustering on our labeling data, two experts rated 15 sample label sets as likely to belong to a strongly grouped set, a poorly grouped set, or a mixed set. For testing, five sets per category were analyzed using Matlab with the Machine Learning and Statistics Toolbox. Experimentation led to the empirical determination that a dendrogram cutoff height of .8375 generates approximately the same number of semantic clusters expected by human experts. The cutoff height severs cluster connections above the specified level, leaving a varying number of smaller groupings (Figure 4).

Having established that a cutoff height of .8375 produces the most cohesive clusters for our data, we applied it to our 15 sample dendrograms, and then calculated the label count for each cluster. The average number of clusters was 4.4, and ranged from 2 to 7, matching our predictions. As implied by our requirements for a label set to be good, we want to see that, regardless of cluster count, the large majority of labels fall into a single cluster. For a label set to fail, the labels must either be highly dispersed among all clusters, or a large proportion of the label set must be split between several clusters.

To obtain an estimate of the level of membership in the main cluster required to indicate good clustering, the 15 samples were divided into two groups: one with high counts in a single cluster and one with high counts in two clusters, plus a single sample in which counts were dispersed among many clusters. Based on observations of membership in well-formed groups, we put in place a final filtering step in which sets with primary cluster membership <= 75% were thrown out. This value can be viewed as representative of the recognition rate, and is well aligned with accuracy levels seen in traditional FER tests using naïve judges.

Semantic Similarity Validation

Most NLP research centers on analyzing texts to extract concepts [7] or perform sentiment analysis [70]. There are algorithms that focus instead on word relatedness. After testing a selection of algorithms based on both cosine similarity measures and ontological approaches, GloVe

[45] and Similarity for WordNet [49] emerged as the analytical devices that most closely met our needs.

A common measure of the performance of a NLP algorithm is accuracy on a word similarity task, which shows the association between word-pair similarity scores returned by the algorithm and those given by human raters [46]. Comparisons between GloVe and several well-known NLP techniques indicate that it outperforms other word space vector models, with a Spearman's correlation coefficient for ranked data ranging from $r_s = 47.8$ to 83.6 across five test sets as demonstrated in [46]. The Adapted Lesk algorithm from Similarity for WordNet [2, 3] is an algorithm which returns a similarity score for a word pair based on the number of shared words in their WordNet definitions. The Adapted Lesk algorithm was tested on a word sense disambiguation task, the results of which cannot be directly contrasted with a similarity task.

Accordingly, we constructed a word pair set specific to our problem and acquired scores from human raters on AMT as the basis for our own comparison. To select the words used, all label pairs from Study One were assigned synonymy scores using both Lesk and GloVe. Unlike other algorithms available in Similarity for WordNet, the Lesk measure makes calculations across synset part of speech boundaries. However, the preponderance of nouns in WordNet renders such correspondences less accurate than comparisons made within the same part of speech [2, 3]. To counter this effect, if the word sense was unaltered we transformed labels to adjective form for more consistent scoring.

After calculating the summed scores of each label, the top-ranked 10% of unique labels from the combined sets of computations were selected for testing on AMT, giving 156 constituent words. Each label was randomly paired with another from the list for a total of 156 pairs. Two same-label pairs were included as gold standard questions. Workers were instructed to rate word pairs based on semantic similarity on a scale from 0 (complete opposites) to 10 (same or interchangeable words). Fifteen workers per word pair were paid $.10 per rating to complete the task. Ratings that fell outside of 1.5 times the interquartile range for each pair were discarded prior to averaging the results. After trimming, the number of responses per pair ranged from 9-15, with an average of 14.45.

The 156 pairs were then assigned synonymy scores using both Lesk and GloVe. Because Lesk calculates scores on a theoretical continuum from 0 to infinity, while GloVe ranges from -1 to 1, a Spearman's rank correlation was calculated to compare results from both algorithms to the averaged scores provided by AMT workers.

