FACE-GRAPH MATCHING FOR CLASSIFYING GROUPS OF PEOPLE

Henry Shu[1], *Andrew Gallagher*[1], *Huizhong Chen*[2], and *Tsuhan Chen*[1]

[1]Cornell University, USA
[2]Stanford University, USA

ABSTRACT

When people gather for a group photo, they are together for a social reason. Past work has shown that these social relationships affect how people position themselves in a group photograph. We propose classifying the type of group photo based on the spatial arrangement and the predicted attributes of the faces in the image. We propose a matching algorithm for finding images from a training set that have both similar arrangement of faces and attribute correspondence. We formulate the problem as a bipartite matching problem where the faces from each of the pair of images are nodes in the graph. Our work demonstrates that face arrangement, when combined with attribute (age and gender) correspondence, is a useful cue in capturing an approximate social essence of the group of people, and lets us understand why the group of people gathered for the photo.

1. INTRODUCTION

People often gather for a photo shot for an underlying social reason. Past works have shown that the spatial arrangement of the faces in a photo provides useful cues as to predicting certain attributes of the faces ([1, 2]). In addition, when harnessed properly, the pairwise spatial positions of faces can also give useful information in predicting the relationships of the individuals in the photo ([2]). Of course, the social relationships and events under which a photo was taken can affect how we humans might categorize the group. Motivated by this observation, our goal in this work is to investigate the relationship between the spatial arrangement of faces in a photo and the type of group that has assembled.

Consider the four photos (a) - (d) in Fig. 1, in which all visual content is removed except the faces, their relative sizes, and some age or gender clues. Using only this information, the reader can attempt to categorize each photo with an appropriate label on the right. Compare your answer to the results shown in Fig. 2. How many photos did you classify correctly? The reader will probably score much better than the 25% random choice. As these simple examples show, facial arrangement and attributes provide an important cue useful for photo type classification. In this work, we demonstrate the usefulness of this cue for group photo classification.

Related work. The usefulness of facial arrangements have been explored before for predicting attributes on single people [2] and relationships between pairs of people [3, 4, 5]. Also, in [1], facial arrangement was used to measure the similarity of two photos. Although the task was for image clustering rather than classification, and the goal was targeted towards human-subjective ranked retrieval assessment, the motivation that facial arrangement has to do with photo similarity is the same. [2] used the least square fit of face sizes and positions to detect group dining photos. Most of these works involve the use of facial positions and attributes. More traditionally, image classification is often conducted with appearance-based features. Examples include face attribute classification [6, 7], occupation prediction based on the kind of clothing a person wears [8], cultural type and urban tribe classification [9]. However, we believe that classifying consumer photos should

