

Automated Prediction of Preferences Using Facial Expressions

David Masip^{1,2*}, Michael S. North³, Alexander Todorov⁴, Daniel N. Osherson⁴

1 Estudis d'Informàtica Multimedia i Telecomunicacions, Universitat Oberta de Catalunya, Barcelona, Spain, **2** Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain, **3** Department of Psychology, Columbia University, New York, New York, United States of America, **4** Department of Psychology, Princeton University, Princeton, New Jersey, United States of America

Abstract

We introduce a computer vision problem from social cognition, namely, the automated detection of attitudes from a person's spontaneous facial expressions. To illustrate the challenges, we introduce two simple algorithms designed to predict observers' preferences between images (e.g., of celebrities) based on covert videos of the observers' faces. The two algorithms are almost as accurate as human judges performing the same task but nonetheless far from perfect. Our approach is to locate facial landmarks, then predict preference on the basis of their temporal dynamics. The database contains 768 videos involving four different kinds of preferences. We make it publically available.

Citation: Masip D, North MS, Todorov A, Osherson DN (2014) Automated Prediction of Preferences Using Facial Expressions. *PLoS ONE* 9(2): e87434. doi:10.1371/journal.pone.0087434

Editor: Oscar Deniz Suarez, Universidad de Castilla-La Mancha, Spain

Received: August 30, 2013; **Accepted:** December 23, 2013; **Published:** February 4, 2014

Copyright: © 2014 Masip et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Research supported by TIN2012-38187-C03-02 and Jose Castillejo grant from the Ministerio de Educación Cultura y Deporte (Spain) <http://www.mecd.gob.es/portada-mecd/>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dmasipr@uoc.edu

Introduction

Recently, social psychologists have shown that people can infer which of two stimuli are preferred by human observers just by viewing covertly recorded videos of the observers' faces [1,2]. Automating these inferences might be useful to the development of electronic devices that respond in human-like ways to their users. Previous research related to this goal has involved face recognition [3], social trait inference [4–7], and the analysis of expression [8,9], but not the prediction of preference from spontaneous videos. Previous work on the automated analysis of facial expressions, moreover, tends to focus on the six basic emotions defined by [10], and the Facial Action Coding System [11]. These studies are mainly limited to exaggerated expressions with posed dynamics. Likewise, publically available face data typically involve exaggerated facial expressions. We propose here to study more mundane stimuli, using low resolution videos acquired in a spontaneous and non-controlled setting. The resulting facial expressions are briefer and vastly more challenging to interpret. Specifically, the present paper makes three contributions. (i) We introduce the problem of automated inference of preferences from videos, (ii) we make available an annotated data set (with frame-by-frame landmark locations) for experimental purposes, and (iii) we propose two simple algorithms (as a baseline) for predicting preferences. Our goal is merely to articulate and illustrate the problem of interpreting spontaneous faces rather than to explore the space of possible algorithms.

Methods

Database Creation

[1] created a video database divided into four categories: *people*, *cartoons*, *animals*, and *paintings*. Eight subjects examined twelve pairs

of images from each category. The two images in a pair were examined serially. When viewing people, they judged which of the two was more attractive. When viewing cartoons, they judged which was funnier. When viewing animals, they judged which was cuter, and when viewing paintings they judged which was aesthetically superior. For details about counterbalancing and experimental design, see [1]. Unknown to the subjects, their faces were covertly recorded while they examined a given pair of images. Only after both images in a given pair were shown and withdrawn did the subject indicate his/her preference; hence, recording occurred while the face was involved in nothing more than examining an image. The recording of the videos was approved by the Institutional Review Board (IRB) of Princeton University, and participants signed a film release authorizing the use of the data for future studies.

In a second phase, 56 new participants tried to guess the original subjects' preferences about the pairs of images just by observing their faces. The second set of subjects did not have access to the pairs of images shown earlier; they made their guesses about preference based only on videos of faces. Henceforth, following the terminology of [1], we call the first set of subjects "targets" and the second second set "perceivers." Each target was viewed by 14 perceivers, drawn from the set of 56.