LABELING AND SEMANTIC SIMILARITY RESULTS

Labeling Results

In the first phase of each of the two expression identification studies described in the previous section, images were judged by naïve workers crowdsourced through AMT platform. A positive vote by two out of three participants determined which FEs were deemed naturalistic. Examples of three reliably recognizable FEs from phase one of Study One are shown in Figure 5.

In the second phase, we gathered single word label sets that approximately 40 participants per image felt best described what the FE was communicating. We generated 40 HITs per image. HITs could be completed by qualified workers at their discretion. Most sets of 40 hits were completed, but some garnered as few as 37 responses.

Label sets were processed using the GloVe (Global Vectors for Word Representations) word vector cosine similarity algorithm, modified to produce pairwise similarity scores between all labels in a set. After scoring, label sets were analyzed for coherence using hierarchical agglomerative clustering, and passing sets had their scores summed as described in the Natural language processing to calculate a semantic centroid sub-subsection.

From the passing sets, the two studies collectively produced 51 unique top labels (Table 3). In several instances, different images had the same semantic centroid. Overlap in label assignments may mean that workers selected less descriptive words more easily retrieved from memory rather than using the suggested thesaurus, and could indicate that FE mapping is not one-to-one.

Study One

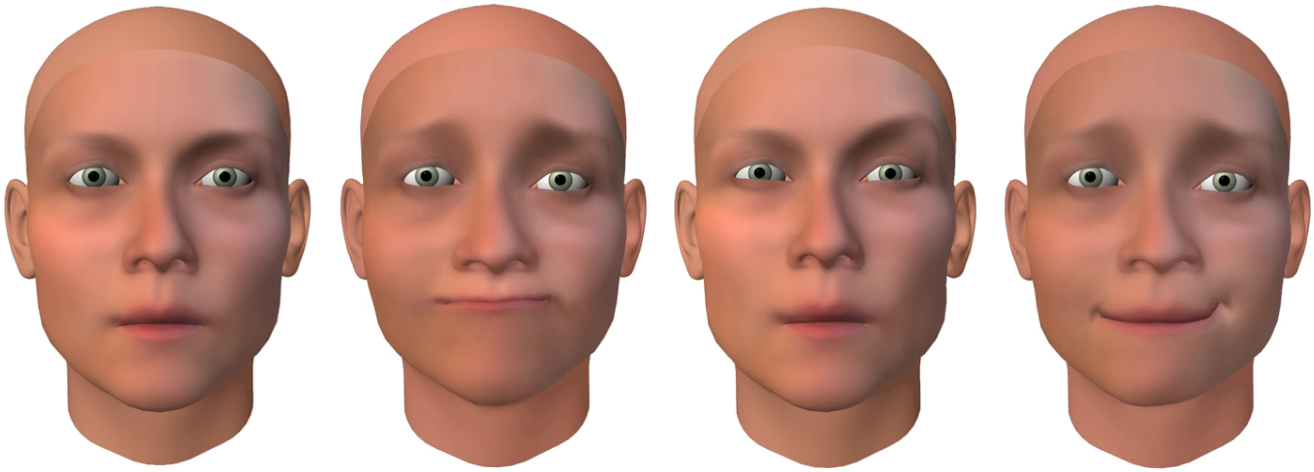
1380 images displaying three AU activations were assigned to AMT for testing in Phase One. AMT judges chose 341 images as recognizable expressions, which then went on to Phase Two testing for label acquisition. Hierarchical agglomerative clustering returned 157 passing and 184 failing label sets. Over all 261,007 label pair similarity scores, the minimum similarity score was -0.3317 and the maximum (excluding same-word pairs) was 0.8901 on a scale from -1 to 1.

Study Two

In Study Two, 229 images displaying two AU activations were tested in Phase One. Out of 114 label sets that went into Phase Two, 63 passed the clustering method and 51 failed. A minimum and maximum similarity score of -0.0001 and 0.8776 respectively were found from the 101,334 total label pair similarity scores.

