Candid Portrait Selection From Video

Juliet Fiss
University of Washington

Aseem Agarwala
Adobe Systems, Inc.

Brian Curless
University of Washington

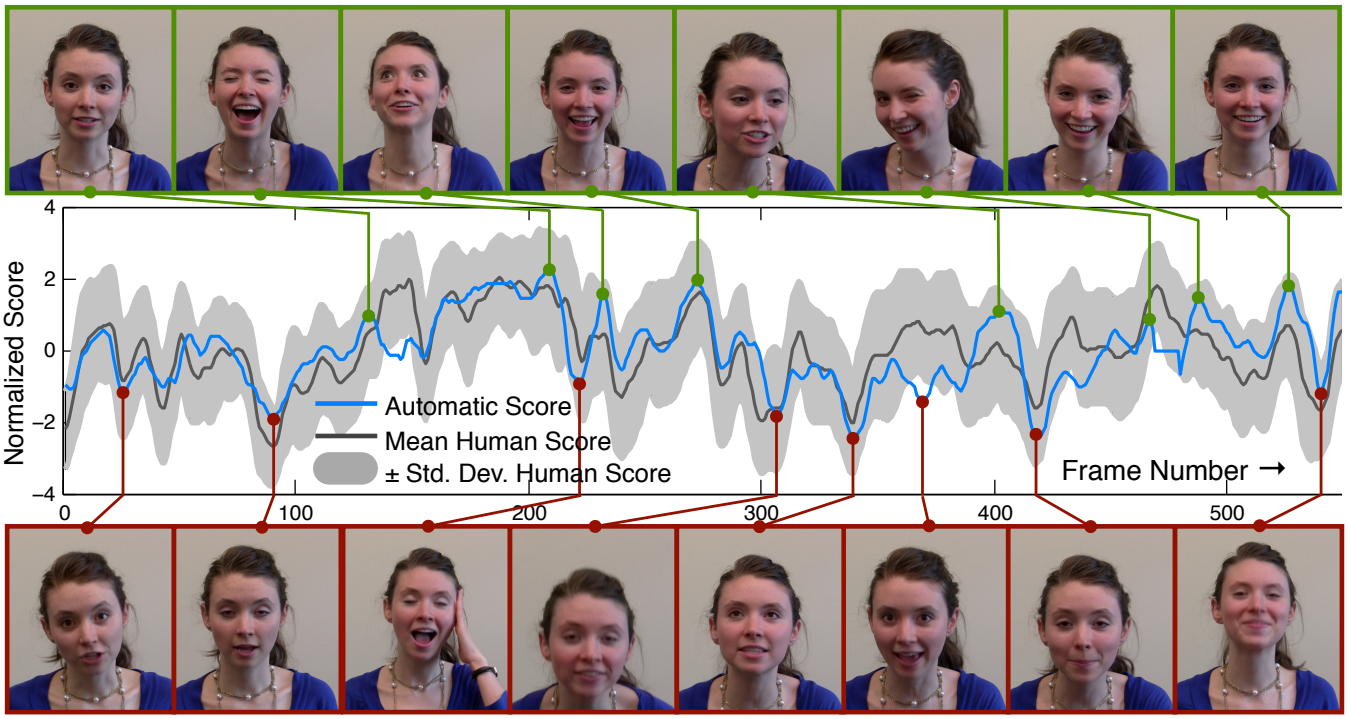

Figure 1: *A plot of the ratings assigned by humans in our psychology study (mean shown in dark gray, per-frame standard deviation shown in light gray), and the ratings assigned by our predictive model (cyan) across the frames of a short video sequence. Both series of ratings have been normalized by their mean and standard deviation. We also show several automatically-selected video frames at peaks (green, top) and valleys (red, bottom) of our predicted rating.*

Abstract

In this paper, we train a computer to select still frames from video that work well as candid portraits. Because of the subjective nature of this task, we conduct a human subjects study to collect ratings of video frames across multiple videos. Then, we compute a number of features and train a model to predict the average rating of a video frame. We evaluate our model with cross-validation, and show that it is better able to select quality still frames than previous techniques, such as simply omitting frames that contain blinking or motion blur, or selecting only smiles. We also evaluate our technique qualitatively on videos that were not part of our validation set, and were taken outdoors and under different lighting conditions.

Links: DL PDF WEB VIDEO DATA

1 Introduction

Cameras sample time very differently than humans do. Still cameras record a frozen moment in time selected by the photographer, and video cameras record a series of moments captured at short, regular intervals. Human vision, on the other hand, is constantly mediated by higher cognitive functions, so that our visual perception of a moment is decidedly different than what a camera captures. This difference is perhaps most glaring when the camera's subject is the human face. It is remarkable how often a photograph or paused video frame of a friend depicts an awkward facial expression that would not be perceived by the human visual system.

While trained photographers are often able to capture the "decisive moment" [Cartier-Bresson 1952], it would be nice if we could simply point our computational cameras at people and automatically acquire only the most desirable photographs. Furthermore, since modern digital cameras can capture very high-resolution video, it would also be nice if we could automatically extract the best moments of people from a captured video. In fact, photographers are starting to record portrait sessions as high-resolution videos; they then select the best frames as a post-process (the cover of the June 2009 issue of Esquire magazine depicting Megan Fox was captured this way [Katz 2009]). In this context, effectively capturing

the moment becomes a filtering task, i.e., selecting the most desirable frames from the very large amount of data. This filtering task can easily become orders of magnitude more time consuming than sorting through images taken with a still camera. Simply viewing 10 minutes of video frames captured at 30 frames per second at a continuous rate of one frame per second would take five hours.

