To appear in ACM TOG 33(6).

Mirror Mirror: Crowdsourcing Better Portraits

Jun-Yan Zhu[1] Aseem Agarwala[2] Alexei A. Efros[1] Eli Shechtman[2] Jue Wang[2]

University of California, Berkeley[1] Adobe[2]

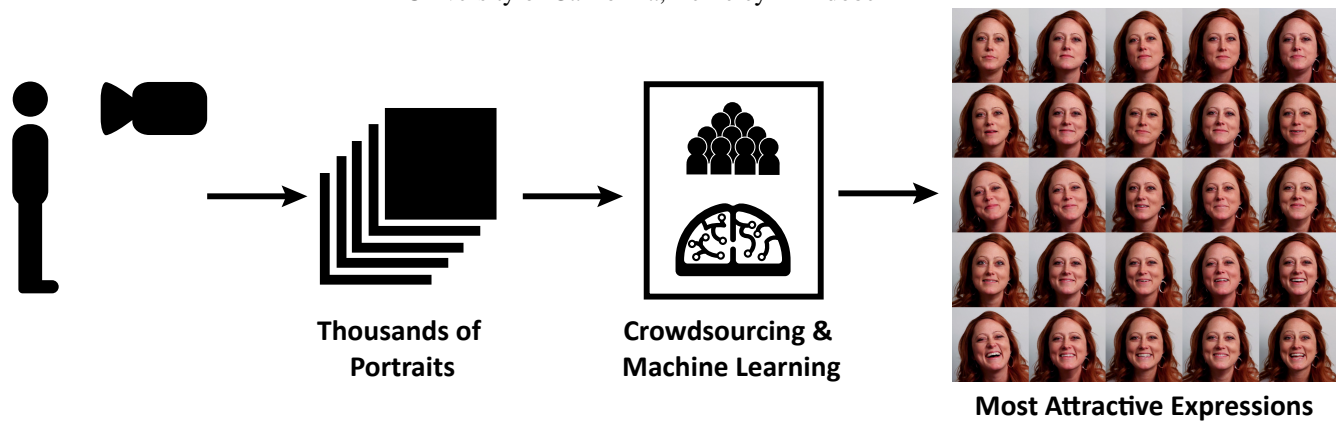


Figure 1: *We collect thousands of portraits by capturing video of a subject while they watch movie clips designed to elicit a range of positive emotions. We use crowdsourcing and machine learning to train models that can predict attractiveness scores of different expressions. These models can be used to select a subject's best expressions across a range of emotions, from more serious professional portraits to big smiles.*

Abstract

We describe a method for providing feedback on portrait expressions, and for selecting the most attractive expressions from large video/photo collections. We capture a video of a subject's face while they are engaged in a task designed to elicit a range of positive emotions. We then use crowdsourcing to score the captured expressions for their attractiveness. We use these scores to train a model that can automatically predict attractiveness of different expressions of a given person. We also train a cross-subject model that evaluates portrait attractiveness of novel subjects and show how it can be used to automatically mine attractive photos from personal photo collections. Furthermore, we show how, with a little bit ($5-worth) of extra crowdsourcing, we can substantially improve the cross-subject model by "fine-tuning" it to a new individual using active learning. Finally, we demonstrate a training app that helps people learn how to mimic their best expressions.

CR Categories: I.3.8 [Computer Graphics]: Applications—;

Keywords: crowdsourcing, portraits, aesthetic visual quality assessment

1 Introduction

Human faces are one of the most common subjects of photographs. Unfortunately, many of us feel anxiety when a camera is pointed in our direction. What should I do to look good? Will my smile look attractive or awkward? We have all experienced the disappointment of not looking our best in other people's photos. While models and actors are taught how to look good when a camera is pointed at them, the rest of us suffer from a lack of feedback; we simply don't know which of our expressions look good to other people. Self-perception in a mirror can be misleading; the image is horizontally flipped, but more importantly, our perception of ourselves is often very different than that of others [Springer et al. 2012] since our perception is influenced by our self-image and internal emotions.

There are a number of approaches to editing and improving faces in photographs as a post-process [Leyvand et al. 2008; Joshi et al. 2010; Yang et al. 2011]; however, we often do not have control of photographs taken by others and posted publicly, and many people are not comfortable with the idea of manipulating expressions in photographs. Instead, our goal is to help people look better in photographs *at the time they are taken.* Specifically, our method offers users feedback on how their range of facial expressions are perceived by others, so that they can be better prepared when a camera is pointed at them. Our method can also be used to select the most flattering pictures of people from a photo collection or video.

Our approach begins by capturing a user's range of facial expressions that are appropriate for portraits. We capture a video of the user while they are shown a twelve minute compendium of videos selected to elicit a range of neutral and positive emotions [Gross and Levenson 1995]. We then use a novel data-driven computer vision model that automatically predicts the scores of the expressions along two axes: attractiveness and seriousness. (We include the serious attribute so that users can see their best expressions across a range of scenarios, from big smiles in social settings to more neutral expressions for professional portraits.) While this method provides a reasonable approximation of the scores of a user's expressions, it cannot capture all the subtle differences between expressions and variation among users. We therefore also describe a novel crowdsourced, active learning scheme to both customize our model to the user's data and select the user's top expressions across a range of seriousness levels. This active learning scheme reduces the cost of data collection by an order of magnitude over random sampling, to about $5.

We provide a number of interfaces and visualizations to inform the user of the results of our models. The first visualization simply shows the user their most attractive expressions across twenty five levels of seriousness (Figures 1,4). Next, we offer a number of tools to explore and visualize the data more deeply. For example, the user can select an expression and suggest a change, e.g., opening the eyes more widely, and see a similar expression with more open eyes and the corresponding change in attractiveness score. The user can also visualize the differences between slices of the data, e.g., the difference between the most and least attractive expressions that

contain open eyes. Finally, we also provide an expression training application, called "Mirror Mirror", for practicing expressions in front of a webcam. The user can see their attractiveness and seriousness scores in real-time, and can practice mimicking their best expressions by selecting one and using a visualization that cross-fades between aligned versions of the current and selected expressions.

We test our method on input videos of eleven subjects, and numerically evaluate our methods on hold-out data. We also include a demonstration of the training app to show that subjects can use it to mimic selected expressions. Finally, we apply our method to select the most attractive expressions of a subject from videos downloaded from the internet, as well as personal photo collections.

2 Related Work

The perception of facial expressions is a well-studied topic [Calder et al. 2012]. The diversity of facial expressions are organized by the Facial Action Coding System (FACS) proposed by Ekman and Friesen [1978]; each action unit describes a specific facial motion (e.g., "cheek raiser") and its underlying muscular basis. More recent work [Du et al. 2014] suggests that there are an even larger range of facial expressions than those encoded by FACS. Of particular interest to our application is the difference between an insincere, voluntary smile and a spontaneous smile, which adds a slight narrowing of the eyes. Studies show that a small percentage of people are able to fake spontaneous smiles (also known as "Duchenne smiles") [Krumhuber and Manstead 2009; Gunnery et al. 2012], which should yield better portraits. The muscular differences in other subtle smile variations (e.g., amused, polite, nervous) have also been observed [Ambadar et al. 2009].

Another area of related research is the differences in social judgments elicited by different faces. Oosterhof and Todorov [2008] algorithmically generate different face shapes and measure their perceived traits (attractive, trustworthy, etc.) as scored by humans. They find that most traits approximately lie in a two-dimensional space that can be modeled as a linear combination of two principal components: valence and dominance. We instead model differences of traits between expressions of a single person, and we choose axes that are more relevant to our application (attractive and serious). However, our experiments also show that other traits that may be desirable in a portrait (e.g., trustworthy, confident) are strongly correlated to our chosen axes. Predicting, ranking, and improving the attractiveness or memorability of the faces of different people is a common research topic [Leyvand et al. 2008; Kagian et al. 2008; Gray et al. 2010; Yang et al. 2011; Altwaijry and Belongie 2013; Khosla et al. 2013]. We instead focus on the attractiveness of different expressions of *the same person*.

