The Meme Quiz: A Facial Expression Game Combining Human Agency and Machine Involvement

Kathleen Tuite and Ira Kemelmacher
Department of Computer Science and Engineering
University of Washington
{ktuite,kemelmi}@cs.washington.edu

ABSTRACT

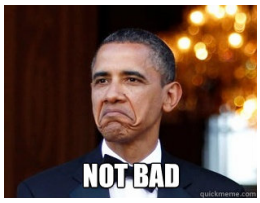
We describe a game with a purpose called *The Meme Quiz* in which a human player mimics popular Internet memes, and the system guesses which expression the player imitated. The purpose of the game is to collect a useful dataset of in-the-wild facial expressions. The game was deployed with 198 players contributing 2,860 labeled images. In contrast to many data-gathering games that use interaction between humans to define the mechanics and verify the data, our game has an online machine learning system at its core. As more people play and make faces, *The Meme Quiz* gathers more data and makes better guesses over time. One main advantage of this setup is the ability to monitor the usefulness of the data as it is collected and to watch for improvement, instead of waiting until the end of the game to process the data. Our contributions are 1) the design and deployment of a game for collecting diverse, real-world facial expression data and 2) an exploration of the design space of data-gathering games along two axes: *human agency* and *machine involvement*, including advantages of building a game around an interactive domain-specific technical system.

Keywords

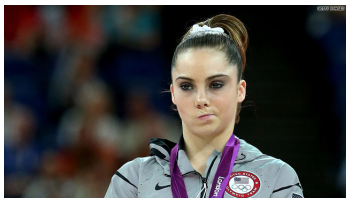
games with a purpose, facial expression recognition, crowdsourcing, computer vision, machine learning

1. INTRODUCTION

Facial expression recognition is an important part of affective computing. Typically, only the six basic expressions of joy, sadness, fear, anger, surprise, and disgust are used in affective computing – a small set of facial expressions by all accounts. We are interested in extending the capabilities of automated expression recognition and in collecting a new dataset of facial expressions that includes many new expressions. To do so, we take advantage of the broad set of facial expressions that appear in Internet memes.



(a) Not Bad Obama



(b) Not Impressed McKayla

Figure 1: Example Internet Memes portraying several different emotions. Do these emotions have obvious names, or is the picture itself a more concise way of conveying the emotion?

Reaction images [1], known by the shorthand *MRW* ("My Reaction When"), are a type of meme that portray an emotion in response to something said or experienced. "Not Bad Obama" and "Not Impressed McKayla", shown in Figure 1, are two recognizable media images that have been elevated to meme status. *MRW* memes can also include non-human faces, such as "Grumpy Cat" in Figure 6(c). These reaction images may have value beyond entertainment; Figure 2 shows a *MRW* meme known as "Success Kid" annotated with a story about a user's breakthrough using reaction memes to convey emotional state to a therapist. Communicative expression-related memes would be useful to affective computing, where a primary goal is to assist people who struggle with reading or communicating emotions in their everyday lives.

Although reaction memes themselves are popular on the Internet, there is no data source of everyday people portraying these same facial expressions. Anticipating that imitating memes would not only generate useful data but also be amusing and compelling, we set out to build a game to crowdsource photos of people imitating meme facial expressions. Since our end goal is to use the collected dataset to train an automated system to recognize these new expressions, we decided to build the training and teaching of this system into the game.

In this paper we present *The Meme Quiz*, a game we developed where the player is asked to act out a meme expression and the system guesses which meme the player is imitating. Over time as the game collects more data, the system improves and is able to guess expressions correctly.

Figure 2: A "My reaction when" meme depicting a story about using MRW memes to convey emotional state to a therapist.

We designed our game such that it does not require the expression recognition technology to work perfectly; the fun of the game comes from the fact that the system occasionally makes mistakes. In fact, our system can start learning immediately and does not need to be bootstrapped with initial data. In the middle of deployment, we were able to adapt the game by adding new memes to impersonate, and we were able to monitor the health of the data over time to make sure the system was in fact learning these novel facial expressions.

Because our game uses online machine learning as its core mechanic, it is different from other crowdsourced data generation games. In the rest of this paper, we explore the space of crowdsourced data-generation games along two dimensions of human agency and machine involvement, and describe how *The Meme Quiz* fits as a game with high agency for both the human and the computer. Our contributions are 1) the design and deployment of a game for collecting diverse facial expression data and 2) an exploration of the design space of data-gathering games along two axes: human agency and machine involvement, including advantages of building a game around an interactive domain-specific technical system.

2. RELATED WORK

This section focuses on background work related to facial expressions and crowdsourcing of these expressions. Games with a purpose are highly relevant, and we discuss many games in Section 3 on our proposed design space for data-gathering games.