The total number of videos in the experiment is 768 (4 categories × 8 targets × 12 pairs of videos × 2). In this paper we consider video pairs as the basic processing unit, yielding 96 pairs for each category. Individual videos lasted three seconds for the people, paintings and animal stimuli, and seven seconds for the cartoons. All videos were recorded at a rate of 24 frames per second; they were acquired via WebCam with 640×480 RGB resolution. The entire data base is available at <http://tlab>.

princeton.edu/databases/ (Princeton Preferences from Facial Expressions Data Set).

Facial Landmark Detection

Our algorithm relies on the dynamics of salient points that reveal the structure of faces. These points are called “landmarks.” Most algorithms for landmark identification focus on local, nonoverlapping regions of the face [12] or else create a joint distribution of potential landmarks over the whole face [13]. Here we rely on the distribution approach developed by [14]. This algorithm is fast (usable in real time), and its source code is publically available. Given the relatively low quality of our videos, it was necessary to modify the original code to improve the localization of the face in the image. A recently trained version of the [15] face detector algorithm was used for this purpose. Sixty-six landmarks were extracted from each frame. Figure 1 provides examples, and Figure 2 shows the landmark numbering.

As noted above, the eight targets (i.e. the subjects in the first phase of the experiment) were recorded covertly. As a consequence, some of the videos suffered from occlusions (e.g., a hand over the mouth) that made them problematic for the analysis of facial expression; see Figure 3 for examples. Relying on visual inspection, we eliminated all pairs of videos in which one or both included such defective frames; in addition one target was eliminated because she chewed gum throughout the experiment. The last row of Table 1 displays the number of surviving video pairs for each category.

Normalization Process

After pruning the data (as above) and performing landmark detection, each frame was normalized via the following procedure. First, the coordinates of the center pixel in each eye were computed as the mean of the six corresponding landmarks (37 to 42 for the left eye, and 43 to 48 for the right eye). All landmarks were then rigidly displaced so that the center of the left eye had coordinates (100,100). Second, the inter-eye distance d was computed and all landmark coordinates were multiplied by $100/d$. This sets the inter-eye distance to 100 pixels.

The beginning and end of a video often displayed exaggerated mobility and movement. This might be due to the cognitive

resources needed to engage the task when the image appears, and to disengage when a judgment is reached. To obtain greater stability, we analyzed just the middle third of each video, discarding frames from the first and last thirds. Other ways of defining a video’s “middle” section (e.g., by discarding frames from just the first and last quarters) yield similar results to those reported below. The use of thirds struck us as the most natural strategy, and we did not attempt to maximize our accuracy by choosing the boundaries accordingly.

Finally, we noticed greater facial mobility to *unattractive* stimuli in the *people* task, and to *noncute* images in the *animals* task. In the experiment [9], preferences were solicited on the basis of attractiveness and cuteness (not their reverse). We therefore switched the sense of preferences in these two domains (both involving the appeal of animate stimuli), and attempted to predict which face in a video pair expressed *less* preference for its stimulus. Specifically, we hypothesized that greater mobility would occur in target faces exposed to the *less* appealing stimulus in a pair. This reversal is left implicit in what follows.

Video Descriptors and Statistical Algorithms

For the data defined above, the goal of a candidate algorithm is to predict which of the two videos in a given pair is associated with preference (e.g., shows the target when s/he is viewing a cartoon that s/he subsequently designates as funnier than the alternative).

Our strategy is to compute a certain statistic for each video then predict the preference-video to be the one with higher value on the statistic. Two statistics were defined for this purpose; each is a plausible measure of the mobility of the face. To describe the two measures, let a video be composed of N frames, $f_1 \dots f_N$. For each frame f_i , define the *center* of f_i as the average x - and y -coordinates of the 66 landmarks appearing in f_i . Define the *dispersion* of f_i to be the average distance of the 66 landmarks to the center. We measured variation in dispersion through time via the following statistics.