Semantic Similarity Results

The purpose of this study was to evaluate how the Lesk and GloVe NLP algorithms score expression word pair similarity in relation to human raters. A Spearman's rank correlation was calculated to determine how closely the synonymy ratings the algorithms generate match human notions of similarity. Lesk performed poorly, $r_s = -0.02$, $p = .76$, showing a slight negative association between Lesk scores and human evaluations. GloVe performed much better, $r_s = .30$, $p < .001$. While the association between GloVe scores and human ratings of our label pairs is much lower than those reported in the authors' own testing [46], it remains the best choice of automated methods we tested.



AU	Weight
1	0.7
12	1.0
17	0.8

AU	Weight
L2	0.6
7	1.0

AU	Weight
1	0.7
12	1.0
L14	0.7

Figure 5. From left to right, the avatar's "neutral" face displaying no muscle activation, alongside the expressions "embarrassed," "curious," and "hopeful." All 218 recognizable expressions are documented in the appendix.

CONCLUSIONS AND FUTURE WORK

FE mapping has traditionally been a hard problem that was solved by using large sets of photographs annotated by experts as input to machine learning algorithms. This paper describes the design of a crowdsourced FE mapping system that uses free response label sets to derive a meaningful textual description of FEs, which are generatively modeled on a 3D avatar parameterized with action unit representations of muscle movements. Initial testing offers a strong indication that it is possible to create a broad lexicon of nuanced FEs with associated signal value labels.

Two studies, using 1609 initial test images, returned 455 FEs considered realistic by AMT workers. Of those, 218 had strongly clustered label sets and were assigned a single "best" label as the semantic centroid, 51 of which were unique. However, images labeled as one of the six basic emotions accounted for 68 out of 157 images in Study One and 27 out of 61 images in Study Two. By employing a subset of AUs in the image generation process that was defined, in part, by the FACS AUs listed in the Emotion Prediction table [17], we may have predisposed our model toward expressing variants of basic emotions. In contrast to work that focuses solely on FEs as conveyers of emotion, however, our framework also yielded terms that represent internal states, like tired, or mindsets, such as curious.

Determining the semantic centroid of a set of words was an unexpectedly difficult task—no rigorously defined process for doing so was found by the authors. Establishing an objective mechanism to perform this calculation so a primary label for a response set could be selected became a substantial part of testing MiFace. The hierarchical agglomerative clustering method we developed to test whether label sets are internally coherent and have a single semantic centroid is intuitively satisfying, as it allows for a natural ordering of the relatedness between labels, and is easily understood as a visualization. In combination with calculating summed all-pairs cosine similarity scores, we have established a unique method of analyzing free-response label sets for coherence and meaning.

Unique Labels		
amazed	disapproving	sad
amused	discouraged	satisfied
angry	disgusted	scared
annoyed	eager	shocked
anxious	embarrassed	skeptical
apprehensive	enraged	smug
arrogant	excited	startled
blase	fearful	stunned
bored	frightened	surprised
cheerful	furious	suspicious
concerned	happy	tired
confused	hopeful	uneasy
contented	interested	unimpressed
curious	joyful	unsure
dejected	outraged	upset
delighted	pleased	vindictive
disappointed	puzzled	worried

Table 3. The 51 unique expression name labels identified by our set analysis method.

Applications to interactive systems for this lexicon development framework include modeling believable facial behaviors with virtual humans and greatly expanding the repertoire of expressions that can be identified by automated FER. Virtual humans can augment a wide range of human activities, acting as digital assistants, teaching coaches, expressive video game characters, or caregiver aides. Automated FER enables feedback for smart software such as interactive digital learning platforms, custom recommender systems, marketing, and security systems, among other assistive systems.

There are several avenues for expanding upon this work. Our label processing methodology is constrained to single word descriptors. An obvious improvement is to accept compound labels and phrases. Additionally, we could provide judges with synonyms to their free-response choices directly within the test environment. Future studies will include FEs generated from a greater number of AUs at varying activation levels, and incorporate animation rather than being limited to static images. We also aim to expand to a larger word corpus such as the Common Crawl, which has 42 billion tokens of web data and 1.9 million words of vocabulary, and integrate ontological features to improve synonymy scores as described in [36].