Name	Subjects	Photos per Subject	Expressions
CK+	127	4 videos	6
Multi PIE	337	2,225 photos	6
MMI	90	20 videos+photos	6+AUs
AM-FED	242	1 videos	2

Table 1: Comparison of facial expression datasets

There are a number of existing facial expression datasets, such as CK+ [15], CMU Multi-PIE [7], and MMI [18], which have been laboriously captured and annotated with emotion and Action Unit (AU) labels from the Facial Action Coding System (FACS). These datasets have fueled facial expression recognition research for over a decade and Table 1 shows a comparison of these datasets.

These standard datasets are often collected in controlled lab environments with, at most, a few hundred subjects. In practical applications, face trackers must work on a wide variety of facial appearances and in many different lighting conditions, and on more than six expressions. Bigger datasets are necessary, as well as datasets captured in the real world in realistic situations, such using a webcam or a front-facing camera on a mobile phone. Our game captures faces in realistic capture conditions, and includes many more expressions.

The AM-FED [16] dataset was also captured "in the wild" by recording subjects' faces as they watched Super Bowl advertisements on their own computers. As an incentive for allowing their reactions be recorded, subjects were shown a chart of their smile activity compared to others, which is an interesting integration of computer vision back into the user experience. The expressions captured in AM-FED are spontaneous (or as spontaneous as they can be when the subjects are aware they are being recorded), but the videos were chosen to elicit joy only, so the dataset does not span a wide range of emotions, or even very extreme emotions. While our own dataset is posed, it includes the basic expressions as well as many more, captured in real world environments.

Capturing spontaneous expressions is difficult, as it requires subjecting users to unpleasant stimuli designed to elicit emotions such as disgust, fear, or pain, and different people might not be sensitive to the same stimuli. Recently, Li et. al. [14] and Yan et. al. [26] have compiled datasets of spontaneous micro-expressions using videos chosen to elicit emotions including joy, surprise, and disgust and encouraging subjects to try to hide their emotions. Zhang et. al. [27] have also captured a 3D dataset of spontaneous expressions by engaging lab subjects in different activities, such as playing an embarrassing game or experiencing an unpleasant smell. We believe acting out expressions is more fun for the participant than being subjected to unpleasant stimuli.

New datasets of labeled examples of facial expressions that span a wide variety of people, capture conditions, and emotions are critical to the future of automated expression recognition and affective computing. Especially compared to bringing subjects to act out expressions in-person, online crowdsourcing has the potential to recruit many more subjects and collect far more data. *The Meme Quiz* is one of many possible ways to realize mechanics and incentives of crowdsourcing facial data.

3. DESIGN SPACE OF DATA-GATHERING GAMES AND SYSTEMS

Before we describe our game, we want to define the design space of games with a purpose (GWAPs), specifically those used for gathering data, and position *The Meme Quiz* within that space.

Games with a purpose, such as the *ESP Game* [25], were first introduced to produce large, labeled datasets to be used as training data for computer vision and machine learning tasks. Since then, games including *BeFaced* [22] and *Motion Chain* [20] have been developed to generate new data. We will call these games *data-gathering* games, with *data-generation* games as a subset. Paid crowdsourcing through micro-task platforms like Mechanical Turk are also a common way to gather and generate data. Not all games with a purpose are data-gathering games; some like *Foldit* [2], *Phylo* [9], and *EteRNA* [13] are about solving puzzles and understanding the human process of finding optimal solutions, rather than simply collecting a dataset. *The Meme Quiz* is ultimately about collecting a dataset, but also understanding the system's learning process.

In order to understand what makes our data-generation game *The Meme Quiz* different from other data-gathering games, we examine a design space split along two different axes: **human agency** and **machine involvement**. Figure 3 shows the game design space split into the axes of the human agency and machine involvement, with the games discussed in this section.

Human Agency. How much choice does the player have? Do players get to be creative and try different options, or is the game expecting one objective right answer from them? In her book *Hamlet on the Holodeck* [17] Murray says of agency, "When the behavior of the computer is coherent and the results of participation are clear and well motivated, the interactor experiences the pleasure of agency, of making something happen in a dynamically responsive world." Games are about giving players choices, especially choices that have a clear and purposeful impact on the game world, so the games mentioned in Figure 3 all have at least some human agency. Agency can blend into creativity and artistic self-expression, with drawing games like *Picard* [23] and with Zitnick's Abstract Scene clipart-based creator[28] discussed below.

Standard non-game labeling tasks on Mechanical Turk, such as HITs from the requester "Tagasaruis" have low human agency, with tasks such as, "Determine the number of people in an image or whether the scene is indoors or outdoors." The human does have to examine the image and enter their answer, but there is a single, objective answer that is expected.

In von Ahn and Dabbish's *ESP Game* [25], a game in which one player tries to guess what another player looking at the same image would provide as a tag, there is slightly more human agency. While the tags should ultimately capture what is important in the image, the player gets to make that choice, and their success in the game is influenced by the choices of other players. The *ESP Game* is more open to gathering subjective labels from players than many labeling tasks on Mechanical Turk.

In *The Meme Quiz*, players have even more agency and responsibility. They must choose a face to imitate and then control their own face to make that expression. They can even change their hair or put on makeup to better match a face, as some of our players have done. In addition to simply providing a label, our players also provide the unique details of their own face.