M_{std} , the standard deviation of the set of dispersions manifested in the frames $f_1 \dots f_N$.

$M_{max-min}$, the difference between the maximum and minimum dispersions manifested in the frames $f_1 \dots f_N$.

We hypothesize that the video with more dispersion corresponds to the preferred picture (cartoon, etc.). Note that $M_{max-min}$

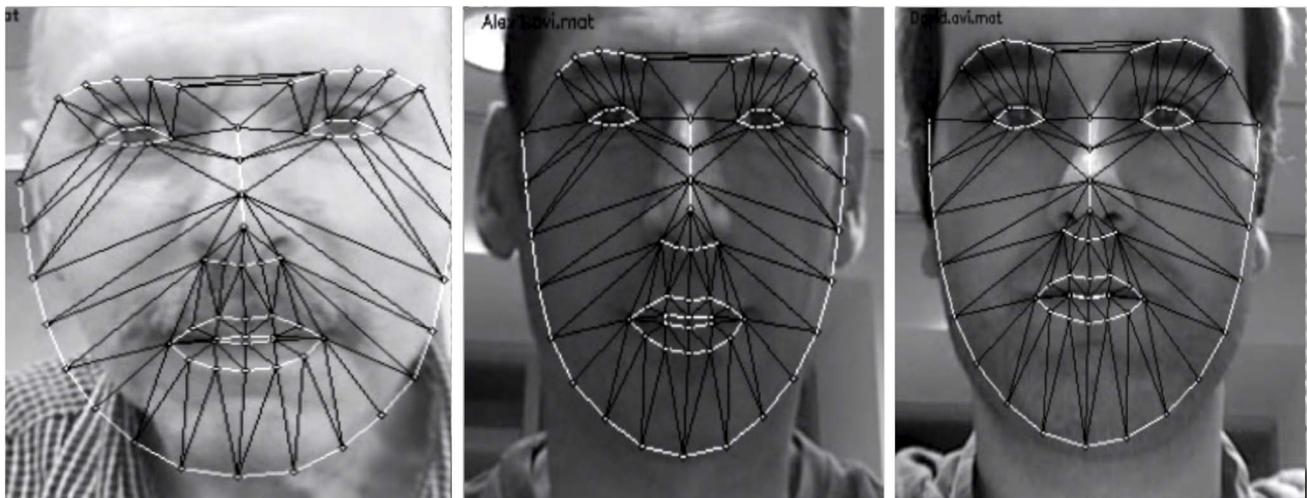


Figure 1. Examples of landmarks assigned to faces. Localization of the landmark points were fitted on the authors pictures (for illustrative purposes).