Finally, the avatar needs to be redeveloped for improved realism. Desired changes include the incorporation of skin wrinkles, higher-fidelity surface textures, and better representation of difficult AUs such as 23, lip tightener. Once it has been redesigned, broader testing of responses across age, gender, race, and ethnicity in both the respondent populations and avatars represented can be performed. However, our primary goal remains building a foundational, ground-truth lexicon of FE mappings using a single avatar.

ACKNOWLEDGEMENTS

The authors would like to thank New York University for providing computing resources for data analysis.

REFERENCES

1. Paul André, Aniket Kittur, and Steven P. Dow. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '14)*. ACM Press, New York, NY, 989–998. DOI:https://doi.org/10.1145/2531602.2531653
2. Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*. Springer-Verlag, London, UK, 136–145.
3. Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificial intelligence (IJCAI'03)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 805–810.
4. Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. 1999. Measuring facial expressions by computer image analysis. *Psychophysiology* 36, 2 (Mar. 1999), 253–263. DOI:https://doi.org/10.1017/S0048577299971664
5. C. Fabian Benitez-Quiroz, Ronnie B. Wilbur, and Aleix M. Martinez. 2016. The not face: A grammaticalization of facial expressions of emotion. *Cognition* 150 (May 2016), 77–84. DOI:https://doi.org/10.1016/j.cognition.2016.02.004
6. Winslow Burleson and Rosalind Picard. 2007. Evidence for gender specific approaches to the development of emotionally intelligent learning companions. *IEEE Intell. Syst.* 22, 4 (Aug. 2007), 62–69. DOI:https://doi.org/10.1109/MIS.2007.69
7. Hiram Calvo, Oscar Méndez, and Marco A. Moreno-Armendáriz. 2016. Integrated concept blending with vector space models. *Comput. Speech Lang.* 40 (Nov. 2016), 79–96. DOI:https://doi.org/10.1016/j.csl.2016.01.004
8. Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '15)*. ACM Press, New York, NY, 600–611. DOI:https://doi.org/10.1145/2675133.2675214
9. Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the 31st Annual CHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM Press, New York, NY, 1999–2008. DOI:https://doi.org/10.1145/2470654.2466265
10. Jeffrey F. Cohn, Adena J. Zlochower, James Lien, and Takeo Kanade. 1999. Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology* 36, 1 (Jan. 1999), 35–43. DOI:https://doi.org/10.1017/S0048577299971184
11. Darren Cosker, Eva Krumhuber, and Adrian Hilton. 2011. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*. IEEE Computer Society, Los Alamitos, CA, 2296–2303. DOI:https://doi.org/10.1109/ICCV.2011.6126510
12. 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 Symposium on Applied Perception in Graphics and Visualization (APGV '10)*. ACM

