FACE2GPS: ESTIMATING GEOGRAPHIC LOCATION FROM FACIAL FEATURES

Mohammad T. Islam, Scott Workman, Nathan Jacobs

Department of Computer Science, University of Kentucky

{tarik, scott, jacobs}@cs.uky.edu

ABSTRACT

The facial appearance of a person is a product of many factors, including their gender, age, and ethnicity. Methods for estimating these latent factors directly from an image of a face have been extensively studied for decades. We extend this line of work to include estimating the location where the image was taken. We propose a deep network architecture for making such predictions and demonstrate its superiority to other approaches in an extensive set of quantitative experiments on the GeoFaces dataset. Our experiments show that in 26% of the cases the ground truth location is the topmost prediction, and if we allow ourselves to consider the top five predictions, the accuracy increases to 47%. In both cases, the deep learning based approach significantly outperforms random chance as well as another baseline method.

Index Terms— facial features, image localization

1. INTRODUCTION

Facial appearance varies dramatically across the globe. This variation depends not only on ethnicity (which is strongly dependent on geo-location), but also on factors such as presence of facial hair, clothing, and expression. The relationship between geo-location and human face appearance provides strong cues which can be leveraged to predict where a person is most likely to be from [1, 2]. Exploring this relationship is an emergent research direction with a wide range of potential applications. For example, a model characterizing this relationship could be used to aid in security and surveillance applications that attempt to determine the identity of a suspect, detect individuals that are out of place, or in defense applications for rapid deployment of an individual to an unknown location. More so, it could be applied to improve the performance of existing image localization algorithms [3, 4], which completely ignore faces as a cue for localization.

We propose a data-driven approach to solving the problem of image geo-localization using an image of a face. We construct a large dataset of face images associated with 50 cities from around the world. Using these images, we apply a deep convolutional neural network (CNN) to predict the location of the city where the image was captured. In Figure 1, we show several example query faces and the probability distribution over cities predicted by our method.

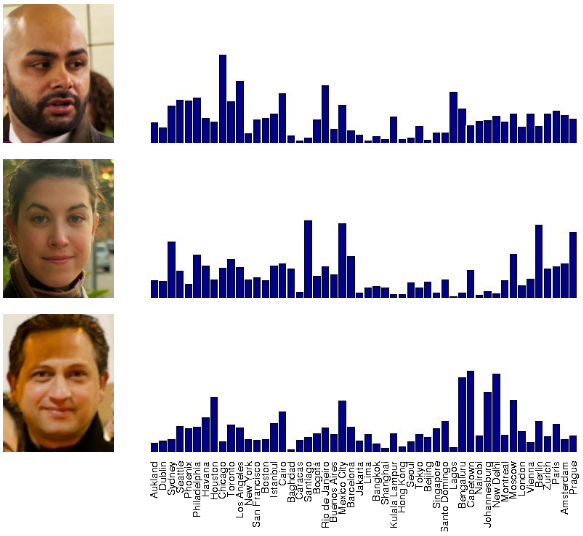


Fig. 1: Latent facial appearance factors provide strong cues for estimating the geographic location of an image.

Our main contributions are as follows: 1) we describe a novel approach which uses a deep convolutional neural network for estimating the geo-location of an image from facial appearance, 2) we present quantitative and qualitative results of our methods versus several baseline approaches on a new dataset of geo-tagged face images, and 3) we analyze the results and make interesting observations in addition to our target task.

2. RELATED WORK

We summarize related work in facial analysis and image localization below.

Facial Image Analysis Automatic analysis has been performed on geometric aspects of the face such as the shape and texture [5], as well as latent factors like race and ethnicity [6, 7, 8, 9, 10], age [11, 12], and attractiveness [13, 14, 15]. More recently, with the emergence of deep learning, faces have been analyzed using various types of deep networks for recognition [16, 17], verification [18], and expression recognition [19].