doi:10.1371/journal.pone.0087434.g001

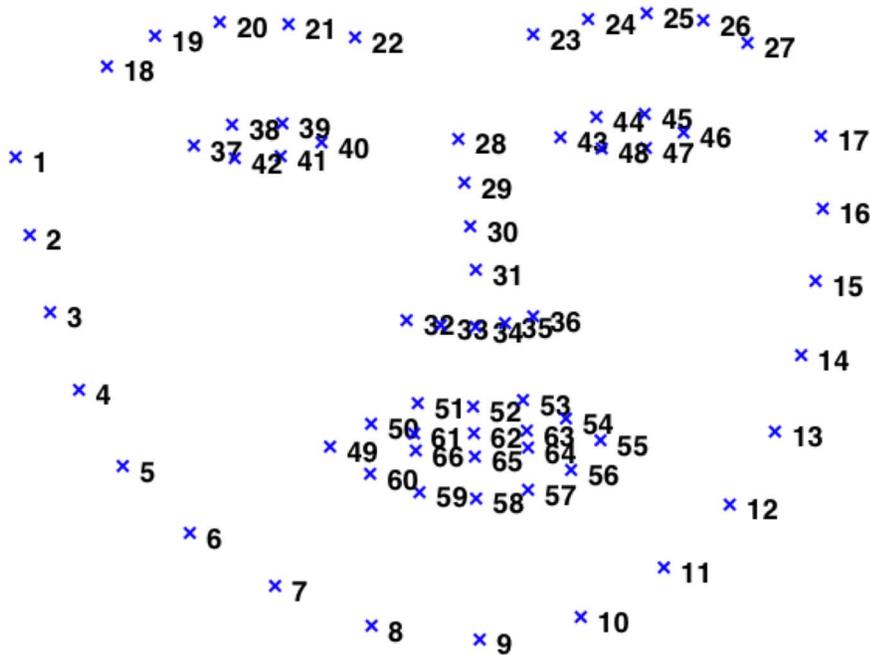


Figure 2. The numbering of the 66 landmarks on a typical face.
doi:10.1371/journal.pone.0087434.g002

is better able to exploit brief, extreme gestures (involving just a few frames) but is sensitive to noise in the landmark locations. M_{std} is more noise resistant because every frame contributes to its value. It is easily verified that the two measures are correlated insofar as the dispersion of the landmarks in time has a Gaussian distribution. Notice that the algorithms based on these statistics do not exploit the temporal order of the frames $f_1 \dots f_N$.

Results of Statistical Algorithms

For each of the four domains, Table 1 shows the percent of video pairs that M_{std} and $M_{max-min}$ accurately label. We did not

apply learning in this first baseline experiment. Instead, each video is predicted as the chosen one if the value of the single statistic (M_{std} or $M_{max-min}$) is the highest in the pair. The ground truth labels are the original choices of the target participants. To illustrate, $M_{max-min}$ correctly labeled two thirds of the cartoons. As a comparison, we computed the probability of obtaining the same or greater success by throwing a fair coin in response to each pair of videos. For example, the probability of such a coin-flipper reaching at least the level of accuracy shown by $M_{max-min}$ on Cartoons is only 0.004 (via a binomial test). Pooling all 235 pairs of videos across the four domains, $M_{max-min}$ correctly classified 58.3% ($P < 0.013$) and M_{std} correctly classified 58.7% ($P < 0.005$).

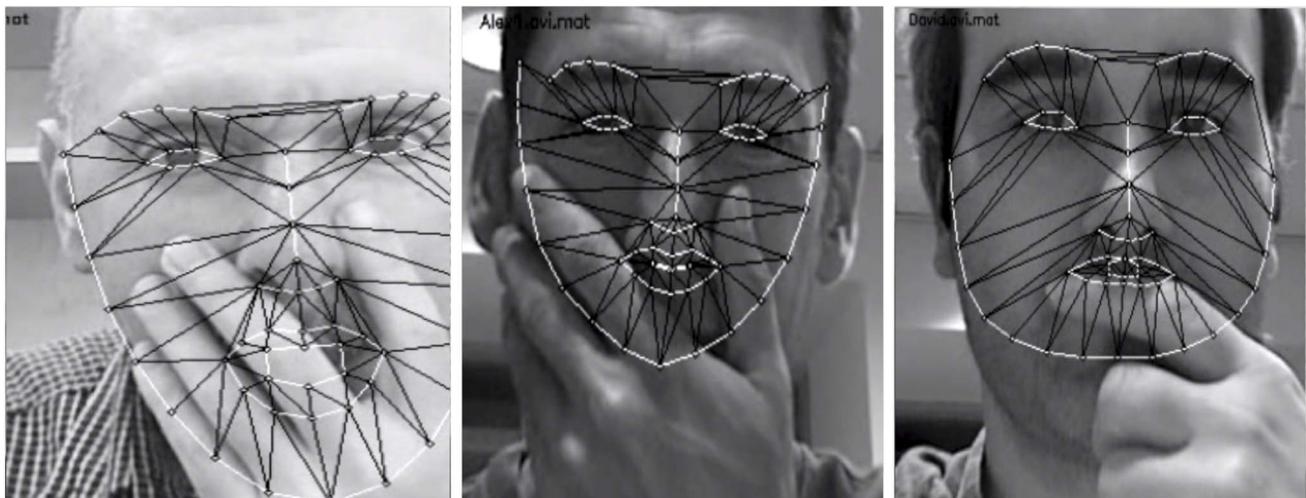


Figure 3. Examples of landmark distortion due to partial occlusion. Given that the participants were unaware of being recorded, some videos presented occlusions that prevented their further processing. The figure shows examples of these distortions on authors' pictures for illustrative purposes.

doi:10.1371/journal.pone.0087434.g003

Table 1. Percent accuracy on the four domains.