Machine Involvement. The game *BeFaced* [22], a casual tile-matching tablet game, is similar to ours in that its purpose is also to collect expressions. Unlike *The Meme Quiz*, *BeFaced* only focuses on the six basic expressions. We consider *BeFaced* to have medium machine involvement, since a live face and expression tracker and runs as part of the game, but is secondary to the tile-matching mechanic. The off-the-shelf expression tracker used by *BeFaced* does not improve over time with more data. In fact, instead of the machine adapting, the players appear to have adapted to the deficiencies in the tracker, avoiding challenging faces and tilting the tablet to make the recognition work [22].

The *ESP Game* [25] on the other hand is primarily concerned with obtaining appropriate labels for images, so there is no system built into the game that actually tries to use the labels. The only machine intervention is identifying taboo words, words that have already been used to label an image, which force players to come up with additional labels.

In contrast to these examples, *The Meme Quiz* lies at the high end of machine involvement. We know that face and expression tracking do not always work, because the space of faces, expressions, lighting, and pose is much more vast than what these systems have been trained on. We chose to build a full face tracking and expression recognition system into our game, and incorporate that unreliability into the game mechanics, making *The Meme Quiz* about experiencing and improving the limitations of facial expression recognition systems. The human player has a large role in acting out expressions, but the machine is heavily involved as well, as the main mechanics of the game depend on the underlying system learning and improving over time.

Additional Examples. *PhotoCity* [24], like *The Meme Quiz*, is a game built around a computer vision pipeline. In the case of *PhotoCity*, the underlying system automatically generates 3D models from player photographs, and computes how much 3D geometry each new photo contributes. The 3D reconstruction system underlying *PhotoCity* is more complex than a face tracker, but it does not learn and improve, it merely incorporates more data into a geometric model. Players can choose where to take photos based on what is convenient or what they are interested in, but they don't have much room for creativity beyond capturing the scene as it exists.

An example of high human agency and creativity is Zitnick and Parikh's database of abstract scenes [28]. Users construct a clipart scene of a sentence like "Jenny threw the beach ball angrily at Mike while the dog watches them both" and the result is many different interpretations of the same scenario. Although there is no machine involvement during the scene creation process, the data is later analyzed to study high-level semantic information and which types of attributes are important for scene understanding.

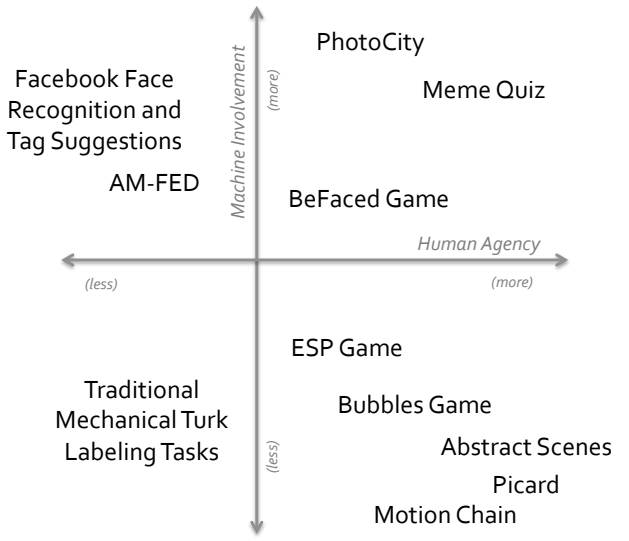


Figure 3: Design space of data-gathering and data-generation games (and other systems). Does the human gets to make choices and express creativity (high agency), or if there is ultimately one objective right answer the human is expected to provide (low agency)? Is there an automated system that processes user data and is in the domain in which the data is expected to be used (high machine involvement)? Does new data change the way the system operates, and does it learn or improve over time?

Spiro's *Motion Chain* [20] is a webcam game for collecting example gestures for gesture recognition, and also allows for human creativity and agency, when players imitate the gestures of other players. The goal of collecting real-world data through a crowdsourcing game is similar to goal of *The Meme Quiz*, but in *Motion Chain*, the mechanics are not based on any gesture recognition system consuming the data, but on players observing and imitating each other.

In the opposite quadrant but also with a very similar goal, the AM-FED [16] facial expression crowdsourcing system has high machine involvement with low human agency. Human agency is low because the participants are asked to watch an advertisement and react the way they normally would if no camera was watching their expression. Machine involvement is fairly high, though, because after the advertisement is finished, the video of the user is quickly processed and analyzed with a face detector and smile tracker, and a chart showing when they smiled is shown to the user, along with a timeline of the video and an average smile graph across many users.

Deng's *Bubbles Game* [5] for bird classification is an interesting example of a game that technically doesn't have a machine or vision system running while the user plays the game, but where the player's actions directly influence how a computer vision system later makes its decisions. In *The Bubbles Game*, a user is guessing which species of bird is shown in an image, but the image starts out blurry and in greyscale. The user can place "bubbles" on the image that make that patch of the image in color and in focus, and the goal is to identify the bird with as few bubbles as possible. From this, the system learns what parts of the image or bird are most discriminative and useful for making its own classifications.

