On Constructing Facial Similarity Maps
** Supplementary Material **

Alex Holub
California Institute of Technology
Pasadena, CA
holub@caltech.edu

Yun-hseuh Liu

Pietro Perona

1. Consistency of Raters

We wish to measure the consistency of our subjects during both the relative and absolute rating methods. We had 5 subjects rate images using the relative method and a different set of 5 subjects rate images using the absolute method.

1.1. Absolute Ratings

Recall that during an absolute rating subjects are required to choose a number from 1 to 7 indicating how similar two faces are to one another. Figure 1 shows the number of times each rating was chosen by each subject. The right-skew in the distribution indicates that subjects are more likely to say faces are dissimilar than similar. In order to assess how consistent subjects were we interleaved trials in which subjects were shown the same two images and required to assess their similarity. If subjects are perfectly consistent they will always indicate the same similarity number (from 1 to 7) between the images. If they are not consistent they will indicate a different number. Figure 2 shows results across 5 subjects for both the condition when they see the hair of the faces and when they do not. We note that subjects seem to be reasonably consistent in rating the similarity of images, and they are even reasonably consistent across subjects.

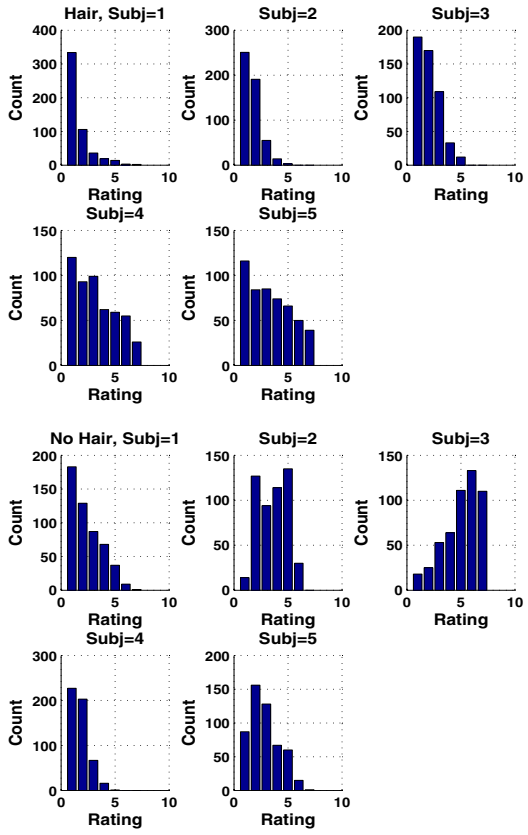

Figure 1. What was the distribution of ratings for the absolute rating task? (Top Set) Faces shown with hair. Each plot is a different subject. The x-axis is the particular rating value chosen (in the range from 1 to 7). The y-axis is the number of times this rating was chosen. Note the right-skew distribution. Subjects were not equally likely to chose any particular rating. Subjects were most willing to say faces were very dissimilar (rating 1). (Bottom Set) Same when images were shown without hair.

1.2. Relative Ratings

We would like to analyze consistency for relative ratings as well. How should we proceed? Again we have interleaved trials where the subject sees the same set of faces and must choose which, from a set of 24 faces, is most similar to a target face. We have a total of 10 sets of images which we repeat 4 times as in the absolute experiments. For one set of images consider each of the 4 trials. The subject may pick 4 different images as being similar, or may consistently pick the same image as being similar. The latter case is, of course, preferable. To graphically illustrate performance, we consider the number of unique groups which each subject selects. In the first case there would be 4 groups (the subject chose 4 unique images), while in the latter there would be only a single group (the subject always chose