	People	Cartoons	Paintings	Animals
$M_{max-min}$	59.1 ($P < 0.088$)	66.7 ($P < 0.004$)	50.0 ($P < 0.556$)	54.0 ($P < 0.336$)
M_{std}	62.1 ($P < 0.032$)	60.9 ($P < 0.046$)	56.0 ($P < 0.240$)	54.0 ($P < 0.336$)
JESP	52.4	65.5	56.2	59.0
# Videos pairs	66	69	50	50

doi:10.1371/journal.pone.0087434.t001

The row labeled “JESP” in Table 1 shows the results obtained by the human perceivers studied in [1]. The row is relative to just the 235 pairs of videos that are free of occlusions and gum-chewing. Performance is similar when all 768 videos are included (as in [1]); with all the data, accuracy is 54.7%, 67.6%, 56.1% and 54.8% for the four domains, respectively.

Overall, the table reveals better-than-chance performance by $M_{max-min}$ and M_{std} for *people* and *cartoons* but scant accuracy for *paintings* and *animals*. Human perceivers do not perform much better than these simple algorithms. To explore the matter further, for each of the 235 pairs of videos, we define the *human accuracy* for that pair to be the percentage of correct classifications on the part of the fourteen perceivers who evaluated that pair. Likewise, we define the $M_{max-min}$ *difference score* to be the difference between the $M_{max-min}$ score on the first minus the second videos – and similarly for M_{std} . The correlation between human accuracy and the $M_{max-min}$ difference score is only 0.04; for M_{std} it is only 0.06. These low correlations suggest little agreement between human and algorithmic inferences. In turn, the low agreement suggests the possibility of designing algorithmic predictors of preference that are more accurate than those offered here.

SM Classification and Results

We next sought to determine whether prediction can be improved by submitting the data to a learning algorithm. Instead of using a single value to describe the average dispersion of the landmarks, we compute the proposed descriptors ($M_{max-min}$ and M_{std}) on each landmark independently. We allow the learning algorithm to weight the contribution of each landmark to the preference prediction. From this perspective we consider each of the 235 pairs of videos to be a sample in a classification problem. The label on a given sample is either 1 or 0 depending on whether the first or second video shows the target’s preference-face. For each pair of videos, we constructed a feature vector for that pair via the following procedure. Let individual video V be composed of N frames, $f_1 \dots f_N$.

- Compute the center c_i of each frame f_i as the average x - and y -coordinates of the 66 landmarks in f_i .
- For each landmark j in frame i , compute the Euclidean distance from j to the frame-center c_i . Gathering these computations for landmark j across the frames $f_1 \dots f_N$ yields a real vector of length N ; the vector records the changing distances between j and the frame centers c_i . There are 66 such vectors, one for each landmark.
- For each of the 66 vectors, compute the difference between its maximum and minimum value across the N frames. In the same way, for each of the 66 vectors compute the standard deviation of its values. Concatenating the two resulting vectors – 66 max-min statistics followed by 66 standard

deviations – yields a 132-dimensional feature vector \vec{V} for the starting video V .

- Given a pair (V, W) of videos, the feature vector for the pair is defined to be $\vec{V} - \vec{W}$, the coordinate-wise difference between the features of V and W .

Relying on these features, a nonlinear Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel [16,17] was applied as a classification rule on the video pairs available in each of the four domains separately. We executed 10 random iterations of a 10-fold cross validation protocol to assess the results. Folds were constructed balancing the number of samples from each class. The dimensionality of the data was reduced by applying Principal Component Analysis on the training set (preserving 99% of the variance). In order to estimate the parameter σ (for the RBF Kernel) and the soft margin C (for SVM), only the training data were used. The 90% of the data reserved for training was split into two subsets, 80% for internal training and 20% for internal validation. The SVM/RBF algorithm was then applied to the 10% testing data, using the two fixed parameters. Table 2 shows the results of 10 applications of the algorithm in this way. It can be seen that predictive accuracy is only slightly higher than for $M_{max-min}$ and M_{std} (applied without training).