It is especially important in a mobile photography context to be able to automatically extract and upload only the desired moments without time spent on user review. The ideal system would adapt to the goals of the user; some users might want flattering photos for a social networking profile, while others might want more photojournalistic, candid photos that effectively communicate the ongoing narrative. The latter problem is especially challenging, since it is less correlated with the type of expression; angry expressions effectively communicate angry moments, but they are rarely flattering. Since the scope of this problem is large, in this paper we focus on the following subproblem: can a computer be trained to identify digital images of human faces from a video stream that effectively communicate the moment? To solve this problem, we collect data by performing a large-scale psychology study, use machine learning techniques to train a predictive model from this data, and perform cross-validation to evaluate our results. We focus on high-quality video, which is appropriate for the target application of portrait selection. A current limitation of our implementation is its slow speed; we leave a real-time, on-camera implementation as future work.

A major contribution of our research is the design and execution of a large-scale psychology study; in total, we collected 318,240 individual ratings of 12,240 video frames depicting 10 actors, from 103 individual human subjects. These videos and ratings are publicly available on the publication website. Simply asking human subjects to select desirable photos is highly subjective and would lead to large variance in the data. Asking subjects to identify good photojournalistic, candid images is also problematic, since the standards of photojournalism are not universally known, and running large-scale studies with photographers as subjects is cost-prohibitive. So, we devised a simple scenario that leads subjects to evaluate the information content of individual expressions. As further described in Section 3.2, we asked subjects to rate video frames as to how important they are for helping two people communicate over a very low-frame-rate videoconferencing connection. Though there is still significant variance in this data, subjects generally agreed on those frames that are the most or least important in this scenario (see Figure 2 for examples). In our opinion, the highly-rated frames correspond to good candid photos that communicate the moment; lending credence to this notion, a small-scale study with photographers shows significant correlation with our large-scale study (Section 3.3).

2 Related work

Prior work on selecting quality portraits from video is limited to identifying posed portraits with predefined desirable and undesirable attributes. For example, many commercially available point-and-shoot cameras use smile detection to trigger the shutter, or blink detection to alert the photographer to a poor photo. Blink or closed-eye detection has also been used for filtering quality video frames in a video chat scenario [Wang and Cohen 2005]. Albuquerque et al. [2007] used supervised learning to classify images from a posed portrait session as either good or bad, then later extended their method to video [Albuquerque et al. 2008] by running it on each frame independently. Their training data was labeled by the authors with specific rules: bad portraits have eyes closed or looking left or right, or have a mouth that is open, speaking, or showing a smile that is too large. All other frames are labeled good. In contrast, our training data are collected from participants in a large scale psychology study, and do not measure adherence to a predefined set of expressions (e.g., many peaks in Figure 1 do not follow these rules). We also incorporate temporal context into our features; as we show in Section 4, features computed over multiple frames are more discriminative for our learning technique than features calculated from individual frames.

Facial expression analysis is a well-studied topic [Fasel and Luettin 1999; Pantic et al. 2000]. Most of this work is focused on either recognition of specific expressions, automatic lipreading, or segmentation of facial behavior in video. Most facial recognition systems use the Facial Action Coding System (FACS) proposed by Ekman and Friesen [1978]. Our problem does not call for a FACS-like system; in contrast to recognizing specific expressions from their building blocks, we wish to obtain a measure of quality of candid photos that generalizes across different facial expressions. The work most relevant to ours attempts to understand facial dynamics from features such as optical flow [Essa and Pentland 1995; Mase and Pentland 1991; Xu and Mordohai 2010], texture, and geometry [De la Torre Frade et al. 2007; Zhou et al. 2010].

Finally, we are not the first to conduct human subject studies and train predictive models for human preferences of visual data. The most related examples to our work are a predictor of aesthetic ratings of short videos [Moorthy et al. 2010], and visual attractiveness of faces [Leyvand et al. 2008; Gray et al. 2010; Whitehill and Movellan 2008].

3 Data Collection

Our data is composed of (1) a series of videos and (2) human ratings of the frames in the videos, obtained through a psychology study.

3.1 Video Collection

We recorded ten actors, six female and four male, performing a wide variety of facial behaviors. Videos include speech, spontaneous laughter, the six basic emotions associated with canonical facial expressions (happiness, sadness, anger, surprise, fear, and disgust), transitions between expressions, and funny faces. Actors were recorded against a uniform background and in semi-uniform lighting conditions. The video was recorded digitally at 30 fps, 1080p. We selected 17 short clips, each lasting 10-40 seconds, for use in our psychology study and further analysis. These selected clips include a variety of both dramatic and non-dramatic behavior. All actors participating in this initial data collection phase were University of Washington students affiliated with the drama department, and ranged in age between early 20s to early 30s. Later, a more diverse set of actors (including older and younger actors, as well as Computer Science students who had no acting training) participated in a second data collection phase that included both indoor and outdoor environments. We used the videos from this second data collection phase to qualitatively evaluate our algorithm.