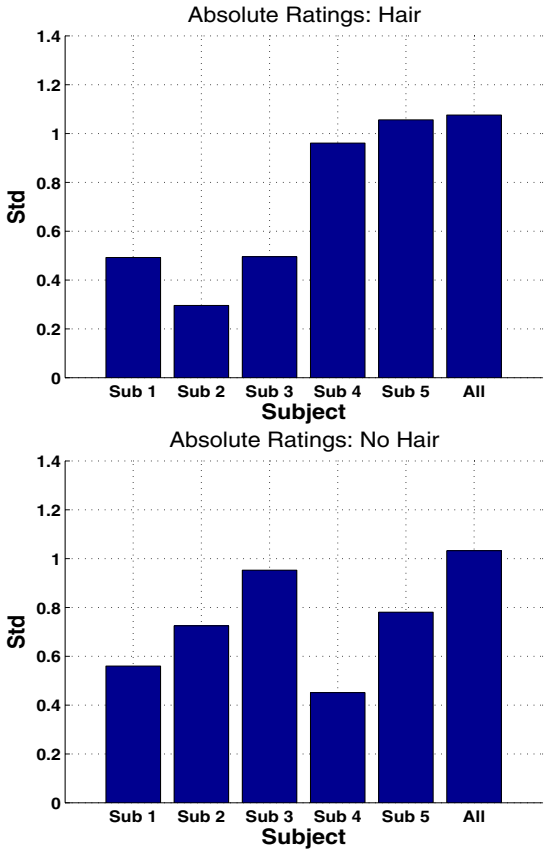

Figure 2. (Top) Results when subjects are shown images with hair. There were a total of 10 unique pairs of faces which were repeated 4 times. The order in which faces were shown to subjects was the same. The repeat pairs were interleaved among a total set of 514 trials. We measured the standard deviation (std) in the responses of each subject to the repeated pairs. A low std indicates that the subjects are consistent, they always chose the same number. We averaged the std across all pairs and plot the results as the first 5 bars. The last bar indicates the std across all subjects and gives an idea to inter-subject consistency. Here the mean score of each subject is subtracted from each score and the std is taken across all subjects. Although there does seem to be some inter-subject variability it does not appear to be extremely drastic. (Bottom) Same as top but when the subjects are shown images without hair. Note that there does not seem to be an appreciable decline in performance.

the same image). We then look at the cardinality of these groups and sort them in decreasing order. Figure 3 shows results averaged over the 10 sets of repeated trials. Note that subjects seem to be reasonably consistent in their choice of the most similar face.

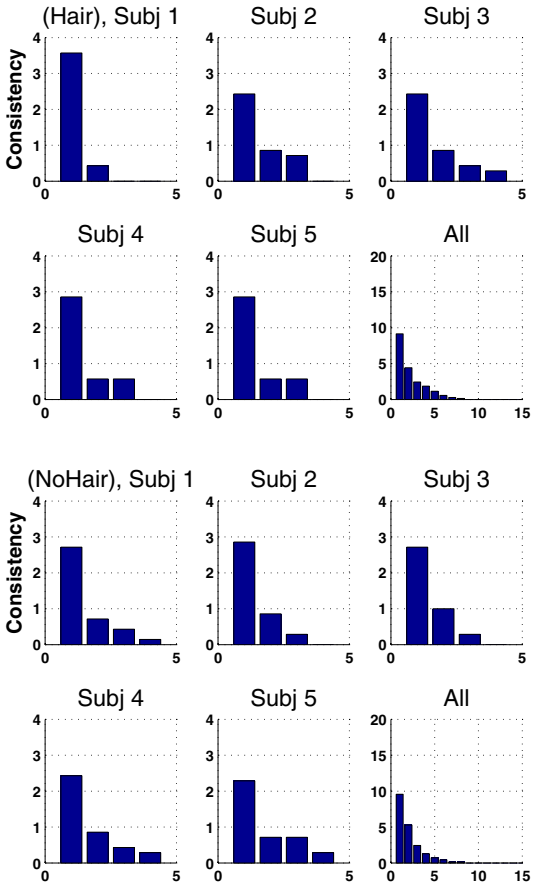

Figure 3. (Top Set) Results when subjects are shown images with hair for relative rating rating experiments. The consistency of the 5 subjects in the relative rating scheme. Perfect performance would be indicated by a single bar of height 4: the subject always picks the same image during the relative experiment. The final bar plot indicates how consistent subjects are between each other. In this case the best performance would be a 16 (4 subjects $\times 4$ interleaved experiments. Again we see that subjects are reasonably consistent between one another: different subjects tend to pick the same faces as being most similar to the target. (Bottom Set) Same as top but when the hair is not shown. The results between hair and no hair conditions did not seem significant.