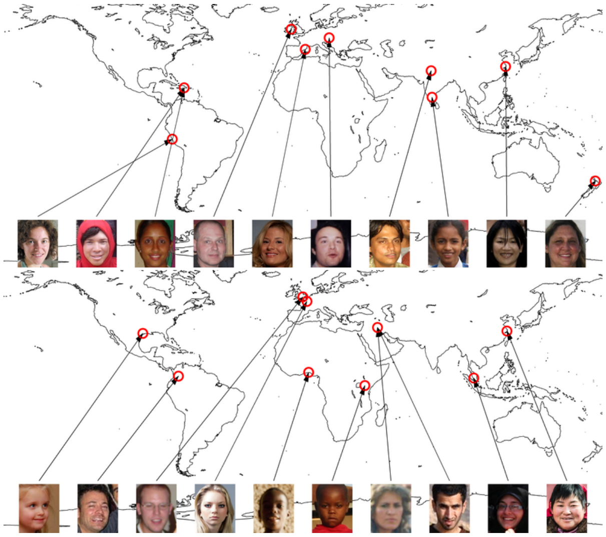


Fig. 2: The most likely locations for a diverse set of faces predicted by our proposed approach.

Image Localization The problem of image localization has been addressed using a wide variety of cues [20, 4, 21, 22]. Hays and Efros [3] use global image features, such as color and texton histograms, to estimate a distribution over geographic location for a query image using a large dataset of geo-tagged reference images. Lin et al. [23] exploit the relationship between ground and aerial image to localize a query image. Baatz et al. [24] use the detected skyline to estimate location by comparing to a digital elevation model. Other approaches use local image descriptors [4, 25]. Many other cues exist such as solar shadow trajectories [26], aggregate light levels [27], and sky appearance [28, 29].

Our work is the first to study the ability of deep networks to estimate the geographic location of an image using facial appearance.

3. ESTIMATING THE GEO-LOCATION OF A FACE

We address the problem of estimating the geographic location of an image given the face of an individual in the image. The first step of our approach is to detect, extract, and normalize the face. We then predict the geographic location using a deep convolutional neural network operating directly on the pixel values of the normalized face patch.

3.1. Pre-Processing

Given an image, we first detect the contained faces using a commercial face detector[1]. The face detector outputs the pose and the confidence of the detected faces along with the locations and confidences of some predefined fiducial points. To normalize the pose of a face, we use the fiducial points corresponding to the centers of the eyes as control points for a similarity transform to align the eyes of the face patch to those of a manually selected, near-frontal face patch.

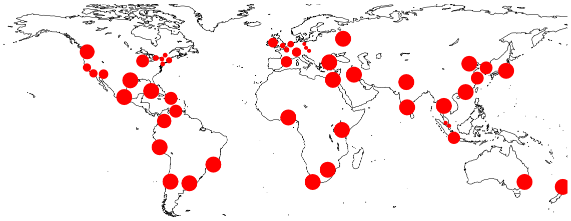


Fig. 3: Sampled city locations from around the world. Each of the red circles denotes sampling region for the city located at its center.

3.2. Data-Driven Approach for Predicting Location

We use a deep convolutional neural network to model the geo-dependency of face appearance. Such networks are known for their ability to learn high-level feature hierarchies from input [30], and are widely used in solving computer vision related problems. Their demonstrated effectiveness in transferring knowledge to other problem domains [31] encouraged us to apply them in our work. In our approach, we fine-tune the CNN architecture introduced in [32], initialized using several publicly available models [33]. The ImageNet model was trained on the ILSVRC 2012 dataset [32] and the Hybrid model was trained on the Places database [34] and a subset of the ILSVRC 2012 dataset. Even though the ImageNet model is used for generic object detection and recognition, and the Hybrid model for both object detection and scene recognition, we show that the knowledge learned by the networks from objects other than human faces can also be used to learn the geo-dependency of human faces.

This architecture consists of five convolution layers, of which the first two, and the last one are followed by a max pooling layer for learning translation invariant features. The first two pooling layers are followed by a local response normalization layer to aid generalization. Rectified linear units (RELUs) are used as the non-linear activation function. The convolutional layers are followed by three fully-connected layers. The first two fully-connected layers employ dropout for learning more robust features and preventing overfitting. We update the final fully-connected layer to have number of outputs corresponding to the number of cities.