3.2 Psychology Study

We performed a psychology study to measure the perceived effectiveness of each frame in our video dataset at communicating the moment. The participants consisted of 103 undergraduate students, 53 female and 50 male, enrolled in an introductory psychology course at the University of Washington. They were not paid, but were rewarded with course credit. Because no background in photography was required for participating in the study, we provided participants with a relatable scenario inspired by Wang and Cohen [2005]. Participants were given the following instructions:

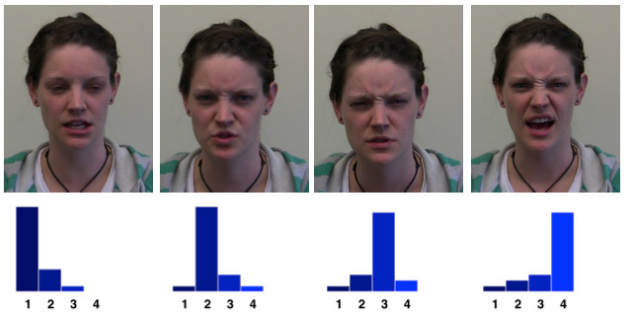

Figure 2: *Four frames of a video, along with a histogram of the ratings assigned by participants in the psychology study.*

> **Background:** Which frames from a video of a person speaking and making facial expressions are the most important for communication? We would like to find out. Your input in this study will help us understand which frames are most important and why.
>
> **Situation:** Imagine that two people are video chatting over a very slow internet connection. The internet connection is so slow that only one frame from the video can be transmitted every few seconds. We would like to transmit only the most important frames from the video to help the two people communicate.
>
> **Instructions:** It is your job to decide which frames from the video are important to transmit to help the two people communicate. You will be shown sequential still frames taken from video where people are communicating in some way. The person in the video may be expressing strong emotion, trying to make someone laugh, or just talking. The video was originally recorded at 30 frames per second.
>
> Rate each frame on a scale from 1 to 4 by pressing the 1, 2, 3, and 4 buttons. Frames rated highly (close to 4) are more likely to be transmitted. Frames rated poorly (close to 1) are less likely to be transmitted. Rate a frame highly (close to 4) if the frame seems important for communicating either: 1. the mental or emotional state of the person in the video –or– 2. a facial expression that the person in the video is clearly intending to make. Sometimes there is very little movement from frame to frame. You may assign the same rating to multiple frames in a row. Try to use all of the possible ratings (1, 2, 3, 4) in each video sequence.

Subjects were also shown example frames for each rating. Frames were shown in temporal order, and subjects were allowed to navigate back and forth between frames. We designed the experiment this way because, in our personal experience, we found it easier to rate frames if we could see their temporal context. The subjects did not hear the audio or see the original video. Each subject rated between four and five video clips, approximately 100 seconds of video per subject. The order in which full videos were presented to subjects was randomized. The mean number of ratings per video frame was 26. However, we manually discarded data from 22 subjects who were clearly not following directions, e.g., repeatedly entering '1, 2, 3, 4, 1, 2, 3, 4, ...' After bad data removal, the mean number of ratings per video frame was 20.8. Figure 2 shows some rating histograms for several frames from one sequence.

Averaged over all rated frames, the mean rating was 2.40 with a standard deviation of 0.93. We found that there was more agreement on frames with very high or very low ratings: frames scoring in the 95th percentile and up had a standard deviation of 0.76; frames scoring in the 5th percentile or below had a standard deviation of 0.71. The middle 5 percent had a standard deviation of 0.99. One of the actors filmed had a full beard, which interfered with the tracking step in our algorithm. Although we collected rated data for one video clip of this actor, we exclude this clip from further analysis. Before further processing, the ratings collected for each video from each participant were individually normalized by subtracting their mean and dividing by their standard deviation. These data were then averaged over all participants. Finally, we reduced noise with a Gaussian lowpass filter of full width half maximum (FWHM) 0.13 seconds. This processed signal served as input to our learning step.

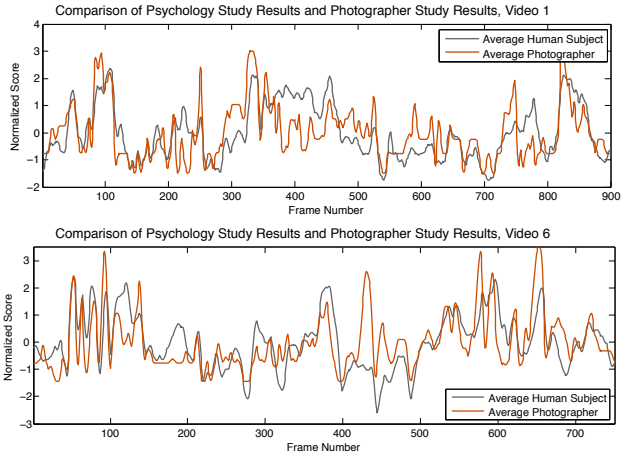

Figure 3: *Plots of the ratings assigned by participants in our psychology study (dark gray), and the ratings assigned by professional photographers (orange) to the frames in Video 1 (top) and Video 6 (bottom). All series of ratings have been smoothed and normalized.*