Although it is not a game, identity recognition and tag suggestion systems, like Facebook, have high machine involvement. They can make intelligent decisions based on context (who is likely to be in a photo based on friend relationships and who is often photographed together) and improve their capabilities over time as more faces are tagged. We consider these systems to have low human agency, where there is an objective right answer of who is who in a photo, and the only user choice is whether or not to post or tag a photo in the first place.

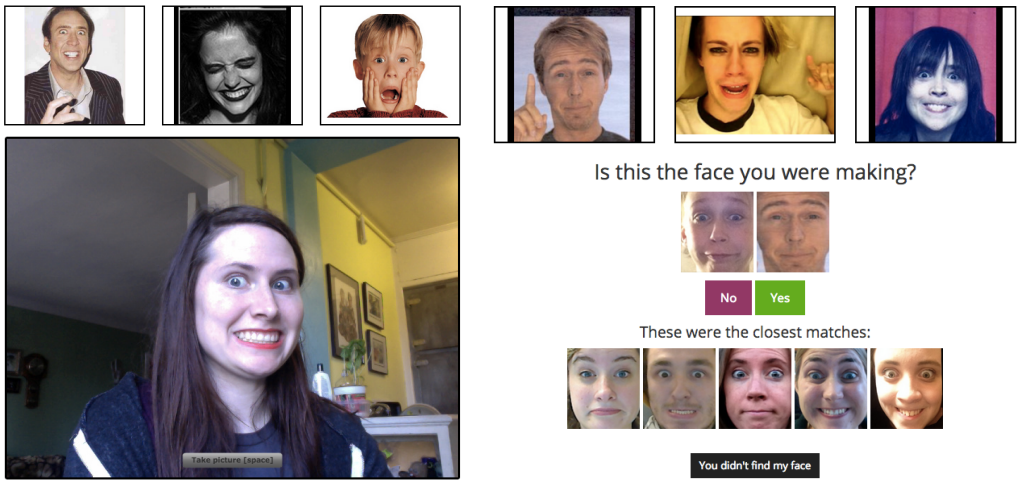
3.1 Advantages of High Machine Involvement

We observed that the design space quadrant of high human agency and high machine involvement has certain advantages. Player agency is an important aspect of game design; according to Sid Meier, "a [good] game is a series of interesting choices." [19] Machine involvement, especially AI/ML/computer vision systems integrated into games in new ways, can create new types of experiences for players. Additionally, these systems can give designers a way to monitor the health of the data as its collected, instead of waiting until the game is over to process and evaluate the data. Conveniently, we found we could adapt the data collection task on the fly by adding more meme expressions without breaking the flow of the game.

4. THE MEME QUIZ

The goal of the game is to have the system correctly guess which facial expression the player is making. The player sees three meme expressions selected at random and chooses one to imitate. The player then takes a photo of herself imitating that expression. The system analyzes the photo and makes a guess, allowing the player to tell it the correct answer. Figure 4(a) shows the acting phase, and Figure 4(b) shows the game making its guess and asking for confirmation from the player. Behind the scenes, the system does not compare the new face directly to the meme, which could be a non-human animal or drawing, but to the faces of everyone who has acted out each of those three expressions before. Figure 4(b) shows the game displaying the top five most similar faces of other players, providing insight into the system's decision and why it might have gotten confused.

When the game was launched, the system had no training examples to learn from, so it guessed randomly. As play continued, there were more faces to compare with and it became easier to find a correctly-labeled similar-looking face. There were some specific ways the system could guess incorrectly, though. For example, the system occasionally thought that a photo of a particular player looked most like that same player making a different expression. Or that the lighting was too different for the faces of otherwise close appearance and expression to seem similar. These are common mistakes for all facial expression recognizers to make, and the best solution is to collect more data to fully understand these distinctions.



(a) Imitating an Expression (b) Verifying the Game's Guess

Figure 4: The quiz interface: (Left) At the top, there are three possible expressions to imitate. Below, the user sees a live preview of their own face and can press a button to take the photo. (Right) After the player takes a photo, the game system makes its guess and shows its answer to the player, who then provides the correct answer. In this example, the game (correctly) guessed the left-most expression, but the top five closest faces include three from the (incorrect) right-most expression, suggesting that this was a difficult example.

4.1 Face Processing

Behind the scenes, the game runs through a standard facial expression recognition pipeline shown in Figure 5 and described below. First, we run the Face++ [8] face detector on the image to find the location of the face within the image and to find the location of landmarks such as eyes, nose, and mouth. Knowing the position of the landmarks, we align and crop the face to a common frame of reference where eyes, nose, and mouth are always in the same place, allowing us to easily compare different faces. Once we have the aligned and cropped face image, we turn that image into a lower-dimensional feature vector by computing Histogram of Gradients (HOG) features [4] in a grid over the face. This feature vector converts the pixel representation of the face to a representation of the shape and intensity of edges and lines in different regions of the face, which can capture changes in skin wrinkles and in mouth/eye/eyebrow shape while being impartial to skin tone and image color variations. This process of detecting, aligning, and computing low-level features to feed into comparison and classification tasks is the same used by Kumar et. al. in their work on searching within collections of faces [11] and on face verification [12].