There is significant work in the computer vision literature on the automatic recognition of facial expressions [Pantic and Rothkrantz 2000]; most of this work focuses on FACS recognition. In contrast, Dibeklioglu et al. [2012] predict whether a portrait contains a genuine Duchenne smile. Both Shah and Kwatra [2012] and Albuquerque et al. [2008] identify smiles from multiple portraits for the purposes of selecting or generating better photographs. None of these techniques can provide a continuous rating of attractiveness of the various facial expressions of an individual. Fiss et al. [2011] select facial expressions from a video stream that best serve as candid portraits. However, they optimize for portraits that convey the moment, and many of the selected expressions are not attractive. Also, their method requires temporal features such as optical flow, and cannot be used on photo collections, which we demonstrate in Section 8. Finally, our approach to using crowdsourcing to collect ranking and scoring data for subjective attributes of images is inspired by Parikh and Grauman [2011], and similar to recent work on font attributes [O'Donovan et al. 2014] and fashion style [Kiapour et al. 2014].

Figure 2: *Left: our video capture set-up. Subjects watch videos (played by an iPad on top of a camera) while we record them. Right: example subject expressions.*

3 Overview

Our system has a number of components that can be organized into two main steps: training and testing.

Training: We begin by collecting a large set of aligned and white-balanced images of unique facial expressions for 11 subjects (Section 4). The first step is to score each image along two attributes: attractiveness and seriousness. We use crowdsourcing to collect randomly-sampled pairwise comparisons for each subject and attribute (Section 4.3), and then perform MAP estimation to compute attribute scores for each image of each subject (Section 5.1). Since we are particularly interested in accurate ranking of the most attractive expressions across different levels of seriousness, we collect additional crowdsourced pairwise comparisons for the highest scoring expressions and re-estimate scores to obtain an even more accurate ranking (Section 5.1.2). These scores for a single subject are used to train a single-subject regression model (Section 5.2) that can estimate attribute scores for an image of the same subject. The model takes as input features of a single image (computed in Section 4.2), and can operate on previously unseen images of the subject. Finally, we take the scores for all 11 subjects and train a cross-subject regressive model that can operate on images of any subject (Section 5.3). This model is more general since it can score a new person's expressions without any additional crowdsourcing; however, it is less accurate than the single-subject model.

Testing: Our system offers a number of applications, such as expression training (Section 6) and visualization (Section 7), for subjects that are not in our training data. For some applications (e.g., Figure 17), we can simply use the cross-subject model to compute attributes. In situations requiring higher accuracy, we first collect images of the new subject's expressions, and use the cross-subject model to compute baseline attribute scores. We use the seriousness scores as-is, since the cross-subject model is accurate enough for this attribute. For attractiveness we use an active learning scheme (Section 5.4) to collect a small number of crowdsourced pairwise comparisons. During this step we re-estimate attractiveness scores for each of the subject's images using both the pairwise comparisons and the cross-subject model as a rough prior. Finally, we train an improved single-subject model from the new scores.

4 Collecting Portrait Data

Our first goal is to collect a set of portrait expressions of a subject and rate them along attributes that provide useful feedback for portrait posing. However, we first need to determine the range of expressions we wish to capture, and select criteria for good portraits. Clearly, attractiveness is a common goal in most casual portraiture.

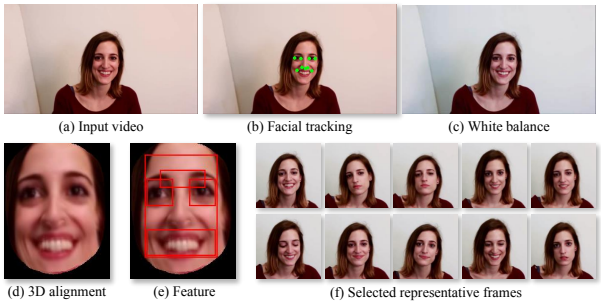


Figure 3: *We pre-process the input video to align the faces, compute features, and reduce data redundancy.*

Also, while most work on facial expression analysis [Ekman and Friesen 1978; Oosterhof and Todorov 2008] include negative attributes like anger and sadness, these attributes are generally not desired in contemporary portraits. We therefore restrict our focus to positive attributes. Along with attractiveness there are a number of positive attributes for portraits; for example, we may wish a professional portrait to appear confident, or a sales person may wish to appear trustworthy.

In initial experiments, we collected measurements on portraits for attractive, confident, and trustworthy attributes. However, like previous work [Oosterhof and Todorov 2008], we found these attributes to be highly correlated, and therefore redundant. Oosterhof and Todorov show that most attributes can be represented as linear combination of two attributes: valence and dominance. Valence is roughly parallel to attractiveness, while dominance is roughly parallel to aggressiveness. We therefore kept the attractive attribute, and chose to add a second attribute that is parallel to aggressiveness but also useful for our portrait application. We found that the highest rated portraits for attractiveness consistently had large smiles; however, it is also useful to be able to pose well for more neutral expressions without large smiles. We therefore added the "serious" attribute, since it is both a useful control for smile strength, and is nearly parallel to aggressiveness.

In the rest of this section, we first describe how we capture portraits that span a range of positive expressions. Next, we pre-process the portraits to normalize their position and color, extract image features used for predicting attribute scores, and eliminate data redundancy. Finally, we use crowdsourcing to collect pairwise comparisons of portraits along the attractive and serious attributes.

4.1 Collecting a Personal Portrait Dataset

We start by collecting a large range of positive facial expressions that may be appropriate for portraits for each subject. We hand-edited together a 12-minute compendium of short videos that ranged across several categories, including funny, scientific, and inspirational topics. The video is shown on an iPad mounted directly above a SLR camera capturing video, so that it appears the subject is looking at the camera (Figure 2). We also asked the subject to make their best portrait expression in several posed categories, such as confident, big open-mouthed smile, etc. Video is often used to elicit emotions for facial analysis [Gross and Levenson 1995; McDuff et al. 2012]. An alternative is to engage in a conversation with the subject [Fiss et al. 2011]; however, mouth motions can make stills unsuitable for portraits. In total, we collected the data of 11 subjects including both male and female subjects ranging in age from 23 to 50.

4.2 Pre-Processing

We perform several pre-processing steps (Figure 3) for each captured video to align the facial data, compute facial features and reduce data redundancy.

Facial tracking and pose normalization: We first perform tracking and pose alignment to place the face in a common reference frame. We use a recently developed face tracker [Xiong and De la Torre 2013] that accurately estimates nine facial feature points and localizes different facial parts such as eyes, mouth and nose (Figure 3b). We apply a median filter with a window size of 5 frames to smooth the estimated points and suppress tracking temporal jitter. Then we align the tracked face to a 3D template model released by [Zhang et al. 2004]. In particular, we estimate a 3D-to-2D transformation matrix between the pre-annotated 3D points in the 3D model and the detected 2D facial points using least squares. Finally, we warp the 2D face into a frontal view (174×224) using the computed transformation matrix. We exclude frames for which the tracker reports tracking failures.

Feature extraction: We extract HOG (Histogram of Oriented Gradients) [Dalal and Triggs 2005] features to capture visual properties of facial expressions in different parts of the face at different scales. Figure 3e shows five bounding boxes we use for HOG extraction, which capture two eyes (4×6 cells), eyebrows and wrinkles (2×6 cells), the mouth (2×6 cells) and the whole face (8×6 cells). The cell size for HOG is 8 pixels. Combining features of different parts results in a 3720-dimensional feature vector.

Select representative expressions: Each video typically contains around $16,000$ frames with highly redundant sampling of common expressions; collecting ratings for each frame is impractical. Therefore, we implement a simple greedy algorithm to select unique expressions from the input video. The algorithm starts by randomly selecting a frame I_i from the video, and then removes any other frame I_j which is very similar to the current frame (i.e., $D(I_i, I_j) > T$ where $D(\cdot, \cdot)$ is an appearance similarity function between two expressions and T is a threshold). After the first iteration, we repeatedly select another random frame and remove duplicates until all frames have been processed. The similarity function $D(I_i, I_j)$ is a weighted dot product between the HOG vectors of frames I_i and I_j (after first centering and whitening the HOG vectors [Hariharan et al. 2012]). As in previous work [Kemelmacher-Shlizerman et al. 2011], we weigh the mouth regions four times as strongly as the other features. We set the threshold T by binary search with the goal of extracting 200 to 250 unique expressions, which we observe empirically to be a good range for avoiding duplicates while avoiding the elimination of subtle but significant facial expression differences. Figure 3f shows several examples of the remaining frames.