3.3 Photographer Study

While our overall goal is candid portrait selection, the instructions of our large-scale study propose a videoconferencing scenario, since we were concerned that portrait selection among non-photographers would yield high variance. To what extent does this methodology choice distort our data? To test this concern, we asked three professional photographers, two male and one female, to perform the same task as in our human subjects study for four of our videos, but with different instructions; unlike in the psychology study, photographers did not see any examples of rated frames:

> **Instructions:** You will be shown sequential still frames taken from a video portrait session. In a video portrait session, video, rather than still shots, is captured. Later, the photographer sorts through the still frames and selects those that would work well as still photographs. The goal of the video portrait session you are about to see is to capture candid, photojournalistic style portraits. The person in the video may be expressing strong emotion, trying to make someone laugh, or just talking. The video was originally recorded at 30 frames per second. Your job is to rate each frame based on how well it works as a candid, photojournalistic style portrait.

The scores provided by the photographers were preprocessed using the same techniques as the scores collected in the psychology study. Figure 3 shows visual overlays of the scores for two video clips

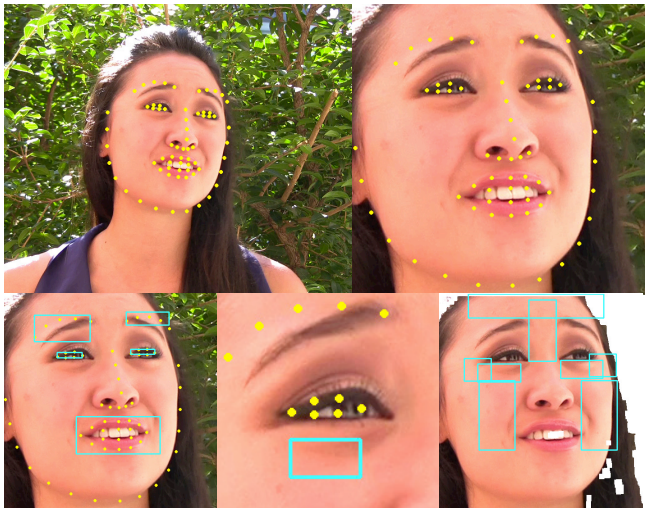

Figure 4: *Top Left: Original video frame with superimposed tracked points. Top Right: Face after pose normalization, i.e., automatically cropped, scaled, and adjusted for in-plane rotation. Bottom Left: Computed bounding boxes for eyes, mouth, and eyebrows. Bottom Center: Computed bounding box for reference patch. Bottom Right: Image and regions used for wrinkle detection. The skin has been segmented by hue. Regions of interest for wrinkle detection include forehead, brow furrow, lower eyelids, crow's feet, and cheeks.*

from these experiments, Video 1 and Video 6. Although there are some discrepancies, in general, the shape of the two curves match well. Additionally, we observed large positive Pearson correlation coefficients of $r_6 = 0.58$, $r_1 = 0.66$, $r_{13} = 0.66$, and $r_3 = 0.55$ for the pairs of ratings supplied by the professional photographers and psychology experiment participants (the subscript of r refers to the video ID). The results of this short experiment suggest that asking the general population to select frames that help people communicate is a reasonable proxy for having professional photographers rate frames based on how well they work as candid, photojournalistic portraits.

4 Predictive Model

The result of our human subjects study is a series of video frames, each with a distribution of ratings from 1 to 4. Our next step is to train a predictive model that can predict the expected rating of a video frame. Our algorithm first tracks the faces in the video, and then normalizes the faces to canonical positions using the tracking data. Then, we compute a series of features that are designed to measure properties that we observed to be correlated with the ratings. Finally, we use supervised learning to train a model to predict a rating given a feature vector.

4.1 Tracking and pose normalization

Face tracking and pose normalization is an essential pre-processing step for feature extraction. We use the face tracker described by Saragih et al. [2009]. This tracker computes six global parameters representing head position and orientation, as well as the locations of 66 points on the face: 12 on the eyes, 10 on the eyebrows, 18 on the mouth, 9 on the nose, and 17 on the jawline (Figure 4, top left). We use the global parameters to normalize and resample each frame for scale (we use 480×480), translation, and in-plane rotation (Figure 4, top right). We then use the tracked points to locate bounding boxes of regions of interest on the face, including the eyes, mouth, eyebrows (Figure 4, bottom left), forehead, cheeks, lower eyelids, crow's feet (wrinkles extending from the outer corners of the eyes), cheeks, and brow furrow (Figure 4, bottom right). Details of how the extents of each bounding box are computed from the locations of tracked points can be found in the pseudocode included with our supplementary material.

In addition, we use the color information in *reference patches* (Figure 4, bottom center) in the upper cheek regions to determine the average hue of the face. The horizontal extents of the patch are determined by the left- and right-most eye points, while the vertical extent is 20 to 50 pixels below the bottom eye point. The hue within the reference patch is used to find a binary mask of areas of the image with matching skin color, i.e., all pixels within an angular distance of 3.6 degrees to the reference patch mean in the hue component of HSV color space. Erosion with a 6×6 structuring element followed by dilation with a 3×3 structuring element is used to clean up the mask. This binary mask is used to remove non-skin pixels in the regions of interest when computing some features, for example, features based on wrinkles (Figure 4, bottom right).