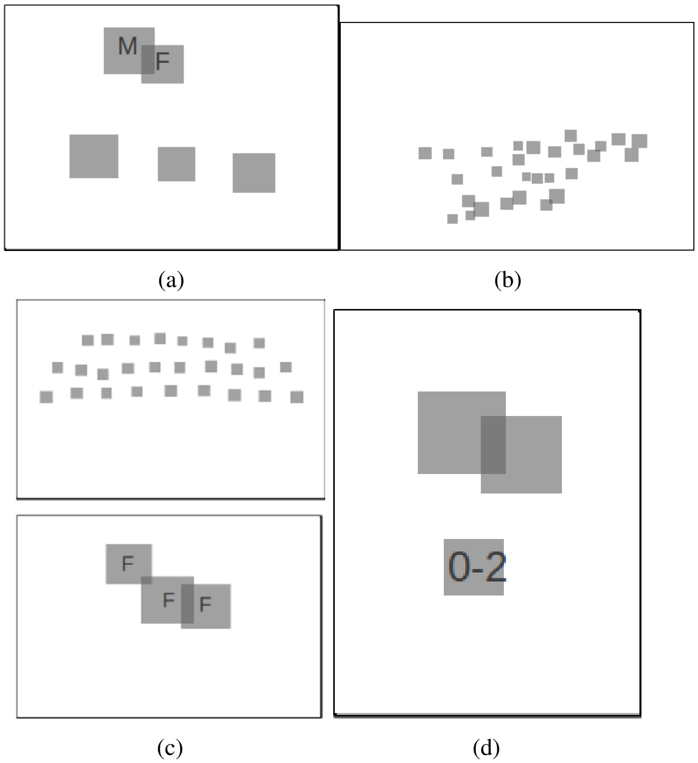
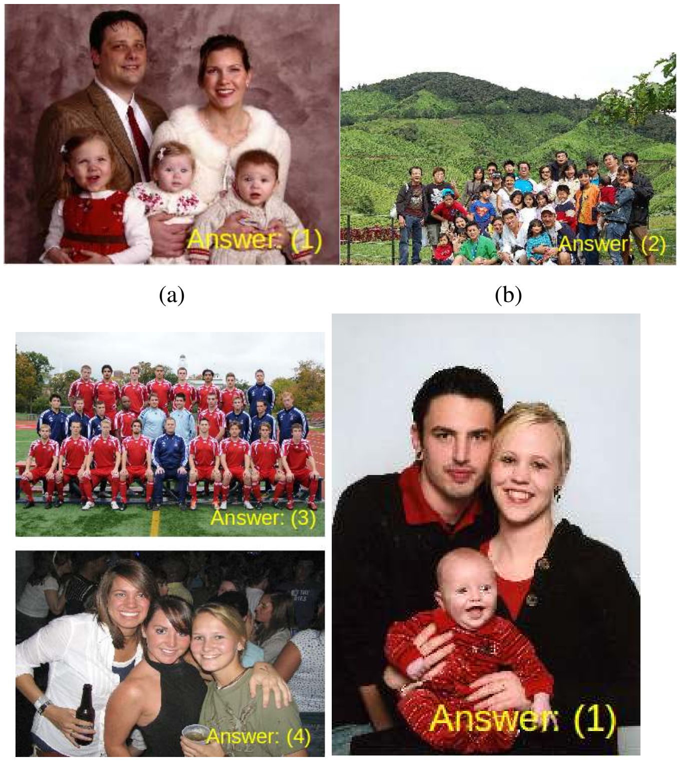


Fig. 1: The label space is (1) Family, (2) Group Field Trip, (3) Sports Team, and (4) Friends Hanging Out



(a) (b) (c) (d)

Fig. 2: Answers to Fig. 1. The label space is (1) Family, (2) Group Field Trip, (3) Sports Team, and (4) Friends Hanging Out

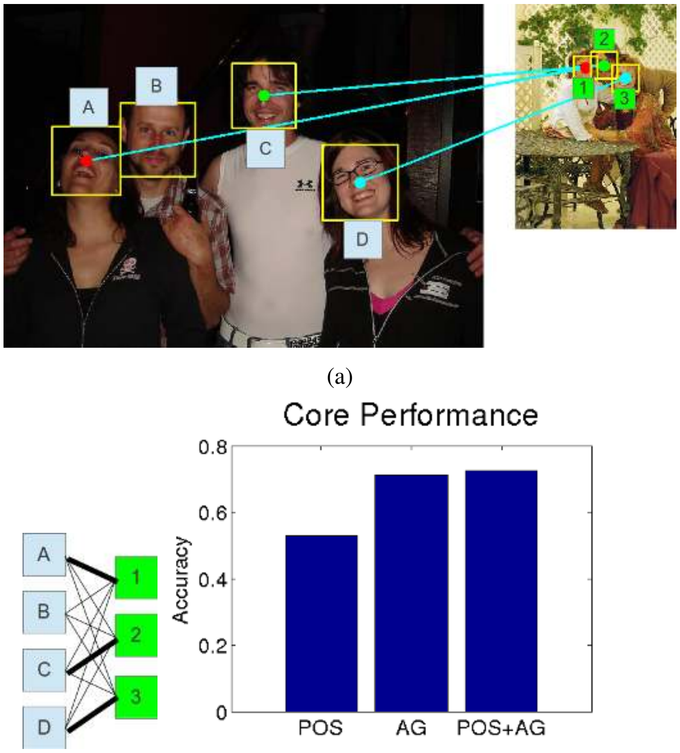
be based on the humans. After all, they are the protagonists of a story the photo tries to convey.

2. GROUND TRUTH AND DATA COLLECTION

Previous work provides some datasets of photos in which something about the photo type is known. For example, [2] gives a rough photo categorization of group, family, and wedding. Often in the previous work, the photo type was derived from the tags that were associated with each photo or the search query terms used in retrieving the photo from such services as Flickr or Google Images. Naturally, the photo types derived this way may be somewhat ambiguous. Indeed, a wedding photo may well be a family photo.