White Balance: Some of our videos are not properly white balanced. To reduce the distortion in color space, we white-balance the selected representative frames using Adobe Lightroom before we collect the annotation data.

4.3 Crowdsourcing Pairwise comparisons

We next collect human response data that allows us to score the unique expressions along the attractive and serious axes for each portrait subject. We use Amazon Mechanical Turk to collect pairwise comparisons (e.g., "Is expression A more attractive than B?"). Pairwise comparisons are a common approach [Tsukida and Gupta 2011] to collecting subjective scoring data since it is much harder for people to provide absolute scores.

We use separate MTurk HITs (Human Intelligence Task) for at-

Figure 4: *Visualizations of the most attractive expressions for three subjects across a range of seriousness (the upper-left is the most serious, the lower-right the least, and seriousness decreases in reading order; attractiveness scores are shown in red). The frames are automatically selected from 12 minutes of video using a combination of crowdsourcing and machine learning.*

tractive and serious attributes, and each HIT only includes portraits from one subject. We provide instructions with two examples of labeled pairwise comparisons from a subject not used in our experiments. Each HIT includes two control questions with obvious answers, along with fourteen unknown comparisons. We discard HITs with incorrect obvious answers, and ban users who fail more than 25% of these tests. No single worker is allowed to complete more than 20 HITs. We pay $\$0.06$ per HIT. Our system always uses this structure for generating HITs; however, we sample expressions to form pairwise comparisons in different ways (random and active) and at different scales in different parts of our system. We discuss this sampling in the next section.

5 Portrait Evaluation

One of the main goals of our system is to output a visualization of the subject's best portrait expressions from a very large input collection of portraits, such as the frames of a video. Our visualization (Figure 4) shows the most attractive expressions across 25 discretized seriousness levels; seriousness scores decrease from the upper left to the lower right in reading order (left-to-right, top-to-bottom), and the most attractive image within each seriousness level is shown. These images can be used directly, or the user can select one and use our training app to learn how to mimic its expression.

Supporting these goals requires two types of portrait evaluation. First, we need a function that can score a portrait for both its attractiveness and seriousness. This score is shown to the user in our expression training app, and could be used to identify the best moment to trigger the shutter on a camera. Second, we need a method to select the most attractive portraits from a large set, i.e., rank them by attractiveness. This ranking is used to visualize a subject's best expressions, and could be used to select the best stills from a video. A ranking can trivially be derived from a scoring function; however, for our application there is a difference in accuracy requirements. For our ranking, the relative ordering of two non-attractive expressions is not important; instead, we want high confidence in our ranking of the top few expressions. At the same time, the scoring function should be reasonably accurate for any portrait.

To accomplish these goals, our method begins by first computing scores for the representative expressions chosen in Section 4.2 using crowdsourced pairwise comparisons. We then use these scored images to compute both single-subject (Section 5.2) and cross-subject (Section 5.3) predictive models. Finally, using the cross-subject model as a rough prior, we learn a more accurate single-subject model with an *active learning* scheme that selects a small number of pairwise comparisons that most increase ranking accuracy (Section 5.4).

5.1 Scoring Representative Expressions

We estimate attractiveness scores $A = \{a_1, ..., a_n\}$ and seriousness scores $S = \{s_1, ..., s_n\}$ for each of n representative expressions. We denote the pairwise comparison annotations as a count matrix $C = \{c_{i,j}\}$, where $c_{i,j}$ indicates expression I_i is preferred over expression I_j by $c_{i,j}$ times. We use the Bradley-Terry model [1952], which models the probability of choosing I_i over I_j as a sigmoid function of the score difference between two expressions, i.e., $P(I_i > I_j) = f(a_i - a_j)$ where $f(u) = \frac{1}{1+exp(-u/\sigma)}$. The scores can be estimated by solving a maximum a-posteriori (MAP) problem [Tsukida and Gupta 2011]

$$A^* = \arg\min_A \left(-\log \Pr(C|A) - \log(\Pr(A))\right), \qquad (1)$$

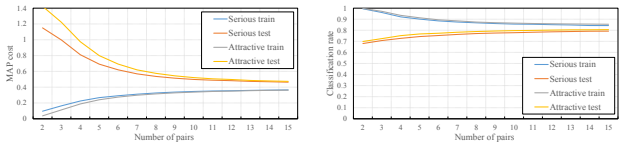
where $-\log \Pr(C|A)$ is the negative log likelihood of the pairwise comparison data given the model, and $-\log(\Pr(A))$ is a model prior term. For now we assume A is a uniform distribution; we improve this prior in Section 5.4. We can therefore rewrite Equation (1) as

$$A^* = \arg\min_A -\sum_{i,j} c_{i,j} \log(f(a_i - a_j)). \qquad (2)$$

We solve this equation using gradient descent (with σ in $f(u)$ set to 1), and then normalize scores to $[0, 1]$ for each subject. The same method is used to estimate seriousness scores S.

5.1.1 Convergence

We need to collect enough pairwise comparisons per subject so that the minimization of the MAP energy in Equation 2 converges to its minimum. As in previous work [O'Donovan et al. 2014], we find that convergence occurs in a linear rather than quadratic number of pairwise comparisons. To determine the actual number required, we reserve 5 pairwise comparisons per expression as hold-out test data, and vary the number of randomly sampled training pairs per expression from 2 to 15. (Note that one pair compares two expressions, so 15 pairs means that we sample $15 \times 2 \times n$ expressions



(a) MAP cost (b) Classification rate

Figure 5: *Convergence of (a) (MAP minimization in Equation 2 and (b) classification rate with varying numbers of training pairs per image, for both training and testing data, and serious and attractive attributes.*

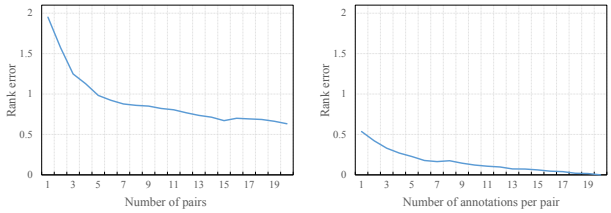
in total, i.e., each expression is seen 30 times.) We evaluate this convergence test on three subjects and report the average MAP cost and the classification rate (percentage of pairwise comparisons correctly predicted) as a function of the number of training pairs (2 to 15). The MAP cost is reported for both testing data (the 5 pairs held-out) and the portion of training data used. As shown in Figure 5, both metrics converge after about 10 pairs per expression.

5.1.2 Ranking

We can use the scores to rank and select the most attractive expressions across a range of serious levels, as in Figure 4. However, MAP convergence does not necessarily mean that the scores are accurate enough to select the best expressions. To explore this question, we first define a *rank error metric* that measures the success of a selection algorithm. We assume the seriousness score of each expression is known, and there are K serious levels (each level is a range of serious values, as computed in Section 5.5). Given a "correct" attractiveness ranking within each serious level we can compute the deviation from this ranking as $\frac{1}{K}\sum_{k=0}^{K}(\pi_k - 1)$, where π_k is the rank of the chosen expression in the k'th serious level in the "correct" ranking. This equation takes the mean of the difference of the rank of the chosen expression (which is 1) and its correct rank π_k. This metric is only concerned with the highest-rated expression in each serious level, since this is the only image shown in our target visualization.