4.2 Features

Through manual inspection of the video data, we observed that the following properties were correlated with higher ratings: (1) intensity of expression, with the natural apex of an expression given the highest relative rating within a sequence of frames; (2) pauses or changes of direction in the motion of the facial muscles, which often corresponds to either the apex of an expression or a pause between expressions; (3) wide smiles and other dramatic expressions that bare the teeth tend to be given high ratings.

We observed that the following properties were correlated with lower ratings: (1) reductions in visual clarity of the expression caused by blinking, motion blur, or the head turned away from the camera; (2) fast head motion causing blur or distortion of the face; (3) motion of the facial muscles or mouth, which often indicates an expression transition or talking.

We therefore designed features to measure these properties, while also being mindful of several other design guidelines: (1) features should generalize well across different people, i.e., be independent of age, skin color, and other human variations; (2) features should be robust to occluders like facial hair and glasses; (3) features should generalize across expressions. For example, a feature that detects peaks in facial dynamics should work equally well for smiles and winces.

With these factors in mind, we designed six basic features that fall into two categories: features based on motion, and features based on texture. Initially, all features are computed based on information from only the current or immediately surrounding frames, using the techniques described in more detail below. After the initial feature computation, we normalize all features per video by subtracting the temporal median and dividing by the temporal mean absolute deviation. Normalizing the features in this way allows us to measure relative changes over time and account for differences among actor appearance and video quality. Finally, each feature is convolved, in the time domain, with a set of averaging and edge-detecting kernels. This filtering step adds temporal context to the features computed at each frame, and the result of each of these filtering operations is used as an additional feature. The specific set of kernels we use includes Gaussian lowpass filters of full width half maximum (FWHM) 0.98 seconds (later referred to as the "long lowpass filter"), 0.38 seconds ("medium lowpass filter"), and 0.16 seconds ("short lowpass filter"), a signum function with support 0.5 seconds, and a ramp function of support 0.5 seconds. In our videos, 1

second = 30 frames. We also use the original, unfiltered features in our training step. The result is a 36-component feature vector that is used to train our predictive model. We now describe the methods used to compute each feature. All features are computed on the output of the pose normalization step described in Section 4.1.

Motion-Based Features. Motion-based features are computed using the optical flow algorithm of Sand and Teller [2006] within patches of interest. Pose normalization removes most of the gross head motion from the optical flow, but this normalization can have errors. Thus, we additionally compensate for head motion by subtracting the average optical flow over the corresponding reference patch (described in Section 4.1) from the average optical flow of the patch of interest. For example, to measure left eye motion, we take the difference of the left eye patch average flow and the left reference patch average flow. The L1-norm of this difference is the optical flow score for that patch; flow scores are then mapped to feature values, as described in the subsections below.

Feature 1: Blink and eye motion detection. We compute the optical flow score for each eye region and average the result to give the feature value; if only one eye is visible, then the score for that eye is used. Minimal optical flow indicates that either the eyes are open and focused on a target, or closed, but not blinking. To account for the fact that the eyes do not appear intentionally focused or fixated in the frames just before and just after a blink, this feature is low-pass filtered over a window of three frames before further processing. An example of this feature is shown in Figure 5.

Feature 2: Mouth motion detection. We compute the average optical flow within regions for the center, left corner, and right corner of the mouth. A large optical flow magnitude indicates that the mouth is moving relative to the rest of the face.

Feature 3: Facial feature motion detection. To measure overall facial feature motion, we sum the optical flow magnitudes for the cheeks, regions surrounding the mouth, and eyebrows. In a facial action such as an eyebrow raise, the optical flow of the facial feature is at a minimum when the expression is held momentarily, the peak activation of the facial action.

Texture-Based Features. Texture-based features are computing using Gabor filter responses and gradient magnitudes within patches of interest that have been converted to grayscale. In addition to detecting important textures on the face, such as wrinkles and teeth, these features serve the additional purpose of producing a very low response when the image of the face is blurry.

Feature 4: Whole face texture feature. For each frame, we compute Gabor filter responses on grayscale images at four scales and eight orientations in regions of the face known to wrinkle: the forehead, brow furrow, crow's feet, lower eyelids, and smile lines. We then average the magnitude of all filter responses in all regions to produce a single scalar feature for each frame that is high when the subject has a lot of facial wrinkles, and low when the subject has fewer facial wrinkles. An example of this feature is shown in Figure 5.

Feature 5: Upper face texture feature. This feature measures specific directional wrinkles that form in the upper face region. Using the mean gradient magnitude, we measure the horizontal wrinkles that form in the forehead, for example when the eyebrows are raised, and the vertical wrinkles that form in the furrow of the brow.

Feature 6: Teeth detection. Since the display of teeth is an important feature of several expressions, most notably smiles, we compute a feature that produces a high response when the teeth are shown. We measure the mean absolute value of the horizontal image gradient within the mouth region. This feature is high when the the edges of the teeth are visible.