To simplify matters, we desire a dataset in which the photo types are as unambiguous as possible. In addition, we seek the types of photos that are common enough to be of sufficient interest as consumer photos. With these goals in mind, a few experimental inspections indicate that four categories of photos fit our objectives well. They are *family*, *group field trip*, *sports team*, and *friends outing*.

We collected 10K photos. Some are from [2] and some are from online image services with relevant keyword queries. A group of human subjects then pick the 1K most unambiguous, properly fit photos for the 4 categories (250 each). For examples of these photos, see Fig. 5.



(a)

(b) (c)

Fig. 3: (a) and (b) are two sample photos showing the face bipartite graph. (c) is the core experiment result.

3. METHOD

Our method works by measuring the spatial similarity of the facial arrangement and the attribute similarity of the faces of the photos. Specifically, let us first consider computing the similarity score of two photos.

3.1. Bipartite Matching

For the two photos shown in Fig. 3(a), first detect the face bounding boxes. Then, we represent the faces as the nodes of a bipartite graph as shown in Fig. 3(b). Associated with each edge is a weight $w_{i,j}$ that captures the cost of matching the respective pair of faces, face i from the first photo and face j from the second, as a corresponding pair. Then, we find a maximum assignment (one in which the node set, say the right hand side in this example, with the smaller number of nodes has all its nodes matched) of minimum weight. Naturally, a matching must respect the one-to-one relationship. This is an example of the minimum weight bipartite assignment problem, which can be readily solved by the Hungarian

algorithm [10, 11].
Edge weights are determined according to Eq 1 with the intent that a smaller weight implies a higher degree of similarity. Here, the weight of edge (i, j) is a linear combination of the positional term and the attribute terms. The weights of the other edges are computed in a similar fashion.

$$w_{i,j} = \alpha \|\mathbf{x}_i - \mathbf{x}_j\| + \sum_l \beta_l h_l(a_l(i), a_l(j)) \qquad (1)$$

Positions. The faces coordinates of each image are first normalized so that the median face sizes in the two images are the same. Then, the faces coordinates within each photo are mean removed. The norm of the positional difference is weighted by α.
Attributes. Each face is associated with it a set of attributes indexed by l. For instance, $a_l(i)$ is the value of attribute l of face i. Function h_l computes the difference of two attribute l values. Each attribute difference is weighted by β_l.
Let us denote by w^* the sum of the weights of the matching edges selected. Fig. 3(b) shows the optimal set of edges selected using positions alone ($\alpha_i = 0$ for all i) as the darkened edges.

3.2. Face Number Discrepancy

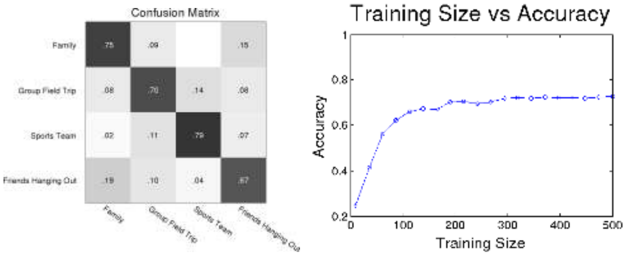
Notice that this matching algorithm does not require the two photos to have the same number of faces. Such requirement would be too stringent. Firstly, the number of training data would greatly decrease when partitioned into photos of different numbers of faces. Secondly, as are evident in the examples we show, photos of different numbers of faces may still have structural similarity that is relevant. On the other hand, allowing face number discrepancy may be unfair in certain cases. Indeed, a two-person photo is very likely to match well with a group photo simply due to chance. To address these issues, we pay an additional cost for any face number discrepancy and define the final similarity of two photos I_1 and I_2 as

$$d(I_1, I_2) \equiv w^*_{I_1,I_2} + \gamma |I_1 - I_2|_{\text{face}}, \qquad (2)$$

where $w^*_{I_1,I_2}$ is the weight of the optimal matching of I_1 and I_2, and $|I_1 - I_2|_{\text{face}}$ is the difference of the numbers of faces in the photos.
With the ability to compute the pairwise similarity score between any two photos, we can readily use many standard classifiers to complete the photo classification algorithm. For simplicity, we use k-NN through this work.