Unfortunately, it is impossible to know whether we have collected enough pairwise comparisons from the crowd to know the "correct" ranking. We therefore generate a baseline ranking as follows. First, we randomly sample 20 pairs per expression for both attractiveness scores and seriousness scores. With this sampling, the MAP error has converged, but the rank error may not have. We therefore generate additional samples that can fine-tune the ranking. We fix the seriousness scores, since these are only used to place expressions into 25 levels, and discard all but the top 10 expressions in each bin. For each pair of these 10 expressions in each bin, we collect an additional 20 pairwise comparisons. That is, we collect 20 redundant opinions for each possible pair. We then re-rank the expressions using this data and our MAP minimization (Equation 2). We show rank error relative to this correct ranking in Figure 6, both for the initial random-sampled comparisons, and the ranking refinement. We can see while 20 random samples is enough to minimize MAP, it does not minimize rank error. Rank error is reduced to around 0.1 after 10 refinement samples, which means that 9 out of 10 visualization expressions are correct. We show the top and bottom ranked attractive/serious expressions for multiple subjects in Supplemental Materials.

This method of generating a "correct" ranking is expensive: $87.8 per subject. We therefore collect this data for only three subjects, as a reference for comparing more efficient methods. In Section 5.4, we show how an active learning scheme can reduce this cost to about $5.



(a) Rank error (random sampling) (b) Rank error (refinement)

Figure 6: *Rank error convergence from method in Section 5.1.2. (a) Mean rank error after varying the number of randomly-sampled pairs per expression. (b) Mean rank error with different numbers of additional pairwise comparisons per expression within each serious level.*

	attractive corr	attractive error	serious corr	serious error
SVR	0.88	0.064	0.90	0.060
GBR	0.88	0.064	0.89	0.063

Table 1: *Accuracy of the single-subject regression model, reported as correlation and mean absolute error, for two regression methods.*

5.2 Single-subject predictive model

Now that we have scores for the representative expressions of a single subject, the next step is to build a model that can predict attractiveness and seriousness scores for new photos of the same subject. We train a subject-specific regression model that predicts scores from facial appearance. We use the HOG features described in section 4.3, and treat scores estimated in Section 5.1 as ground-truth. We experimented with two popular regression models — Support Vector Regression (SVR) [Smola and Schölkopf 2004] and Gradient Boosted Regression Trees (GBR) [Friedman 2001] — and evaluate both methods on all the 11 subjects using 10-fold cross-validation where each fold has 20 to 25 test images. We report correlation and mean absolute errors in Table 1. The two methods produce similar results, and we use SVR since it is more efficient. We also tried adding tracking landmark point coordinates (normalized by face size) to our feature vector, as suggested by Khosla et al. [2013], but found that it barely boosted prediction performance.

A natural criticism of our approach is that smile and open-eye detectors could be adequate for predicting attractive expressions. To explore this question we use an off-the-shelf smile detector [Jiang et al. 2011], and build our own open-eye detector using the facial tracker landmarks by taking the mean distance of the two points on top of each eye from their corresponding points on the bottom. Larger distances correspond to open eyes; we experimentally confirmed that this metric works well. We train an SVR on our score data using only the 2-dimensional output of the smile and open-eye detector, combined. Its correlation with the correct attractiveness scores is only 0.47, indicating that our model (with correlation 0.88) is understanding much more than smile size and blinks. The smile detector output has a -0.51 correlation with seriousness, so it is somewhat effective at modeling that attribute.

Our smile and open-eye detectors may not be state-of-the-art; we simulate "ideal" detectors by manually selecting expressions with open eyes and smiles. We show a histogram of the expressions by attractiveness score in Figure 7. We can see that while open-eye and

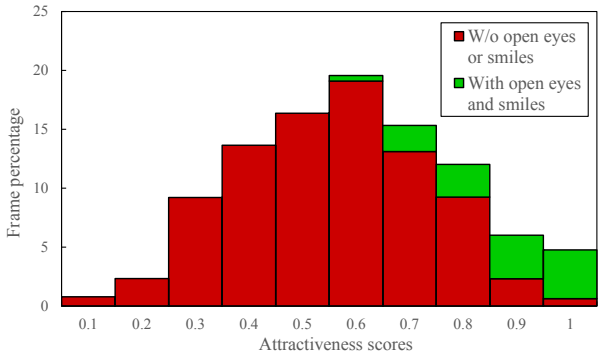


Figure 7: *Attractiveness scores for three subjects, discretized into 10 bins. Green portions of the histogram indicate open eyes and smiles; the red are the rest.*

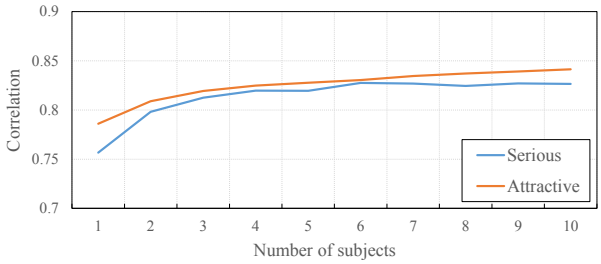


Figure 8: *Correlation between the expression scores computed in Section 5.1 and scores from cross-subject models trained with fewer numbers of subjects. Since there are multiple ways to select x subjects (e.g., for $x = 3$, there are $\binom{11}{3}$ combinations), we randomly select at most 50 combinations and average them to produce plot values.*

smile detectors can filter out the worst images, they miss many of the more attractive expressions.

5.3 Cross-subject predictive model

Our single-subject model can predict attractiveness and seriousness scores for one subject given 10 pairs per expression for both attractive and serious attributes. This crowdsourcing costs on average $21.6 for a single subject to achieve a good scoring function, and $87.8 to accurately rank the top expressions, which is too expensive for real-world applications. Given differences between humans and their facial expressions, it is challenging to build a sufficiently accurate completely automatic model for new subjects without any crowdsourcing. However, we should be able to share information between the single-subject models to build a reasonably effective cross-subject model that can at least serve as an initial condition. We therefore combine features and labels from different subjects, and train a cross-subject SVR model to predict attractive and seriousness scores using the same method as in Section 5.2.

To evaluate the model we hold-out one subject and train on the others, and then average the results of all 11 subjects. The correlation score between the single-subject scores and cross-subject prediction is 0.84 for "attractive", and 0.83 for "serious". The cross-subject model can also be evaluated by its rank error of 1.00; this is significantly higher than the rank errors in Figure 6, and suggests that this model alone is not sufficient to accurately select the most attractive expressions.

Adding data for more subjects may improve the cross-subject model. We plot the correlation between the scores computed in Section 5.1 and versions of the cross-subject model trained with fewer subjects (from 1 to 10) in Figure 8. We can see that seriousness has converged. Attractiveness has mostly converged, but adding a few more subjects will probably slightly improve the model. Also, while our subjects do include a variety of races, genders, and ages, it is likely that there are people for whom our current model will not perform well.

In the end, we use the cross-subject serious model to predict subject-specific seriousness scores since high accuracy is usually not required for this attribute (in our main visualization seriousness scores are only used to assign portraits to serious levels). In the next section we improve the attractiveness score with a small amount of crowdsourcing guided by active learning.

5.4 Active Learning

We wish to collect a small amount of crowdsourced data to improve the ranks and scores for photos of new subjects that are first computed with the cross-subject model. The problem of selecting the optimal data to collect during a learning procedure is called *active learning*, and is well-studied. Though most of the literature addresses collecting class labels for objects, several papers address pairwise comparisons while learning to rank data [Ailon 2012; Jamieson and Nowak 2011; Liang and Grauman 2014]. Most of these techniques address learning a ranking function that operates on data features, and thus can generalize to new data. In our case, we only wish to rank existing representative expressions. Chen et al. [2013] update the Bradley-Terry model we use in Section 5.1 to better handle the crowdsourced setting by taking worker quality into account. We could use their method to produce rankings, but our situation is still unique for several reasons. For one, we are most interested in accurate ranking of the most attractive expressions. Two, our expressions are organized into serious levels, and relative ranking within a serious level is most important; on the other hand, the scores of expressions in different serious levels should still be comparable. Three, while there are subtle differences in the attractiveness of expressions across different subjects, there are also significant commonalities (e.g., open eyes and smiles are usually more attractive). We can therefore use scores from the cross-subject model to predict scores that can serve as a prior.

Nonetheless, our active learning scheme follows the same principles of most previous work. We more frequently sample pairs with high uncertainty [Ailon 2012], which corresponds to pairs with similar attractiveness scores. We add to this scheme a preference for sampling more attractive expressions, and a preference for sampling images of similar seriousness scores. (While only sampling pairs within the same serious level would quickly optimizing ranking error, the scores of different levels would drift from each other; we therefore use a soft preference.) Finally, we use scores from the cross-subject model as a prior.