After the face image has been transformed into a feature vector, we can compare it to other faces that have been processed in the same way. For a person imitating one of three memes, we look at all the faces also acting out those memes, find the face that is most similar to the new face, and set the system's guess to be the label for the closest face. Essentially, each face represents a point in N-dimensional space, and we are computing the new face's k-nearest neighbors (k-*NN* [3]) and checking their expression labels. The game makes its guess based on the single closest face.

As the game collects more data, there are more face points to compare to, filling out an N-dimensional face space. During the game, we use feature vectors and a nearest neighbor classification algorithm. In our evaluation, we use these same feature vectors to train linear SVM classifiers. We did not use SVMs during the live game because $k-$NN was much simpler to implement; unlike SVMs, which would have required retraining after each new face or use of an online, incremental SVM.

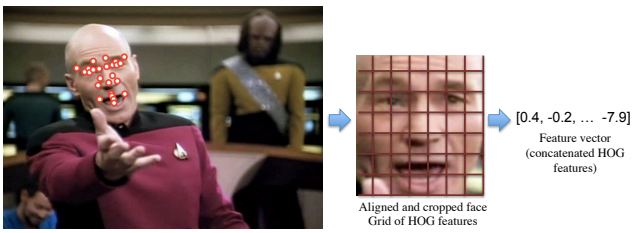


Figure 5: Every face that is uploaded to the game goes through this pipeline. First, a face detector finds locations of landmarks on the face, such as eyes, eyebrows, nose, and mouth. Using those landmark locations, we align the face to a common reference frame and crop the image to just the face region. Then we compute a grid of Histogram of Gradient (HOG) features and concatenate them together to get our final feature vector that we use for classification.

5. GAME DEPLOYMENT

We built a website and an iOS application for *The Meme Quiz* and seeded the set of expressions with about twenty popular Internet memes. Over time, we expanded the set to 298 expressions, including memes suggested by users, expressive faces from movies, celebrities making extreme ex-

pressions, and the six basic expressions. Figure 7 shows the number of memes active in the game over time, and how adding new faces impacted performance, which we will discuss more in the next section.

198 players played 2860 rounds (one photo per round), averaging 14.4 rounds/photos per player. Two-thirds of the photos (1893 out of 2860) were captured through the iOS application. There were 110 women, 61 men, and 27 people who did not indicate a gender. Players spanned a wide age range: 13-15 years: 17 players, 16-19 years: 24 players, 20-29 years: 55 players, 30-39 years: 26 players, 40+ years: 5 players, and 7 players of unknown age.

We asked users for the country in which they were born, as well as the country in which they currently lived. There is a spread, and the most common responses for current country was: United States: 98 players, unknown/decline to state: 24 players, Mexico: 8 players, Brazil: 6 players, France: 5 players, Columbia: 4 players, UK: 4 players, Russia: 3 players, Turkey: 3 players, Iran: 3 players, and so on. For birth country, the top responses were: Unknown/decline to state: 133, United States: 44, Ukraine: 3, China: 2, Korea: 2, India: 2, and so on.

Ideally, we would like to have a broad and dense sampling of expressions across all gender, age, and ethnicity permutations.

Our consent form, which every player was required to consent to before playing, allowed players ages 13 and older to participate. Ethically, collecting faces for a public research database, especially one known to contain minors, it a touchy subject. In our defense, our subjects are explicitly told that they are contributing their face to an anonymized public research database, and can opt not to participate. One alternate approach is to perform research on privately-owned user data collected by Facebook, such as the 4,000-person, 4.4-million photo dataset used by Taigman et. al. [21]. A second alternate approach is to scrape the Internet for images of emoting faces, such as the dataset collected for the 2013 Kaggle Facial Expression Recognition challenge [6], which contains over 30,000 faces. In our analysis of this dataset, we discovered that around 20% of these images were duplicates, many were mislabeled, and many were stylized stock photographs of models overacting each expression. In both of these cases, there are thousands of people who have no idea their faces are in these datasets. We value transparency and would like to see it become more common to intentionally build and contribute to computer vision research datasets.

6. EVALUATION

Our evaluation focuses on studying the system's learning process. Because of the large number of expressions included in the game, not all have enough examples to be usefully analyzed, so the evaluation in the following section will focus on the 26 expressions that have 30 or more examples. Figure 6 shows eight of these expressions and the average blend of imitations. We generated these average images by computing the average RGB pixel value for each pixel in a stack of faces. Conveniently, the face images are pre-aligned because of the preprocessing step to detect, align, and compute features, so eyes, mouth, and other landmarks line up. It is possible to compute an even sharper average blend using a technique called Collection Flow [10].