4. EXPERIMENTS

Here, we describe the experiments we conduct to evaluate our method. We randomly split our dataset into 50% for training and 50% for test. The face attributes we use are age and gender. We use the predictions from the algorithm proposed by [12], in which there are 7 age bins roughly representing different stages of life. For h_{gender} we use the binary gender difference, and for h_{age} we use the bin difference which roughly captures how far apart two ages are. Finally, we use leave-one-out cross validation on the training set to tune the parameters required in our algorithm.



(a) The confusion matrix for POS+AG

(b) How training size affects test accuracy.

Fig. 4

4.1. Main

The core performance results are summarized in Fig. 3(c). We carry out three sets of experiments. For the positions only experiment, which we denote by POS, we turn off the face attributes ($\beta_{\text{gender}} = \beta_{\text{age}} = 0$). For the age and gender experiment, denoted by AG, we turn off the position contribution ($\alpha = 0$). When combining positions with age and gender, which we denote by POS+AG, we attain an accuracy of 72.6%. Fig. 4(a) shows the confusion matrix of the 72.6%-accuracy experiment. Compared to the 52.8% accuracy of POS, AG achieves an accuracy of 71.2%. It is perhaps not too surprising that age and gender attributes seem to play an important role in photo classification. From human intuition there is no shortage of plausible reasons behind it. Indeed, a family photo tends to include a wider range of faces of disparate age ranges and genders. On the other hand, a photo of a group of friends hanging out tends to have most of the faces belonging to roughly the same age groups, and they tend to be either all males or all females. Likewise, for a sports team photo, in most of the cases it consists of a majority of all males or all females, as official sport teams are rarely coed. For a group photo of field trip, the faces usually contain a more even mix of males and females, and the age range is wider as well. Positional cues have their own merits, however. The performance of POS at 52.8% outperforms the random guess accuracy of $1/4 = 25.0\%$. Considering that no appearance-based cues are used in POS, this result quantitatively supports our hypothesis that facial position arrangement gives a nontrivial cue that can be helpful for photo type classification. In addition, POS+AG does give a significant, albeit small, improvement of 1.4%.
We also use purely appearance-based features as a rudimen-

tary baseline for our task of photo type classification. The feature we use are GIST [13], with RBF SVM as the classifier whose parameters we tune by cross validation on the training set. The performance result is a mediocre 42.3%, compared apple-to-apple with those results shown in Fig. 3(c). We do not find this result surprising. After all, a dominating factor for determining the type of a consumer photo is the humans in the photo. As such, a method that directly analyzes the humans in the photo may likely work better than one that does not.

4.2. Horizontal Symmetry

It is interesting to point out that we can effectively double the number of training data for POS by symmetrically flipping each training photo left-to-right. This observatoin comes from the assumption that, everything else being equal, people have no preference for the left or right side in a photo shot. Indeed, allowing such symmetry turns out to improve the performance of POS by 2-3%. Throughout this work, the experiments we conduct for POS and POS+AG take advantage of this horizontal symmetry assumption.