Our method is initialized by computing baseline seriousness and attractiveness scores $S^0 = \{s_1^0, ..., s_n^0\}$ and $A^0 = \{a_1^0, ..., a_n^0\}$ from the cross-subject model. We fix the seriousness scores and do not attempt to improve them, since they are already reasonably accurate and only used to assign expressions to serious levels. We then iterate through active learning rounds $t = 1, ..., T$. In each round we first select n pairs to sample via crowdsourcing. These

samples are selected by sampling a probability distribution

$$\Pr(I_i, I_j) \sim e^{-||a_i - a_j||^2/2\sigma_a^2} \cdot e^{-||s_i - s_j||^2/2\sigma_s^2} \cdot e^{-[(1-\tilde{a}_i)^2 + (1-\tilde{a}_j)^2]/2\sigma_h^2} \quad (3)$$

where

$$\tilde{a}_i \propto \frac{a_i \sum_j e^{-||s_j - s_i||^2/2\sigma_{\tilde{a}}^2}}{\sum_j a_j e^{-||s_j - s_i||^2/2\sigma_{\tilde{a}}^2}} \quad (4)$$

and $\sigma_{\tilde{a}}$ in Equation 4 is set to the std. deviation of the seriousness scores. The first factor prefers to sample expressions with similar attractiveness scores, i.e., similar ranks. The second factor prefers to sample similar seriousness scores. The third factor prefers to sample more attractive expressions, according to the current estimate of their scores. We use $\tilde{a}_i$ because directly using a_i leads to under-sampling the more serious levels, since serious and attractiveness scores are negatively correlated. Equation 4 normalizes each score a_i by a local weighted average of attractiveness scores, where scores with similar seriousness scores (i.e., close to s_i) are weighted higher. As a result, attractiveness scores that are unusually high for the local range of seriousness are more likely to be sampled. Note that we rescale $\tilde{a}_i$ to [0, 1] after we calculate Equation 4. We use σ_a, σ_s and σ_h to weight the relative importance of each factor. (We describe how each parameter is set later.)

Once we have selected samples within a round t, we update the scoring model before iterating. First, new crowdsourced labels are added to the existing crowdsourced annotation data: $c_{i,j} = c_{i,j} + 1$. Next, we minimize Equation 1 to compute scores. However, in this case we can use the cross-subject model as a more suitable prior than a uniform distribution. We assume a Gaussian distribution $\Pr(A) \sim N(A^0, \sigma_c^2 I)$ as the prior model of A, where I is the identity matrix. That is, we encourage each expression's score to be similar to the cross-subject score. We can thus re-write the MAP Equation 1 as

$$\begin{aligned} A^t &= \arg\min_A -\log \Pr(C|A) - \log(\Pr(A)) \\ &= \arg\min_A -\sum_{i,j} c_{i,j} \log(f(a_i - a_j)) + \frac{1}{2\sigma_c^2} \sum_i ||a_i - a_i^0||^2 \end{aligned} \quad (5)$$

where parameter σ_c controls the emphasis of the cross-subject prior relative to the data-fitting term, and σ in the sigmoid function f is set to the std. deviation of the prior scores A^0. We solve Equation 5 using gradient descent. Notice that $-\log \Pr(C|A)$ increases its influence as we sample more pairs; we start from the cross-subject model and increasingly rely on personalized crowdsourced data as it arrives. Many expressions with low attractiveness scores may never be sampled at all, and simply be scored by the cross-subject model. On the other hand, highly attractive pairs of expressions may be sampled multiple times with different workers.

5.4.1 Simulated Pairwise Comparisons

Our method has four parameters; we set these to minimize the ranking error on pairwise comparisons generated with a simulation, since online optimization with crowdsourcing would be prohibitively expensive. We take the scores generated by random-sampling pairs in Section 5.1, and assume they are ground-truth. We then simulate a Mechanical Turk active learning experiment by generating pairwise labels according to these scores, plus some noise. We label $c_{i,j} = 1$ if a Gaussian random number generator (with bias $a_i - a_j$ and variance σ^i_{worker}) produces a positive number, as suggested by Thurstone's Law [Tsukida and Gupta 2011]. We model each worker's labeling noise with a Gaussian kernel σ^i_{worker}, where the noise std. deviation of the i'th worker (σ^i_{worker}) is sampled from another Gaussian distribution $N(\sigma_{worker}, \sigma^2_{worker})$. We fit the overall variation in worker noise (σ_{worker}) to actual data from our random sampling experiments by performing a grid search on σ_{worker} between $[0.0, 0.8]$. We can see in Figure 9 that our simulation is fairly accurate compared to crowdsourced data. We then set the parameters σ_a^2, σ_s^2, σ_h^2 and σ_c^2 to values that minimize the ranking error by the end of round 20. The optimized parameters are $\sigma_a^2 = 0.02, \sigma_s^2 = 0.5, \sigma_h^2 = 0.1$, and $\sigma_c^2 = 0.5$. Note that the simulation is only run once to set these parameters; it does not need to be run again for new subjects.

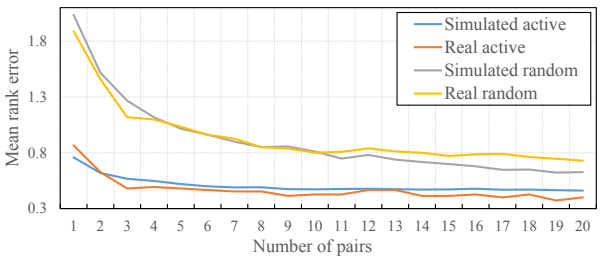


Figure 9: *Mean rank error averaged across three subjects versus the number of pairwise comparisons per expression for four conditions: active learning versus random sampling, across both real (crowdsourced) and simulated data.*

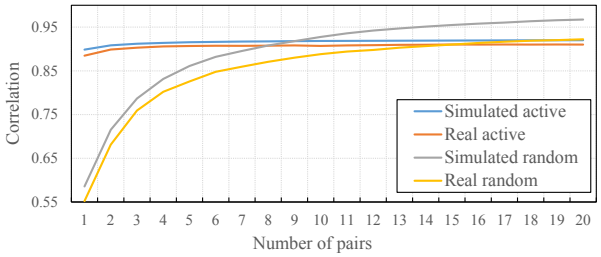


Figure 10: *Correlation between the scores computed in Section 5.1.2 and scores computed using either active learning or random sampling, across both real and simulated data. Correlations are averaged across three subjects.*

5.4.2 Evaluation

We can now evaluate performance over a series of sampling rounds, where each round samples n pairs. We consider four conditions: active learning versus random sampling, across both simulation data and real Mechanical Turk data. Performance can be measured with both mean rank error and the correlation with the attractiveness scores computed in Section 5.1, averaged across three subjects. We show these performance metrics in Figures 9 and 10.

We can see that active learning strongly outperforms random sampling, especially in early rounds, for both simulated and real data. Our active learning scheme can achieve a reasonable accuracy (0.52) with just 5 pairs per expression, while random sampling still has rank error 0.73 after 20 pairs. Using 5 pairs within active learning reduces the crowdsourcing cost to \$5.6, on average, for a subject. Also, after 5 pairs the active learning scheme gives accurate scores, with a correlation over 0.9.