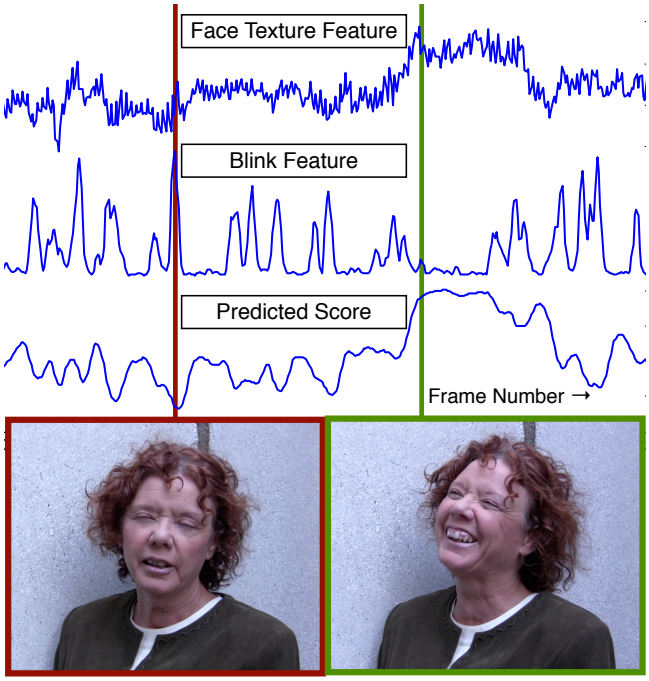

Figure 5: *Top: Whole face texture feature (upper), blink and eye motion feature (middle), and overall predicted score (lower) tracked over time for a short video clip. Bottom Left: Frame with the highest value for the blink feature, which negatively contributes to overall score. Bottom Right: Frame with the highest value for the whole face texture feature, which positively contributes to overall score. Note that the blink feature does not have a high value at the bottom right frame because the eyes are narrow for an extended period of time.*

We also experimented with features based on locations of tracked points, parameters returned by the tracker, as well as facial symmetry. However, we found that these features were not robust enough for our purposes. For example, the geometry of the mouth was not tracked accurately and consistently enough to derive expression information from the tracked points alone. However, the tracked points did provide enough information to localize the mouth so that our features could be computed within the correct region. Finally, we ignore the computed features for frames where the tracker failed. The tracker reports whether it was successful at each frame. In the cases where it failed, we set the feature values to their medians. In our validation set, there are only a few frames where tracker failure was an issue.

4.3 Learning

The features computed in the previous section comprise a 36-component vector describing each video frame and its temporal context; the next step is to perform regression to predict a mean rating from a feature vector. We experimented with probit and logistic regression [McCullagh and Nelder 1989], LASSO [Tibshirani 1996], and Gaussian processes [Williams and Rasmussen 1995]; we found that probit regression achieved the lowest error. Our human-provided mean scores are scaled linearly to $[0, 1]$ on a per-video basis before being passed into the regression algorithm.

According to the feature weights learned by our regression system, the top 10 most influential features were as follows: (1) blinking feature, long lowpass filter. (2) mouth motion feature, long lowpass

Figure 6: *Top: Eight highest-scoring peak frames in the score predicted using our method for Video 6. Middle: Eight highest-scoring peak frames in the mean human score from the psychology study for Video 6. Bottom: Eight highest-scoring peak frames in the mean score from the photographer study for Video 6. All sets of frames have been sorted temporally.*

Figure 7: *Examples of frames with the highest automatic score from very expressive video clips. Top Left: Frame captured while smiling and talking. Top Right: Frame captured while gesturing and telling a funny story. Bottom Left: Frame captured while laughing. Bottom Right: Frame captured while acting out the emotions of surprise and disgust as part of a funny story.*

Figure 8: *Two highest scoring (top) and two lowest scoring (bottom) frames from a video portrait shoot.*

filter. (3) blinking feature, medium lowpass filter. (4) whole face motion feature, medium lowpass filter. (5) mouth motion feature, medium lowpass filter. (6) upper face wrinkle feature, medium lowpass filter. (7) upper face wrinkle feature, short lowpass filter. (8) blinking feature, short lowpass filter. (9) whole face wrinkle feature, short lowpass filter. (10) mouth motion feature, short lowpass filter. In general, the long, medium, and short lowpass filtered versions of the base features were the most influential. The noisy, unfiltered versions of the base features were less influential, and the 1D edge detecting filters were least influential. However, adding all of these features positively impacted the learning results.

4.4 Evaluation

We evaluate our predictive model with cross-validation experiments on 16 videos that were rated by participants in the psychology study. During each fold of validation, we exclude any videos from the training set that feature the same actor as the test video. We compute the mean squared error (MSE) between the predicted and human-provided scores for the test and training sets. As a baseline, we predict the mean of the training set for every frame and compute the MSE. Averaged over all validation folds, we saw a median 31 percent reduction in MSE for the test sets, and a median 34 percent reduction in MSE for the training sets compared to the baseline. Full numerical results for all videos in the validation set are provided with the supplementary material. To select individual frames from a continuous score, we use a peak detection algorithm, `peakdet`[1], then return the N top-scoring peak frames.