Press, New York, NY, 101–108. DOI:https://doi.org/10.1145/1836248.1836268

13. Charles Darwin and Paul Ekman. 2009. *The expression of the emotions in man and animals*. Oxford University Press, New York, NY.
14. Deepak P. and Prasad M. Deshpande. 2015. *Operators for similarity search: Semantics, techniques and usage scenarios*, Springer International Publishing, Cham, Switzerland. DOI:10.1007/978-3-319-21257-9
15. Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. In *Proceedings of the National Academy of Sciences (PNAS '14)*. National Academy of Sciences, Washington, DC, E1454–E1462. DOI:https://doi.org/10.1073/pnas.1322355111
16. Paul Ekman. 1994. All emotions are basic. In *The Nature of Emotion: Fundamental Questions*. Oxford University Press, New York, NY, 15–19.
17. Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. 2002. A human face. In *Facial Action Coding System: The manual on CD ROM*.
18. Clarence (Skip) Ellis and Paulo Barthelmess. 2003. The neem dream. In *Proceedings of the 2003 conference on diversity in computing (TAPIA '03)*. ACM Press, New York, NY, 23–29. DOI:https://doi.org/10.1145/948542.948548
19. Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM Press, New York, NY, 4647–4657. DOI:https://doi.org/10.1145/2858036.2858535
20. Beverley Fehr and James A. Russell. 1984. Concept of emotion viewed from a prototype perspective. *J. Exp. Psychol Gen.* 113, 3 (Sept. 1984), 464–486. DOI:https://doi.org/10.1037/0096-3445.113.3.464
21. Gretchen N. Foley and Julie P. Gentile. 2010. Nonverbal communication in psychotherapy. *Psychiatry (Edgmont)* 7, 6 (Jun. 2010), 38–44.
22. Nico H. Frijda. 1987. Emotion, cognitive structure, and action tendency. *Cognition & Emotion* 1, 2 (Apr. 1987), 115–143. DOI:https://doi.org/10.1080/02699938708408043
23. Takeo Fujiwara, Rie Mizuki, Takahiro Miki, and Claude Chemtob. 2015. Association between facial expression and PTSD symptoms among young children exposed to the Great East Japan Earthquake: A pilot study. *Front. in Psychol.* 6 (Oct. 2015). DOI:https://doi.org/10.3389/fpsyg.2015.01534
24. Jeffrey M. Girard, Jeffrey F. Cohn, Laszlo A. Jeni, Michael A. Sayette, and Fernando De la Torre. 2015. Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behav. Res. Methods* 47, 4 (Dec. 2015), 1136–1147. DOI:https://doi.org/10.3758/s13428-014-0536-1
25. Steven L. Gordon. 1981. The sociology of sentiments and emotion. *Social psychology: sociological perspectives*. Transaction Publishers, New Brunswick, NJ, 562-592.
26. Jonathan Haidt and Dacher Keltner. 1999. Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. *Cognition & Emotion* 13, 3 (May 1999), 225–266. DOI:https://doi.org/10.1080/026999399379267
27. Shlomo Hareli and Ursula Hess. 2012. The social signal value of emotions. *Cognition & Emotion* 26, 3 (Apr. 2012), 385–389. DOI:https://doi.org/10.1080/02699931.2012.665029
28. Arlie Russell Hochschild. 2012. *The managed heart: commercialization of human feeling*, Univ. of California Press, Berkeley, CA.
29. Rachael E. Jack, Oliver G. B.Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. Facial expressions of emotion are not culturally universal. In *Proceedings of the National Academy of Sciences (PNAS '12)*. National Academy of Sciences, Washington, DC, 7241–7244. DOI:https://doi.org/10.1073/pnas.1200155109
30. Philip L. Jackson, Pierre-Emmanuel Michon, Erik Geslin, Maxime Carignan, and Danny Beaudoin. 2015. EEVEE: The empathy-enhancing virtual evolving environment. *Front Hum Neurosci* 9 (Mar. 2015). DOI:https://doi.org/10.3389/fnhum.2015.00112
31. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM Press, New York, NY, 1301–1317. DOI:https://doi.org/10.1145/2441776.2441923
32. Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Trans. Affective Comput.* 4, 1 (Jan. 2013), 15–33. DOI:https://doi.org/10.1109/T-AFFC.2012.16
33. Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting references between text and charts via crowdsourcing. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM Press, New York, NY, 31–40. DOI:https://doi.org/10.1145/2556288.2557241
34. Walter S. Lasecki, Young Chol Song, Henry Kautz, and Jeffrey P. Bigham. 2013. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM Press, New