Our method for ranking portraits has a number of components. The active learning probability for selecting pairwise comparisons in Equation 5 has three different factors, and we also use our cross-subject model as a prior. How much do each of these components contribute to the success of our method? We answer this question by turning off individual components and comparing performance

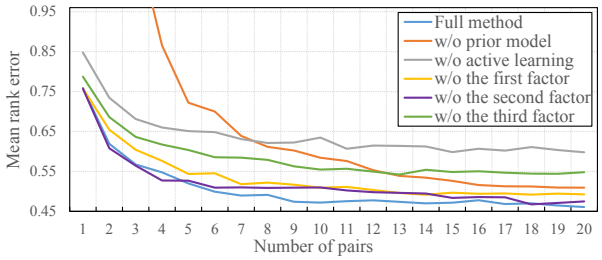


Figure 11: *Performance achieved after removing individual components of our active learning scheme, computed on simulated pairwise comparisons. We compare: (1) the full active learning scheme, (2) active learning without a cross-subject prior, (3) random sampling plus a cross-subject prior, and (4-6) active learning with a cross-subject prior while removing one of the three factors in Equation 3.*

.

using the simulated pairwise comparison data described in Section 5.4.1 (Figure 11). We can see that each part of our method does contribute to reducing mean rank error more quickly. The cross-subject prior has the most significant effect, while comparing expressions with similar seriousness scores has the least significant effect.

5.5 Visualization details

Finally, we give some technical details on how the visualization in Figure 4 is generated.

We first divide seriousness scores into K serious levels, and display the most attractive expression for each serious level. We could simply evenly sample the range of seriousness scores to create serious levels. However, the most attractive expressions tend to be less serious, while there are larger numbers of serious expressions in our input data. The most serious levels may not contain any expressions that are attractive. We therefore divide the seriousness scores into levels based on the idea that the sum of attractiveness scores in each serious level should be about the same.

To compute the number of expressions in each serious level, we first sort attractiveness scores so that their associated seriousness scores are in descending order. We compute the sum of all attractiveness scores, and divide by K to get the target sum of attractiveness for each serious level. Then, we iterate through the sorted attractiveness scores and sum them until we reach expression a_i such that the sum exceeds the target sum for a serious level; the number of expressions in this serious level is set to either i or $i+1$, depending on which minimizes the difference between the current and target sum. The process is repeated until all expressions are assigned to serious levels. We also found it useful to increase the influence of the most attractive expressions during this binning process by first exponentiating each attractiveness score to a power p. We set $k = 25$ and $p = 4$ in all our experiments.

6 Expression Training App

We demonstrate a simple app, called "Mirror Mirror", for training subjects to mimic their best expressions. The app takes input from a webcam and displays the current expression along with its attractiveness and seriousness scores, computed in real-time (about 15fps). Seriousness scores are computed with the cross-subject model after computing features for each input frame; attractiveness is computed from the improved single-subject model computed after active learning. We place a SeeEye2Eye[1] device, which contains a pair of mirrors, on the monitor so that the subject can simultaneously look into the camera and see the camera output.

Figure 12: *Two examples from two subjects of using the cross-fade ability of the expression training app to mimic target expressions; the subjects triggered the capture themselves once they were happy with their expression. We show, from left to right, the target expression aligned to the captured expression, the captured expression, the target expression composited into the current expression, and a* 50% *blend between the previous two images.*

In training mode the app shows the visualization in Figure 4, along with scores of each portrait. The subject can select a target expression to mimic. The app then shows three windows; the current expression, the target expression, and an aligned and a blended cross-fade between the two. The cross-fade oscillates between the target and current expression once per two seconds, so that the subject can examine differences between the two expressions. The target expression is aligned to the current expression and blended to remove visible seams and color differences that might distract from perceiving expression differences. We also show a similarity score between the current and target expression that the user can try to increase. The system automatically saves frames when similarity scores reach new highs; the subject can also pause the system to see fine-grained differences at a frozen moment of time. We show a screen capture of such a session in the supplemental video. We show examples in Figure 12 that demonstrate that subjects can accurately mimic target expressions using our interface.

After alignment we blend the target expression into the current one by performing color histogram transfer between the two images; we then blend with Laplacian pyramids [Burt and Adelson 1983]. We compute the similarity score between the target and current expression with a weighted sum of the difference in attractiveness scores, the difference in seriousness scores, and the projection errors of face alignment landmarks.

[1] `http://www.bodelin.com/se2e`

Figure 13: *We show a comparison of average images of unattractive (left) and attractive (right) portraits organized into 10 bins by eye size (top to bottom, we show 6 of 10 bins). Eyes of equivalent size look different between the two sides.*

Figure 14: *Given a query image (middle) we show expressions that are similar but less or more attractive; expressions are sorted by attractiveness score in increasing order. We show examples for three subjects. (Zoom to see subtle differences; attractiveness scores are shown in red.)*

7 Data Analysis and Visualization

In this section we use our collected and rated portraits to provide users with useful visualizations, glean insights on the properties of attractive portraits, and explore differences between crowd and subject perception of attractiveness.

7.1 Eyes open

In previous work [Albuquerque et al. 2008; Wang and Cohen 2005] it is common to assume that open eyes yield good images, and closed eyes do not. Our analysis shows that the situation is more nuanced. In Section 5.2 we created a simple open-eye detector, and found its correlation with attractiveness scores is only 0.45. It is also useful to visualize the difference between attractive and unattractive photos with the same eye size (Figure 13). We show average images of a single subject grouped into attractive (right) and unattractive (left) clusters by score. The y-axis of the visualization is organized by how open the eyes are; very open eyes are at the top, and closed eyes at the bottom. If we look at the middle bins, we can see a substantial difference in the appearance of good and bad eyes, even though they are open to the same degree. On the left, the eyes appear drugged; the upper eyelid is lowered more substantially than on the right, while the lower eyelid is lower. These bad images usually correspond to expressions in transition (e.g., half-way through a blink). On the right, we can see the same eye size made naturally. Note that smiles often involve narrowing of the eyes.

This observation is consistent with a recent viral video on principles of portrait posing by Peter Hurley[2] that recommends "squinching" (raising the lower rather than the upper eyelid to narrow the eyes). We can see that good eyes of the same size as bad eyes exhibit more squinching.

7.2 Subject Preferences and Poses

When subjects are asked to rank their own best portraits, are their opinions consistent with the crowd? We asked four subjects to rank their top three portraits from the visualization in Figure 4. Their average rank compared to the first, second, and third choices of the crowd are 10, 11, and 10.7. These ranks suggest that subject preferences are not generally consistent with other viewers. An open question is whether friends of the subject, rather than strangers, would also have different opinions.

Second, we examine the success of subjects at posing upon demand. The beginning of our video designed to elicit emotions asks subjects to first pose for three styles of portraits; an open-mouth smile, a closed-mouth smile, and a neutral professional photo. Then, for seven subjects we look at the top ten attractive portraits, and use the video timeline to determine if they came from portrait posing, or from natural responses to videos. We find that, on average, 7.9 of these ten expressions come from natural response, and 2.1 expressions are posed. The mean rank of the single top posed expression in these top ten is 6.6, versus 1.4 for natural expressions. This difference suggests that subjects do not generally show their best expressions when asked to pose. An alternate explanation is that subjects choose to convey something different with their expressions than what the crowd wishes to see.

7.3 Improving Expressions

A subject may like a specific expression, but wish to see if there are similar expressions that the crowd finds more attractive. We therefore generate the visualization shown in Figure 14, where a query expression is shown in the middle, and less and more attractive expressions that are similar to the query are shown on the left and right, respectively. This visualization lets the subject see subtle differences between similar expressions and how they may be improved (or worsened).

To create the visualization from a query expression we retrieve the top two most similar expressions who scores are higher than the query, and two that are lower. Similarity is computed as in Section 4.2.

7.4 Changing One Feature

Another scenario arises when a subject is interested in a specific expression, but wishes to know how changing one feature of the face affects attractiveness. For example, the subject can ask to see different eye or smile sizes, with all other aspects of the face the same.

[2] http://www.youtube.com/watch?v=ff7nltdBCHs

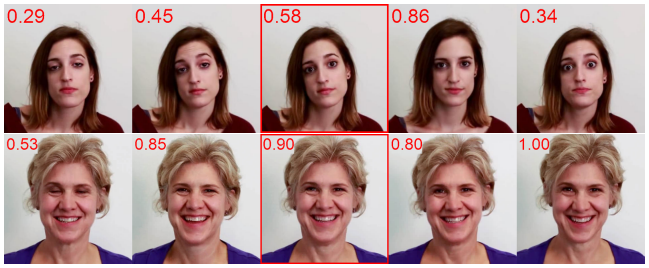


Figure 15: *Given a query image (middle) we show expressions that are similar but with different eye sizes, increasing from left to right, for two subjects. (Zoom to see subtle differences; attractiveness scores are shown in red.)*

In Figure 15 we show examples of different eye sizes, increasing from left to right, for a specific query image (middle). We can see in the first row that increasing the eye size slightly increases attractiveness, but opening the eyes too widely introduces awkwardness.