Figures 1 and 6 show example results that have been sorted temporally for videos 3 and 6 from our validation set. Figures 5, 7, 8, and 9 demonstrate the results of our algorithm on videos that were not part of the validation set, and were recorded outdoors with a handheld camera. Many of the frames selected as top peaks fall outside of the criteria defined as "good" by Albuquerque et al. [2007] (these

[1] `http://www.billauer.co.il/peakdet.html`

Figure 9: *Three highest scoring (top) and three lowest scoring (bottom) frames from a video of an actor singing a humorous song.*

Figure 10: *Top row: high-scoring peak frames that would not meet the criteria for "Good Frame" defined by Albuquerque et al. [2007]. Bottom row: low-scoring valley frames that do not contain the obvious artifacts of blinking or motion blur.*

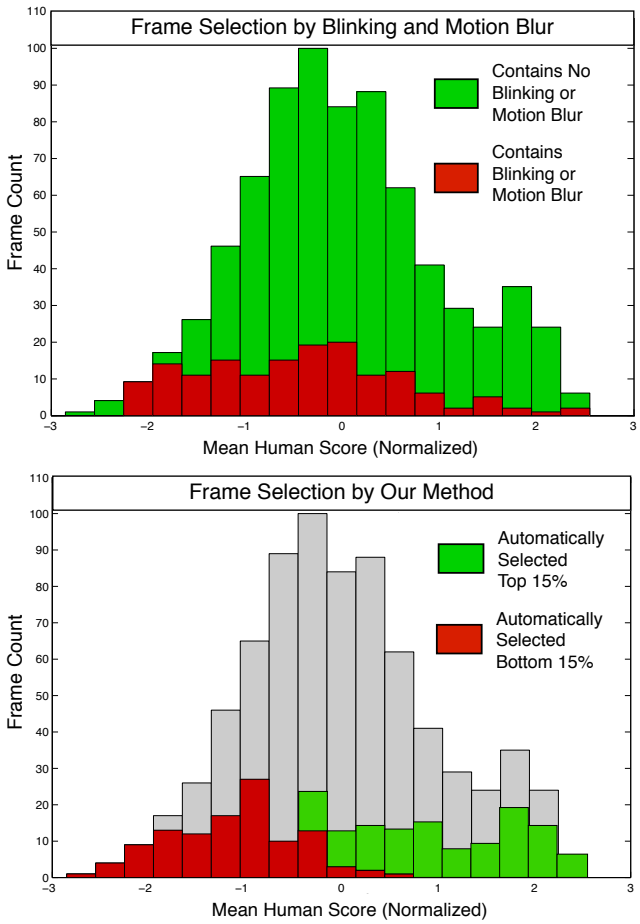

Figure 11: *Comparison of selecting best and worst frames using manual blink and motion blur classification versus our method.*

criteria include open eyes looking at the camera, with a neutral or slightly smiling mouth). The top row of Figure 10 shows frames selected as top peaks by both psychology participants and our algorithm that do not meet these criteria. The bottom row of Figure 10 shows examples of frames rated poorly by both humans and our algorithm that do not contain the obvious detractors of blinking and motion blur. The bottom center frame fails as a portrait because of the slightly defocused eyes, which were moving in the original video. Albuquerque et al. [2007] would likely classify this frame as "good." Figure 7, Bottom Left depicts the subject laughing with closed eyes; the top row of Figure 1 shows similar frames that were rated highly by users. Figure 7, Bottom Right shows a peak frame taken from a video of an actress expressing surprise and disgust while telling a funny story. While this expression is not particularly flattering, it is the apex of an expression that conveys the intent of the actress and the content of the story she is telling. Full sets of results, along with the original video clips, can be found in our supplementary materials.

In Figure 11, we compare the results of our algorithm to simply removing all frames that contain blinking or motion blur for video 6 from our validation set. The histogram of normalized mean human ratings is shown for the video in both the upper and lower parts of the figure. In the top histogram, all frames that contain blinking or motion blur (detected manually) are highlighted in red (for these frames, mean human rating = -0.43, std. dev. = 1.03). All other frames are highlighted in green (mean = 0.09, std. dev. = 0.97). In the bottom histogram, we use our method to predict the bottom 15 percent of frames, highlighted in red (mean = -1.13, std. dev. = 0.65), and top 15 percent of frames, highlighted in green (mean = 1.03, std. dev. = 0.80). For a perfect predictor, the highlighted regions would perfectly correspond to the tails of the distribution. The comparison of the two distributions shows that while blinking and motion blur are correlated with low scores, they are less accurate predictors of human score than our method. Furthermore, frames containing blinking and motion blur comprise only a small fraction of the total frames in the video (about 20 percent in this case), so one is still left with a challenging quality filtering task after removing them.

4.5 Limitations

Our predictor will not perform well if the face tracker fails to accurately track the face. We have found the tracker to be fairly robust to video resolution, lighting and skin tone variation, as well as some occluders including moderate facial hair (as shown in Figure 9), hair falling into the face, and glasses. However, the tracker often fails in the presence of other occluders, including thick beards and hands obscuring more than a third of the face. As face trackers improve, we hope to leverage this technology to expand the input domain of our system. For our system to work on subjects with full beards, we may find that we need to train specifically for this case, or modify our texture features around the mouth. The tracker and other aspects of the algorithm will also fail on heavily compressed

videos and interlaced videos. Since frames taken from these degraded videos would not make good portraits, we opted to evaluate our results on only high quality video.