Conclusion

In this paper we introduce the problem of automatically inferring preferences from spontaneous facial expressions. We make available an annotated database, and propose baseline methods to infer preferences. The simple descriptors $M_{max-min}$ and M_{std} perform better than chance in two domains (*people*, *cartoons*), and at approximately the same modest level as human perceivers. Classification based on a standard learning algorithm yields only limited improvement. The question immediately arises whether the faces in [1] hold further information that can be exploited to reveal preference. Developing more successful algorithms than ours would provide an affirmative answer. Failure would suggest that faces are often opaque, and it would invite hypotheses about which social circumstances allow more emotional information to invade the face. Research in this area provides a rare point of convergence between Computer Science and Social Psychology.

Table 2. Results using SVM/RBF: mean accuracies and 95% confidence intervals.

	People	Cartoons	Paintings	Animals
SVM-RBF	59.4 ± 3.2	65.4 ± 3.3	58.6 ± 4.5	45.8 ± 2.7

doi:10.1371/journal.pone.0087434.t002

Author Contributions

Conceived and designed the experiments: DM MN AT DO. Performed the experiments: DM MN DO. Analyzed the data: DM AT DO.

Contributed reagents/materials/analysis tools: DM MN DO AT. Wrote the paper: DM DO AT.

References

1. North MS, Todorov A, Osherson DN (2010) Inferring the preferences of others from spontaneous, low-emotional facial expressions. *Journal of Experimental Social Psychology* 46: 1109–1113.
2. North MS, Todorov A, Osherson DN (2012) Accuracy of inferring self-and other-preferences from spontaneous facial expressions. *Journal of Nonverbal Behavior* 36: 227–233.
3. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: A literature survey. *Acm Computing Surveys (CSUR)* 35: 399–458.
4. Rojas Q, Masip D, Todorov A, Vitria J (2010) Automatic point-based facial trait judgments evaluation. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE*, 2715–2720.
5. Rojas M, Masip D, Todorov A, Vitria J (2011) Automatic prediction of facial trait judgments: Appearance vs. structural models. *PloS one* 6: e23323.
6. Brahnam S, Nanni L (2009) Predicting trait impressions of faces using classifier ensembles. In: *Computational intelligence*, Springer. 403–439.
7. Brahnam S (2005) A computational model of the trait impressions of the face for agent perception and face synthesis. *AISB Journal* 1: 481–508.
8. Fasel B, Luetttin J (2003) Automatic facial expression analysis: a survey. *Pattern Recognition* 36: 259–275.
9. Tian YI, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23: 97–115.
10. Ekman P (1993) Facial expression and emotion. *American Psychologist* 48: 384–384.
11. Ekman P, Friesen WV (1977) Facial action coding system. Consulting Psychologists Press, Stanford University, Palo Alto.
12. Vukadinovic D, Pantic M (2005) Fully automatic facial feature point detection using gabor feature based boosted classifiers. In: *Systems, Man and Cybernetics, 2005 IEEE International Conference on. IEEE*, volume 2, 1692–1698.
13. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23: 681–685.
14. Saragih JM, Lucey S, Cohn JF (2009) Face alignment through subspace constrained mean-shifts. In: *Computer Vision, 2009 IEEE 12th International Conference on. IEEE*, 1034–1041.
15. Viola P, Jones MJ (2004) Robust real-time face detection. *International journal of computer vision* 57: 137–154.
16. Hearst MA, Dumais S, Osman E, Platt J, Scholkopf B (1998) Support vector machines. *Intelligent Systems and their Applications, IEEE* 13: 18–28.
17. Schölkopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization and beyond. the MIT Press.