To create the visualization from a query expression we select the top two most similar expressions whose eye sizes are larger, and two that are smaller. However, in this case we turn off the HOG eye window when computing similarity, since we do not want the eye appearance to be too similar.

8 Results

We tested our method on nine subjects who are not paper authors, and two authors; three of these were also tested using the more expensive, randomly-sampled method in Section 5.1.2. We have already numerically evaluated our active learning scheme, and shown results in Figure 4. Note that all our results shown in Figures are generated by active learning, rather than random-sampling. Visualizations for additional subjects are included as supplemental materials. We tested our training app on four subjects, and show results of mimicking expressions in Figure 12 and the supplemental video.

We also show that our method works on imagery that we did not capture specifically for this paper; in each case, we use only the cross-subject model without any additional crowdsourcing. First, we downloaded a YouTube video[3] on portrait posing; in this video the photographer freezes the frame nine times to indicate good portraits. We select the ten most attractive frames after running peak detection on the attractiveness score signal (to avoid repeating multiple frames of the most attractive expression). Remarkably, nine out of ten selected expressions are the same as those selected by the photographer (Figure 16). We show a plot of the attractiveness scores rated by our cross-subject model over time in Figure 16.

Next, we try two personal photo collections (Figure 17). The first comes from a public person photo dataset [Gallagher and Chen 2008], which already has faces labeled. The second comes from a personal photo collection; we use Picasa to isolate and identify the subjects, and then automatically remove non-frontal faces (angles larger than 15°) using the pose estimates from the face tracker. We compute the attractiveness score on all faces of specific subjects, and show the ten most and least attractive photos. Note that these photo collections are already partially filtered, so there are fewer very bad photos.

Finally, we add an experiment combining our method with the Photobios feature in Picasa [Kemelmacher-Shlizerman et al. 2011] (see supplemental video). We filter the representative expressions to images with attractiveness scores greater than 0.6, and set their dates in order of decreasing seriousness. The resulting Photobio shows a smooth animation of attractive expressions from the most serious to the least.

[3] `https://www.youtube.com/watch?v=yrC9eUwPIoo`

9 Limitations and Future Work

Our method has a number of limitations. While our videos were selected to elicit a wide range of expressions, there is no guarantee that our input video is not missing good expressions of subjects, or that all good expressions can be triggered by watching videos. Also, we only investigate the influence of expression on attractiveness; there are many other factors, such as lighting, camera viewpoint and angle, makeup, and hair. These other factors may not be independent of expression. Though we demonstrate some results on faces captured from an angle, our current methods are not trained on profile or near-profile views.

The most fundamental question about our expression training app is whether it actually helps people pose better for portraits. Conducting this user study accurately would require evaluating the attractiveness of photos from portrait photography sessions before and after using the app; the second session should not be immediately after the training session, to avoid improvements that are only short-term. We leave this more ambitious user study to future work. Our expression training app is only a proof-of-concept for now; it remains an open question whether people can be trained to make certain expressions, or how training compares to other alternatives (such as remembering certain happy or funny moments).

Finally, while we describe methods to select the best expressions, a subject may wish to slightly modify an expression to increase its attractiveness. Using our scoring model to optimize image edits or warps is a promising avenue for future work.

10 Conclusion

We describe a method that uses a combination of crowdsourcing and machine learning to provide users feedback on their best portrait expressions, and to select their most flattering ones from photo collections and videos. While the graphics and vision communities have focused extensively on improving photos through post-processing, we believe there are numerous opportunities to improve photos *before* they are taken. For example, we could identify which photos or very short videos are most effective at eliciting attractive expressions, and play them before snapping a picture. Our large, and often unexplored, collections of photos and videos also offer a large opportunity for identifying flattering content.

Acknowledgements

This work was supported in part by an Adobe Research Grant and ONR MURI N000141010934. We thank Peter O'Donovan for code, Andrew Gallagher for public data, and our subjects for volunteering to be recorded. Figure 1 uses icons by Parmelyn, Dan Hetteix, and Murali Krishna from The Noun Project. The YouTube frames (Figure 16) are courtesy Joshua Michael Shelton.

References

AILON, N. 2012. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research 13*, 1, 137–164.

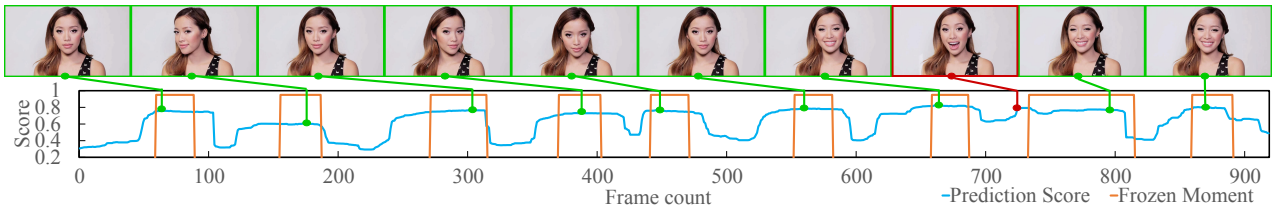


Figure 16: *The ten most attractive expressions selected by our algorithm run on an Internet video about portrait posing (top). We also show a plot of attractiveness over the frames of the video (bottom); the orange rectangles indicate freeze frames used by the photographer to indicate expressions they select. Remarkably, nine out of ten of our selects come from these freeze-frame regions of the video.*