The three key questions we investigate in our evaluation are:

1. How well does the system learn over time?
2. Which expressions are learned best? Which expressions are confused?
3. How does our dataset compare to existing datasets and are we collecting useful data?

6.1 Learning Over Time

We made a game around teaching the system to recognize facial expressions, but with the challenges of so many expressions to master and so many possible facial appearances from "in the wild" players, did the system successfully learn? Instead of six basic expressions to master, the system was exposed to hundreds of new expressions (although only three at a time). It saw hundreds of different human players, each with different face shapes, hairstyles, accessories, and glasses. It was also exposed to many different lighting environments and shadows cast on faces in different directions. Would the game be able to make sense of the underlying expressions in spite of all these variations?

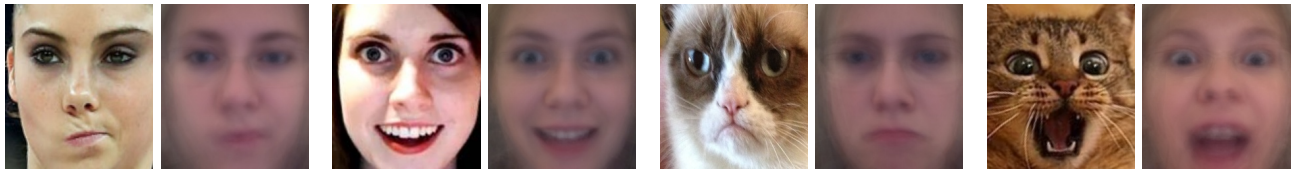
As the game deployment progressed, we tracked how often the system guessed correctly. Figure 7 shows a plot of accuracy over time. If the game were guessing randomly, it would have an accuracy of 33%, but the accuracy was significantly higher than that and increasing over time; over the course of the 2,860 rounds played, accuracy has progressively increased to 60.69%. The best results of a recent Kaggle competition on facial expression recognition [6] come in at 71.2% accuracy on six expressions.

At five points during deployment, we expanded the set of memes (these points are represented by the vertical bars in Figure 7). As can be seen, immediately following each expansion, accuracy takes a temporary dip. For example, when we added 185 new expressions, mostly celebrities making goofy faces and extreme expressions, accuracy dropped from 52% to 48% as the system had no information about these new faces. But over time, players imitated these expressions and the system was able to correctly guess them.

Being able to monitor the health of the system during the deployment and to verify that the game was actually improving at its expression recognition task was both essential and encouraging. Furthermore, our learning system was flexible enough to allow us to change what it was learning on the fly, and to rebuild its own knowledge base when we gave it completely new challenges.

6.2 Which Expressions Learned Best/Worst?

With so many memes to recognize (298 by the end of the deployment), and certain memes that either showed the same expression or looked very similar, we wanted to know which expressions were being learned most reliably. Some expressions might do worse than others if they either had very little data, or if they looked too similar to other expressions.



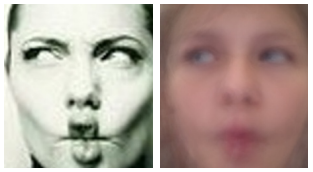
(a) Not-Impressed-McKayla (57 photos)

(b) Overly Attached Girl (43 photos)

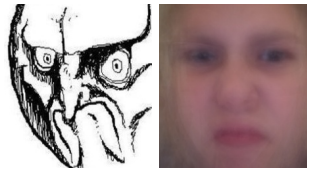
(c) Grumpy Cat (43 photos)

(d) OMG Ribbons Cat (42 photos)

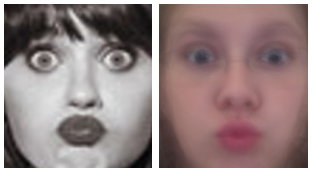
(e) Not Bad Obama (42 photos)



(f) Angelina Fish Lips (41 photos)



(g) 'No.' Rage Face (40 photos)



(h) Zooey Kiss (39 photos)

Figure 6: Example expressions and their average blends.

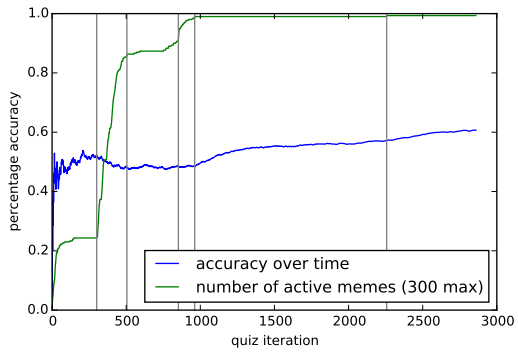


Figure 7: Over almost 3,000 quiz iterations, the game's accuracy (shown in blue) in choosing the correct expression out of three climbs from 33% (random) to above 60%. With more quiz rounds, this number would likely continue to increase. The green line shows how many meme expressions are active in the game, with the vertical bars indicating when more memes were added. Adding new memes, which the system knowns nothing about, temporarily hurts performance until the game collects enough examples to learn those memes.