York, NY, 1203–1212. DOI:https://doi.org/10.1145/2441776.2441912

35. Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime facial animation with on-the-fly correctives. *ACM T. Graphics* 32, 4 (Jul. 2013), 1. DOI:https://doi.org/10.1145/2461912.2462019
36. Wei Lu, Yuanyuan Cai, Xiaoping Che, and Yuxun Lu. 2016. Joint semantic similarity assessment with raw corpus and structured ontology for semantic-oriented service discovery. *Personal and Ubiquitous Computing* 20, 3 (May 2016), 311–323. DOI:https://doi.org/10.1007/s00779-016-0921-0
37. 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. *Comput. Hum. Behav.* 37 (Aug. 2014), 94–100. DOI:https://doi.org/10.1016/j.chb.2014.04.043
38. Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW '10)*. IEEE Computer Society, Los Alamitos, CA, 94–101. DOI:https://doi.org/10.1109/CVPRW.2010.5543262
39. Catherine Lutz. 1988. *Unnatural emotions: Everyday sentiments on a Micronesian atoll & their challenge to western theory*, University of Chicago Press, Chicago, IL.
40. Daniel McDuff, Rana El Kaliouby, and Rosalind W. Picard. 2012. Crowdsourcing facial responses to online videos. *IEEE Trans. Affective Comput.* 3, 4 (Jan. 2012), 456–468. DOI:https://doi.org/10.1109/T-AFFC.2012.19
41. 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. *Psychiat. Clin. Neuros.* 59, 1 (Feb. 2005), 4–10. DOI:https://doi.org/10.1111/j.1440-1819.2005.01325.x
42. Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10).* Association for Computational Linguistics, Stroudsburg, PA, 26–34.
43. Magalie Ochs, Catherine Pelachaud, and Gary Mckeown. 2017. A user perception-based approach to create smiling embodied conversational agents. *ACM Trans Interact. Intell. Syst.* 7, 1 (Jan. 2017), 1–33. DOI:https://doi.org/10.1145/2925993
44. Doris Peham et al. 2015. Facial affective behavior in mental disorder. *J. Nonverbal Behav.* 39, 4 (Dec. 2015), 371–396. DOI:https://doi.org/10.1007/s10919-015-0216-6
45. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation, version 1.2. (2014). Retrieved October 16, 2016 from http://nlp.stanford.edu/projects/glove/
46. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '14)*. Association for Computational Linguistics, 1532–1543. DOI: 10.3115/v1/D14-1162
47. Robert Plutchik. 2001. The nature of emotions. *Am. Sci.* 89, 4 (Jul. 2001), 344–350. DOI:https://doi.org/10.1511/2001.4.344
48. Marie Postma-Nilsenová, Eric Postma, and Kiek Tates. 2015. Automatic detection of confusion in elderly users of a web-based health instruction video. *Telemed. J. e-Health* 21, 6 (Jun. 2015), 514–519. DOI:https://doi.org/10.1089/tmj.2014.0061
49. Princeton University. 2015. About WordNet. (Mar. 2015). Retrieved November 10, 2016 from https://wordnet.princeton.edu/
50. 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. *J. Nonverbal Behav.* 35, 1 (Mar. 2011), 1–16. DOI:https://doi.org/10.1007/s10919-010-0095-9
51. James A. Russell. 1995. Facial expressions of emotion: What lies beyond minimal universality? *Psychol. Bull.* 118, 3 (Nov. 1995), 379–391. DOI:http://dx.doi.org/10.1037/0033-2909.118.3.379
52. James A. Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol. Bull.* 115, 1 (Feb. 1994), 102–141. DOI:https://doi.org/10.1037/0033-2909.115.1.102
53. Saba Safdar, Wolfgang Friedlmeier, David Matsumoto, Seung Hee Yoo, Catherine T. Kwantes, and Hisako Kakai. 2009. Variations of emotional display rules within and across cultures: A comparison between Canada, USA, and Japan. *Can. J. of Behav. Sci.* 41, 1 (Jan. 2009), 1–10. DOI:https://doi.org/10.1037/a0014387
54. Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 6 (Jun.