Because we consider our current implementation a prototype, we made no attempt to optimize for speed. However, we anticipate that it should be possible to create a real-time implementation in the near future. In our implementation, processing takes on the order of minutes per frame, but is performed concurrently, so that processing an entire video clip also takes on the order of minutes or tens of minutes, depending on the length. The bottlenecks in our implementation are the computation of optical flow and the Gabor filter responses; both of these operations have been computed in real-time in other implementations. In our prototype system, we compute these metrics on the entire frame as a pre-processing step; a speed-optimized implementation would limit the regions of computation to specific areas of interest. While we hope that our technique can soon be implemented in real-time on a computational camera, we would also like to emphasize the utility of the offline implementation that selects portraits from pre-recorded video.

5 Conclusions and Future Work

In this paper, we explored whether an algorithm can perform a task that is subjective yet fairly straightforward for a human: select frames from a video of a human face that effectively communicate the moment and work well as candid portraits. To automate this task, we collected a large dataset of human ratings, and trained a predictive model to select those frames that are most or least effective as candid portraits. While our algorithm can currently be used as a post-process filter on video portrait sessions, we also hope that the release of our data and algorithm will lead to a real-time, on-camera implementation.

Furthermore, we believe that our methods are a step in an exciting direction. Models of human perception are becoming increasingly important in automated computer graphics, but human perception is notoriously challenging to model and understand. While our specific features are tuned to faces, we believe that our overall methodology could be applied to many related problems. For example, an extension of our method would be to use full-body motion and texture features to predict perfectly-timed action shots of athletes and dancers. More generally, human visual preferences are important but tricky to model; we think performing large-scale human studies plus machine learning is the right way to create effective perception-aware algorithms.

Acknowledgements: We would like to thank Geoffrey Loftus and Julie Anne Seguin for their guidance on designing and facilitating the psychology studies featured in this paper, and we would like to thank Jason Saragih for the use of his face tracker.

References

ALBUQUERQUE, G., STICH, T., AND MAGNOR, M. 2007. Qualitative portrait classification. *Proc. Vision, Modeling, and Visualization (VMV'07)* (11), 243–252.

ALBUQUERQUE, G., STICH, T., SELLENT, A., AND MAGNOR, M. 2008. The good, the bad and the ugly: Attractive portraits from video sequences. In *Proc. 5th European Conference on Visual Media Production (CVMP 2008)*.

CARTIER-BRESSON, H. 1952. *The Decisive Moment*. Simon & Schuster.

DE LA TORRE FRADE, F., CAMPOY, J., AMBADAR, Z., AND COHN, J. F. 2007. Temporal segmentation of facial behavior. In *International Conference on Computer Vision*.

EKMAN, P., AND FRIESEN, W. V. 1978. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 577 College Avenue, Palo Alto, California 94306, 1978.

ESSA, I., AND PENTLAND, A. 1995. Facial expression recognition using a dynamic model and motion energy. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, IEEE Computer Society, Cambridge, MA, 360–367.

FASEL, B., AND LUETTIN, J. 1999. Automatic facial expression analysis: A survey. *PATTERN RECOGNITION 36*, 1, 259–275.

GRAY, D., YU, K., XU, W., AND GONG, Y. 2010. Predicting facial beauty without landmarks. In *Proceedings of the 11th European conference on Computer vision: Part VI*, Springer-Verlag, Berlin, Heidelberg, ECCV'10, 434–447.

KATZ, D. 2009. Good morning, Megan Fox. *Esquire* (June).

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

MASE, K., AND PENTLAND, A. 1991. Lipreading by optical flow. *Systems and Computers 22*, 6, 67–76.

MCCULLAGH, P., AND NELDER, J. A. 1989. *Generalized linear models*, 2nd ed. London: Chapman & Hall.

MOORTHY, A. K., OBRADOR, P., AND OLIVER, N. 2010. Towards Computational Models of Visual Aesthetic Appeal of Consumer Videos. In *Proc. ECCV*.

PANTIC, M., MEMBER, S., 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.

SAND, P., AND TELLER, S. 2006. Particle video: Long-range motion estimation using point trajectories. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on 2*, 2195–2202.

SARAGIH, J. M., LUCEY, S., AND COHN, J. 2009. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*.

TIBSHIRANI, R. 1996. Regression Shrinkage and Selection Via the Lasso. *Royal. Statist. Soc B 58*, 1, 267–288.

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*, DCC '05, 309–318.

WHITEHILL, J., AND MOVELLAN, J. R. 2008. Personalized facial attractiveness prediction. In *FG*, 1–7.

WILLIAMS, C. K. I., AND RASMUSSEN, C. E. 1995. Gaussian processes for regression. In *Neural Information Processing Systems*, 514–520.

XU, L., AND MORDOHAI, P. 2010. Automatic facial expression recognition using bags of motion words. In *Proceedings of the British Machine Vision Conference*, BMVA Press, 13.1–13.13. doi:10.5244/C.24.13.

ZHOU, F., DE LA TORRE, F., AND COHN, J. F. 2010. Unsupervised discovery of facial events. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.