Because the game only tests three expressions at a time, we are interested in evaluating the full knowledge of the system at the end of deployment. We focus on the top 26 expressions with 30 or more examples, ending up with 930 samples. Instead of the nearest-neighbor classification used online in the game, we fed these 930 faces into a 26-class SVM, which we then used to predict the expression label of a new face. We randomly split our dataset into training and testing sets (66% of the data in the training set, 33% in the remaining testing set) and repeated this training/testing ten times, then averaged the results.

Overall, expressions were correctly predicted 38.89% of the time. Note that this prediction is choosing from twenty-six labels (thus a baseline accuracy of 3.8%), not just three labels as in the game itself.

Meme A	Meme B	Percent
Borat	Bunny Teeth	21.5%
Bunny Teeth	Borat	21.2%
Borat	Overly Attached Girlfriend	19.8%
About to Interrupt	Bro Paul Ryan	19.2%
Crazy Eyes Nick Cage	Happy Kenneth	17.4%
Liz Lemon Eye Roll	Grumpy Cat	17.1%
No.	Not Bad Obama	16.8%
Grumpy Cat	Not Bad Obama	16.6%
Mr. Bean	Zooey Kiss	16.6%
OMG Ribbons cat	Einstein Tongue	16.2%
No.	Grumpy Cat	15.9%
Crazy Eyes Nick Cage	Overly Attached Girlfriend	14.6%

Table 2: Expressions commonly confused with other expressions over 14% of the time.

The confusion matrix shown in Figure 8 illustrates which expressions were correctly predicted and which were classified as different expressions. The dark diagonal shows where expressions were successfully classified. Dark squares off axis indicate expressions that were confused. The most frequently confused expressions are listed in Table 2. Many of these confusions make sense; Grumpy Cat and Not Bad Obama expressions both have a similar appearance of the lines around the mouth coming from Grumpy Cat's frown and Obama's mouth shrug. Borat and a picture of Ellen Page making "bunny teeth" also look similar and are understandably confused.

In the future, we imagine our data being used to distinguish between similar-looking expressions, e.g. learning to look at the eyebrows to differentiate "grumpy" and "not bad". We intend to perform a hierarchical classification of expressions, and compare it to the six basic expressions and the FACS basis space of expressions based on face muscles.

6.3 Comparison with CK+

The facial expression dataset from Cohn-Kanade (CK+) [15] is a standard in facial expression recognition research, cap-

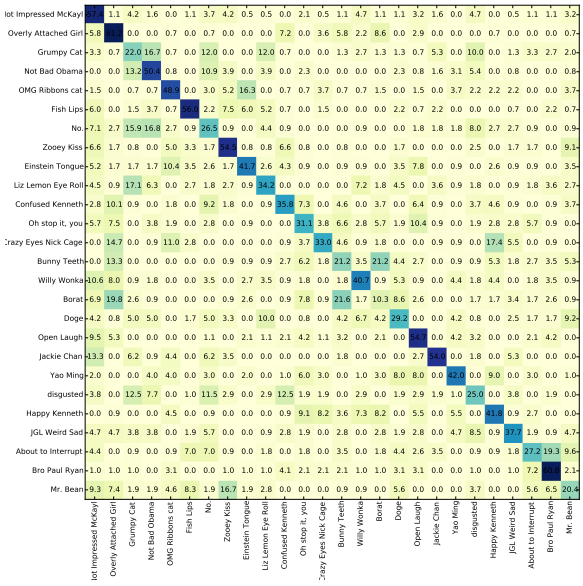


Figure 8: We trained and tested (using cross-validation) a multi-class SVM on the 26 expressions. (Each expression has between 30 and 57 training examples.) This confusion matrix shows which expressions were commonly categorized as other expressions, such as Grumpy Cat with Not Bad Obama. The confusion matrix is not symmetric.

turing 127 subjects making the six basic expressions of joy, sadness, anger, fear, surprise, and disgust. We included photos of people making these expressions in our prompt set to directly compare with CK+.

For the comparison we used a subset of our data – the 317 photos with the six expressions in CK+ – and compared it to the 327 photos of these expressions in CK+, not including neutral expressions. For each dataset, we trained a classifier for each expression, and then classified faces from the opposite dataset. We found that when training classifiers using CK+ data, we were only able to recognize 48.58% of *Meme Quiz* faces. In contrast, when we use *Meme Quiz* data as the training set, we are able to recognize 75.84% of CK+ faces. This means that CK+ data is not diverse enough (many of the faces have the same lighting) to serve as training data for real world scenarios such as ours. In contrast, this small subset of our data, with its real-world faces captured in diverse lighting environments, still works to predict standard examples of these expressions fairly reliably.

7. LIMITATIONS AND FUTURE WORK

7.1 Verification

The goodness of our system relies on (most) players making their best effort to imitate each expression and providing the correct answer at the end of each quiz round. The goal of the game, whether or not players choose to play this way, is to teach the system enough about each expression to have it correctly guess as often as possible. However, players can lie, or they can game the system and attempt to teach the system incorrect concepts. We have no explicit verification method for dealing with scenarios like this, and can only hope there is enough good data to treat misinformation as outliers.