2015), 1113–1133. DOI:https://doi.org/10.1109/TPAMI.2014.2366127

55. José Serra, Verónica Orvalho, and Darren Cosker. 2016. Behavioural facial animation using motion graphs and mind maps. In *Proceedings of the 9th International Conference on Motion in Games (MIG '16)*. ACM, New York, NY, 161–166. DOI:https://doi.org/10.1145/2994258.2994270
56. Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *J. Pers. Soc. Psychol.* 52, 6 (Jun. 1987), 1061–1086. DOI:http://dx.doi.org/10.1037/0022-3514.52.6.1061
57. Judith Sinzig, Dagmar Morsch, and Gerd Lehmkuhl. 2008. Do hyperactivity, impulsivity and inattention have an impact on the ability of facial affect recognition in children with autism and ADHD? *Eur. Child Adolesc. Psychiatry* 17, 2 (Mar. 2008), 63–72. DOI:https://doi.org/10.1007/s00787-007-0637-9
58. Petr Slovák and Geraldine Fitzpatrick. 2015. Teaching and Developing Social and Emotional Skills with Technology. *ACM Trans. Comput.-Hum. Interact.* 22, 4 (Jun. 2015), 1–34. DOI:https://doi.org/10.1145/2744195
59. Andreas Sonderegger, Klaus Heyden, Alain Chavaillaz, and Juergen Sauer. 2016. AniSAM & AniAvatar: Animated visualizations of affective states. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM Press, New York, NY, 4828–4837. DOI:https://doi.org/10.1145/2858036.2858365
60. Matteo Sorci, David McCallum, and Alastair Gordon. 2011. Say it to my face! In *Proceedings of 2011 Australian Market & Social Research Society National Conference (AMSRS '11)*. AMSRS, Sydney, Australia, 1–21.
61. Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock. 2006. The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*. 423-426.
62. Alexander Todorov, Ron Dotsch, Jenny M. Porter, Nikolaas N. Oosterhof, and Virginia B. Falvello. 2013. Validation of data-driven computational models of social perception of faces. *Emotion* 13, 4 (Aug. 2013), 724–738. DOI:https://doi.org/10.1037/a0032335
63. Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.* 37 (Mar. 2010), 141–188. DOI:https://doi.org/10.1613/jair.2934
64. Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Trans. Affective Comput.* 3, 1 (Jan. 2012), 69–87. DOI:https://doi.org/10.1109/T-AFFC.2011.27
65. Justin Walden, Eun Hwa Jung, S. Shyam Sundar, and Ariel Celeste Johnson. 2015. Mental models of robots among senior citizens: An interview study of interaction expectations and design implications. *Interact. Stud.* 16, 1 (Aug. 2015), 68–88. DOI:https://doi.org/10.1075/is.16.1.04wal
66. Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. In *Proceedings of the International Conference on Computer Graphics & Interactive Techniques (SIGGRAPH '11)*. ACM Press, New York, NY, 1–10. DOI:https://doi.org/10.1145/1964921.1964972
67. Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE T. Affect. Comput.* 5, 1 (Jan. 2014), 86–98. DOI:https://doi.org/10.1109/TAFFC.2014.2316163
68. Marzanna Wiechetek Ostos, Françoise Schenk, Tania Baenziger, and Armin von Gunten. 2011. An exploratory study on facial emotion recognition capacity in beginning Alzheimer's disease. *Eur. Neurol.* 65, 6 (Jun. 2011), 361–367. DOI:https://doi.org/10.1159/000327979
69. Hui Yu, Oliver G.B. Garrod, and Philippe G. Schyns. 2012. Perception-driven facial expression synthesis. *Comput. Graph.* 36, 3 (May 2012), 152–162. DOI:https://doi.org/10.1016/j.cag.2011.12.002
70. Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 1 (Mar. 2008), 39–58. DOI:https://doi.org/10.1109/TPAMI.2008.52
71. Shuo Zhou, Timothy Bickmore, Michael Paasche-Orlow, and Brian Jack. 2014. Agent-user concordance and satisfaction with a virtual hospital discharge nurse. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents (IVA '14)*. Lecture Notes in Computer Science. Springer Verlag, Berlin, Germany, 528–541. DOI:https://doi.org/10.1007/978-3-319-09767-1_63