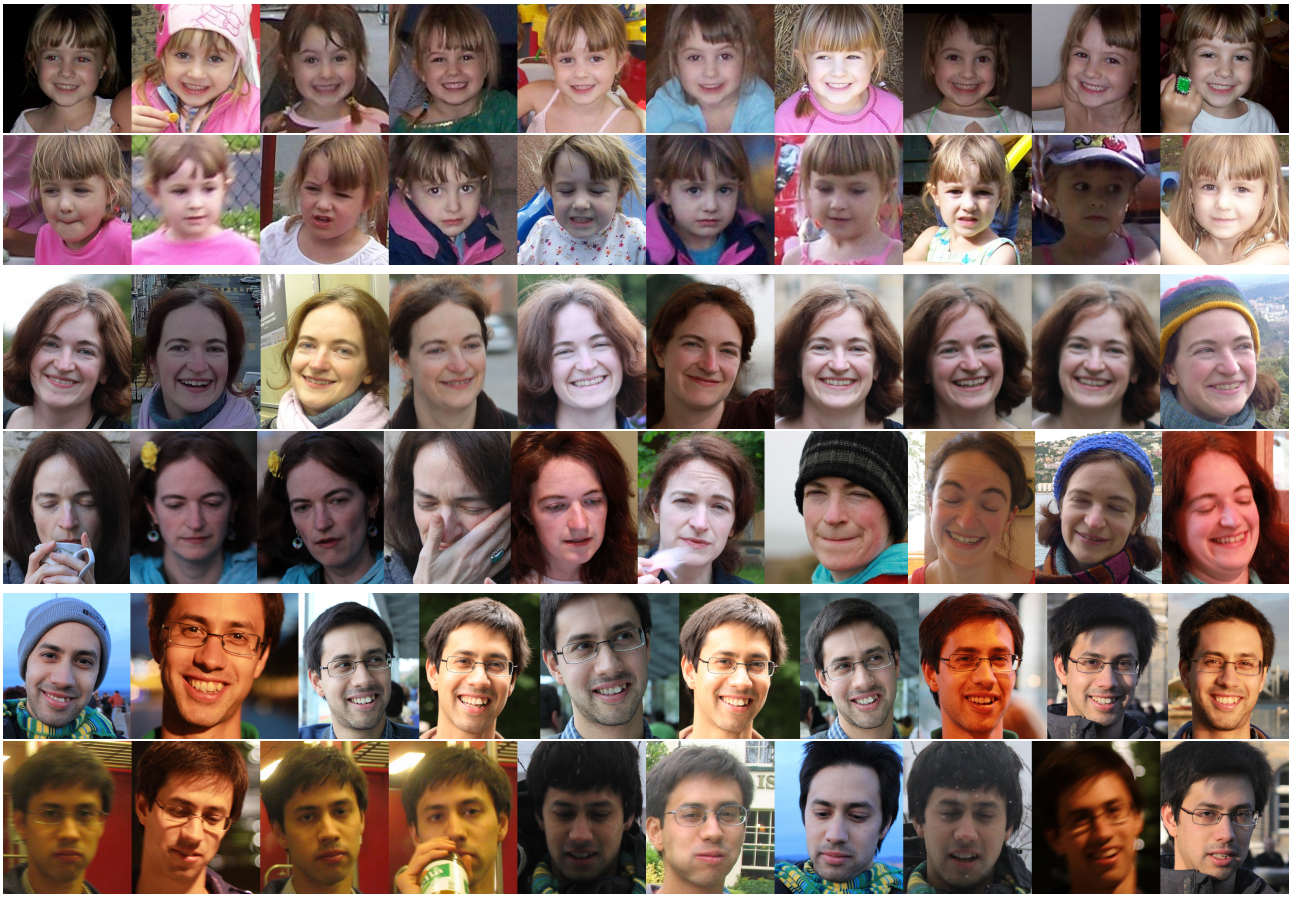


Figure 17: *We show two rows for each of three subjects from personal photo collections: the ten most attractive, and the ten least. We select these expressions from 111, 101, and 85 images of each subject, respectively.*

ALBUQUERQUE, G., STICH, T., SELLENT, A., AND MAGNOR, M. 2008. The good, the bad and the ugly: Attractive portraits from video sequences. In *European Conference on Visual Media Production*.

ALTWAIJRY, H., AND BELONGIE, S. 2013. Relative ranking of facial attractiveness. In *IEEE Winter Conference on Applications of Computer Vision*, 117–124.

AMBADAR, Z., COHN, J. F., AND REED, L. I. 2009. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior 33*, 1, 17–34.

BRADLEY, R. A., AND TERRY, M. E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*, 3/4, 324–345.

BURT, P. J., AND ADELSON, E. H. 1983. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics 2*, 4, 217–236.

CALDER, A., RHODES, G., JOHNSON, M., AND HAXBY, J. 2012. *Oxford Handbook of Face Perception*. Oxford University Press.

CHEN, X., BENNETT, P. N., COLLINS-THOMPSON, K., AND HORVITZ, E. 2013. Pairwise ranking aggregation in a crowd-sourced setting. In *ACM International Conference on Web Search and Data Mining*, 193–202.

DALAL, N., AND TRIGGS, B. 2005. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*.

DIBEKLIOGLU, H., GEVERS, T., AND SALAH, A. A. 2012. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, no. 3, 525–538.

DU, S., TAO, Y., AND MARTINEZ, A. M. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Science*.

EKMAN, P., AND FRIESEN, W. V. 1978. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

FISS, J., AGARWALA, A., AND CURLESS, B. 2011. Candid portrait selection from video. *ACM Transactions on Graphics 30*, 6, 128:1–128:8.

FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

GALLAGHER, A., AND CHEN, T. 2008. Clothing cosegmentation for recognizing people. In *IEEE Conference on Computer Vision and Pattern Recognition*.

GRAY, D., YU, K., XU, W., AND GONG, Y. 2010. Predicting facial beauty without landmarks. In *European Conference on Computer Vision*. 434–447.

GROSS, J., AND LEVENSON, R. 1995. Emotion elicitation using films. *Cognition & Emotion*.

GUNNERY, S. D., HALL, J. A., AND RUBEN, M. A. 2012. The deliberate duchenne smile: Individual differences in expressive control. *Journal of Nonverbal Behavior 37*, 1, 1–13.

HARIHARAN, B., MALIK, J., AND RAMANAN, D. 2012. Discriminative decorrelation for clustering and classification. In *European Conference on Computer Vision*. 459–472.

JAMIESON, K. G., AND NOWAK, R. D. 2011. Active ranking using pairwise comparisons. In *Neural Information Processing Systems*, 2240–2248.

JIANG, B., VALSTAR, M. F., AND PANTIC, M. 2011. Action unit detection using sparse appearance descriptors in space-time video volumes. In *International Conference on Automatic Face & Gesture Recognition*, 314–321.

JOSHI, N., MATUSIK, W., ADELSON, E. H., AND KRIEGMAN, D. J. 2010. Personal photo enhancement using example images. *ACM Transactions on Graphics 29*, 2, 1–15.

KAGIAN, A., DROR, G., LEYVAND, T., MEILIJSON, I., COHEN-OR, D., AND RUPPIN, E. 2008. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision research 48*, 2, 235–43.

KEMELMACHER-SHLIZERMAN, I., SHECHTMAN, E., GARG, R., AND SEITZ, S. M. 2011. Exploring photobios. *ACM Transactions on Graphics 30*, 4, 61.

KHOSLA, A., BAINBRIDGE, W. A., TORRALBA, A., AND OLIVA, A. 2013. Modifying the memorability of face photographs. In *International Conference on Computer Vision*.

KIAPOUR, M. H., YAMAGUCHI, K., BERG, A. C., AND BERG, T. L. 2014. Hipster wars: Discovering elements of fashion styles. In *European Conference on Computer Vision*. 472–488.

KRUMHUBER, E. G., AND MANSTEAD, A. S. R. 2009. Can duchenne smiles be feigned? new evidence on felt and false smiles. *Emotion 9*, 6, 807–820.

LEYVAND, T., COHEN-OR, D., DROR, G., AND LISCHINSKI, D. 2008. Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics 27*, 3, 38:1–38:9.

LIANG, L., AND GRAUMAN, K. 2014. Beyond comparing image pairs: Setwise active learning for relative attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*.

MCDUFF, D., KALIOUBY, R. E., AND PICARD, R. W. 2012. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing 3*, 4, 456–468.

O'DONOVAN, P., LIBEKS, J., AGARWALA, A., AND HERTZMANN, A. 2014. Exploratory Font Selection Using Crowdsourced Attributes. *ACM Transactions on Graphics 33*, 4.

OOSTERHOF, N. N., AND TODOROV, A. 2008. The functional basis of face evaluation. *Proceedings of the National Academy of Science 105*, 32, 11087–11092.

PANTIC, M., AND ROTHKRANTZ, L. J. M. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 1424–1445.

PARIKH, D., AND GRAUMAN, K. 2011. Relative Attributes. In *International Conference on Computer Vision*.

SHAH, R., AND KWATRA, V. 2012. All smiles : Automatic photo enhancement by facial expression analysis. In *European Conference on Visual Media Production*.

SMOLA, A. J., AND SCHÖLKOPF, B. 2004. A tutorial on support vector regression. *Statistics and computing 14*, 3, 199–222.

SPRINGER, I. N., WILTFANG, J., KOWALSKI, J. T., RUSSO, P. A. J., SCHULZE, M., BECKER, S., AND WOLFART, S. 2012. Mirror, mirror on the wall: self-perception of facial beauty versus judgement by others. *Journal of cranio-maxillo-facial surgery 40*, 8, 773–6.

TSUKIDA, K., AND GUPTA, M. R. 2011. How to analyze paired comparison data. Tech. Rep. UWEETR-2011-0004, Dept. of Electrical Engineering, University of Washington.

WANG, J., AND COHEN, M. F. 2005. Very low frame-rate video streaming for face-to-face teleconference. In *Proceedings of the Data Compression Conference*, 309–318.

XIONG, X., AND DE LA TORRE, F. 2013. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 532–539.

YANG, F., WANG, J., SHECHTMAN, E., BOURDEV, L., AND METAXAS, D. 2011. Expression flow for 3d-aware face component transfer. *ACM Transactions on Graphics 30*, 4, 60.

ZHANG, L., SNAVELY, N., CURLESS, B., AND SEITZ, S. M. 2004. Spacetime faces: High-resolution capture for modeling and animation.