We could use humans to verify each others' faces by having them act as the 'computer' and guess which expressions other players are making. This could give us useful information about which expressions are incorrectly labeled, which players are not good actors, or which players are not good at reading expressions.

An alternate, partially automated solution would be to use the trained expression classifiers to look for outliers and present those to a human for verification (either inside or outside of the game context). Essentially, the game's technology has a built-in tool for identifying the most confusing or questionable faces, which would simplify the verification task for the human.

7.2 Engagement

We struggled with questions of how to engage users. What would make people want to take photos of themselves? What would make them keep playing? Could we generate a compelling artifact, e.g. a picture of their face swapped with a meme, that they could share with their friends to draw new users to the site? Would the idea of "teaching the computer" be compelling enough, or would it be too nerdy?

We tried many different approaches, including generating swaps and sharable artifacts (which did draw new people to the site) but we were not rigorous enough in our studies to draw any conclusions. However, despite the numerous surface changes we made, the core mechanic of *The Meme Quiz* remained constant, and the system kept on learning.

8. CONCLUSION

We have presented *The Meme Quiz*, a game for teaching computers to recognize facial expressions by having players imitate Internet memes and quiz the computer. We have successfully evaluated the learning process of our game, confirming that, despite the wide variety of expressions and faces it was seeing, it did improve over time, even when we adapt the game to include more memes. Furthermore, the expressions it confused were often similar-looking and point to strategies one could employ in the future to help distinguish similar expressions. Compared to existing datasets of facial expressions, our data are more diverse and potentially provide a better training basis than the existing datasets themselves.

We have situated our game in a design space of data-gathering games, which we have split along two axes of *human agency* and *machine involvement*. Our system has high human agency with players acting out expressions of their choice, as well as high system involvement, with a face recognition system evaluating each face and learning as it goes. Observed benefits of being in the high-human-agency, high-machine-involvement quadrant include being able to monitor the health of our data over time, as well as make adaptations to the game without impeding the system's learning process.

9. ACKNOWLEDGMENTS

The authors would like to thank Daniel Avrahami for his extensive editing help and for his encouragement to see this research through to publication. This paper would not have made it to FDG without Daniel.

10. REFERENCES

[1] J. Asuncion. Know your meme: Reaction images. `http://knowyourmeme.com/memes/reaction-images`. Accessed: 2015-04-24.

[2] S. Cooper, A. Treuille, J. Barbero, A. Leaver-Fay, K. Tuite, F. Khatib, A. C. Snyder, M. Beenen, D. Salesin, D. Baker, and Z. Popović. The challenge of designing scientific discovery games. In *FDG '10: Foundations of Digital Games*. ACM Request Permissions, June 2010.

[3] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[5] J. Deng, J. Krause, and L. Fei-Fei. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 580–587. IEEE Computer Society, 2013.

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 2014.

[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[8] M. Inc. Face++ research toolkit. www.faceplusplus.com, Dec. 2013.

[9] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, J. Waldispühl, et al. Phylo: a citizen science approach for improving multiple sequence alignment. *PloS one*, 7(3):e31362, 2012.

[10] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection flow. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1792–1799. IEEE, 2012.

[11] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *Computer Vision–ECCV 2008*, pages 340–353. Springer, 2008.

[12] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[13] J. Lee, W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, A. Limpaecher, S. Yoon, A. Treuille, and R. Das. Rna design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 111(6):2122–2127, 2014.

[14] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[16] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected" in-the-wild". In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 881–888. IEEE, 2013.

[17] J. H. Murray. *Hamlet on the holodeck: The future of narrative in cyberspace*. Simon and Schuster, 1997.

[18] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[19] A. Rollings and D. Morris. *Game Architecture and Design with Cdrom*. Coriolis Group Books, 1999.

[20] I. Spiro. Motion chain: a webcam game for crowdsourcing gesture collection. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 1345–1350. ACM, 2012.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014.

[22] C. T. Tan, H. Sapkota, D. Rosser, and Y. Pisan. A game to crowdsource data for affective computing. *Proceedings of Foundations of Digital Games*, page 11, 2014.

[23] K. Tuite, T. Pavlik, S. B. Fan, T. Robison, A. Jaffe, Y.-E. Liu, E. Andersen, and S. Tanimoto. Picard: A creative and social online flashcard learning game. In *Proceedings of the International Conference on the Foundations of Digital Games*, FDG '12, pages 231–234, New York, NY, USA, 2012. ACM.

[24] K. Tuite, N. Snavely, D.-Y. Hsiao, N. Tabing, and Z. Popović. PhotoCity: training experts at large-scale image acquisition through a competitive game. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Request Permissions, May 2011.

[25] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Request Permissions, Apr. 2004.

[26] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu. Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[27] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[28] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3009–3016. IEEE, 2013.