4.3. Effect of Training Size

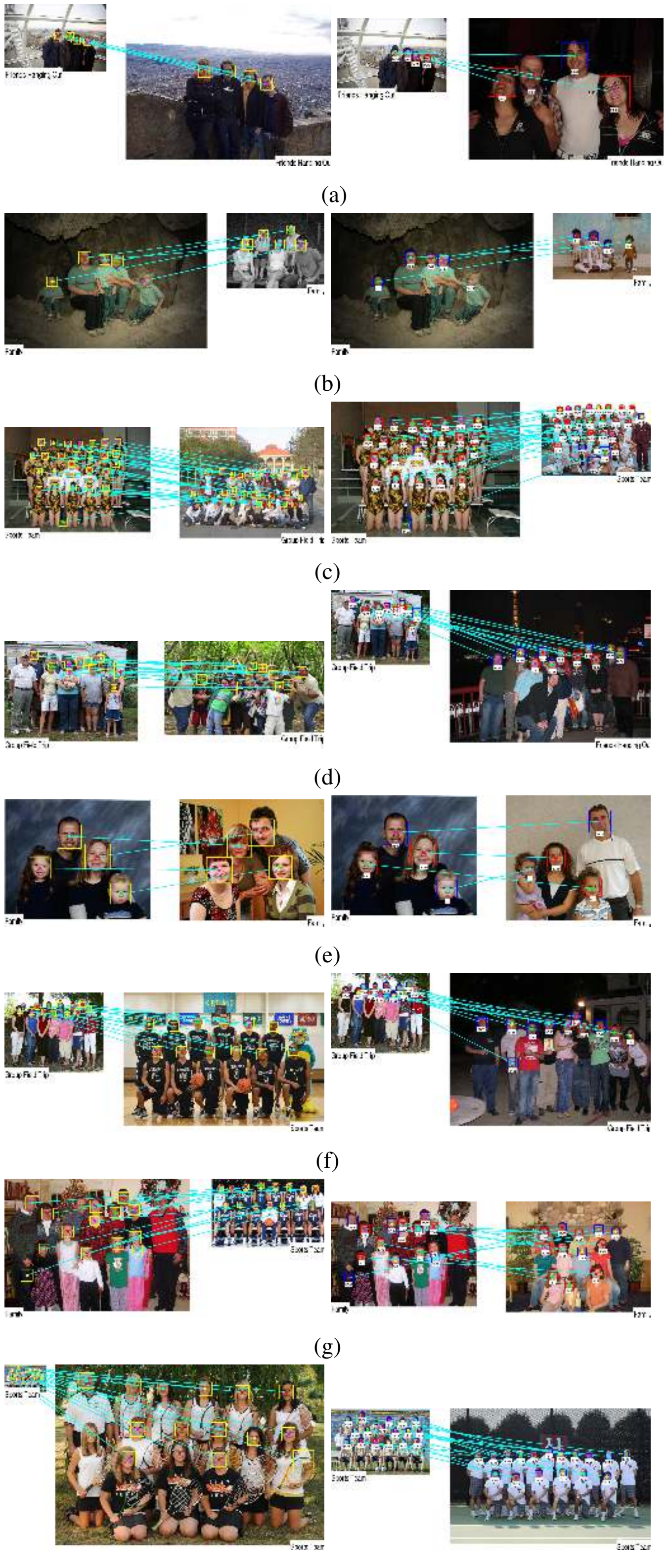
While fixing the same test set, we artificially change the size of the training set down to as few as 10 photos. Fig. 4(b) gives the result for both POS and AG. In both cases, we see that the accuracy of $> 50\%$ for POS and that of $> 70\%$ for AG are both achieved in as few as 100 and 200 training photos, respectively. In general, we find that contextual cues usually require much less training data than appearance-based features to attain their optimal classification performance.

5. CONCLUSION

In this work, we demonstrate the usefulness of facial arrangement and attribute (age and gender) cues in photo classification. Of course, there are limitations. For example, we may wonder how likely a truly randomly chosen consumer photo from the internet will be, say, a field trip photo given that it is quintessentially similar to the field trip photos in the training set. Nevertheless, the performance results from our work confirms the benefits of such contextual cues and encourages future work to build on it.

Acknowledgements

This work was supported in part by National Science Foundation grant DMS-0808864.


(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5: Shown in each of (a) - (h) are two image pairs. In each pair, the left image is the test query and the right is its most similar image from the training set. The left pair of images is based on POS, and the right pair of images is based on POS+AG, in which the gender and age predictions are shown as well. The ground truth photo types are provided at the bottom of each image. Best viewed in magnification in color.

6. REFERENCES

[1] M. Abdel-Mottaleb and L. Chen, "Content-based photo album management using faces' arrangemen," in *Proc. ICME*, 2004.

[2] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009.

[3] P. Singla, H. Kautz, J. Luo, and A. Gallagher, "Discovery of social relationships in consumer photo collections using markov logic," in *Proc. CVPRW*, 2008.

[4] X. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," .

[5] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing people in social context: recognizing people and social relationships," in *Proc. ECCV*, 2010.

[6] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. ICCV*, 2009.

[7] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Describable visual attributes for face verification and image search," in *Proc. PAMI*, 2011.

[8] Z. Song, M. Wang, X.S. Hua, and S. Yan, "Predicting occupation via human clothing and contexts," in *Proc. ICCV*, 2011.

[9] A. C. Murillo, I. S. Kwak, L. Bourdev, D. Kriegman, and S. Belongie, "Urban tribes: Analyzing group photos from a social perspective," in *Proc. CVPRW*, 2011.

[10] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, 1955.

[11] Alexander Melin, "http://www.mathworks.com/matlabcentral/fileexchange/11609," .

[12] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. CVPR*, 2013.

[13] A. Oliva abd A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," in *Proc. IJCV*, 2011.