We use the dataset described in Section 4.1 and formulate the problem as a one-of-many classification task. To train the network, we minimize a multinomial logistic loss function using stochastic gradient descent (SGD). This results in a categorical probability distribution over locations. We set the learning rate for all the layers except for the final fully-connected layer low to encourage knowledge retention from

[1] http://www.omron.com/ecb/products/mobile/

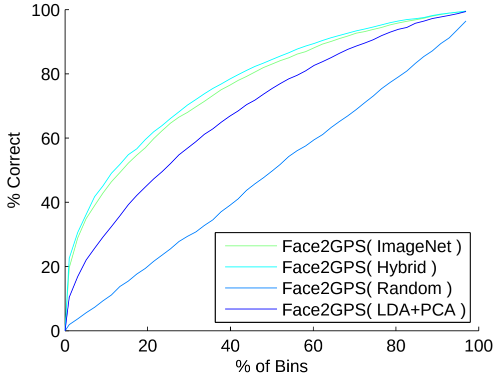


Fig. 4: Performance comparison of different classification methods.

the initializing model. It also helps the final layer learn the correct weights for classification taking advantage of the geo-dependent appearance features learned by the previous layers. We start with a base learning rate of 0.001 and increase it progressively. The learning rate for the SGD solver is set to 0.001 with a step size of 20000. We implement our methods using Caffe [33].

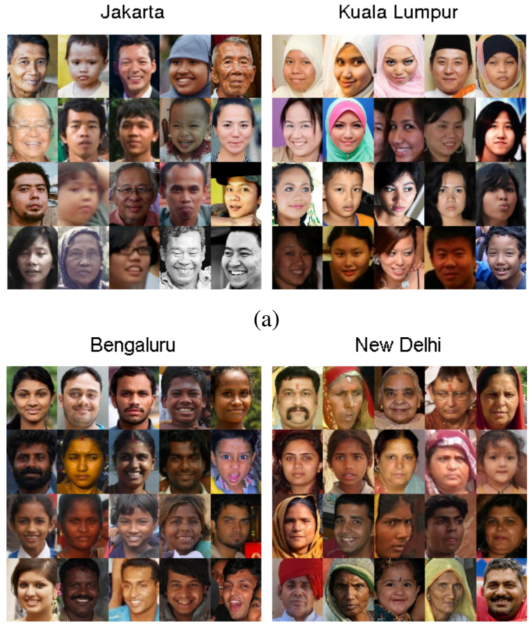
Once the training is complete, we apply the images from our test set at the input layer of the trained network one at a time and obtain a probability distribution over the target cities. For predicting the most likely location for a given test face, we find the city with maximum probability in this distribution. See Figure 2 for a few examples of inferences made using this approach. We can see from the figure that the predicted locations of the faces are reasonably accurate as they match the expected appearance of a person from the corresponding locations.

4. EVALUATION

We evaluate our proposed approach against two baselines, linear discriminant analysis (LDA) and random predictions, on a large real-world dataset.

4.1. Evaluation Dataset

We constructed an evaluation dataset using a subset of GeoFaces [1], a dataset built from 3 million geo-tagged Flickr images. For each image, all faces were detected and automatically filtered, to minimize the number of non-frontal faces, using the commercial tool previously mentioned. Each face was then aligned to a reference face with a similarity transform, using the eyes as control points. We then filtered out the faces that are not front facing using the technique described in [1]. The final filtered dataset consists of more than $800k$ geo-tagged near-frontal face patches. We used these images to construct a suitable test dataset. We first selected 50 large cities from around the world. We then computed non-overlapping, circular sampling regions around each and randomly sampled, with replacement, disjoint sets of training ($n = 3000$) and testing images ($n = 50$) for each city. The sampling regions for all cities are shown in Figure 3.



(a)

(b)

Fig. 5: Montages of people predicted to be from cities that are similar in ethnic distribution, but differs in overall appearance.

4.2. Quantitative Results: Prediction Accuracy

We compared the predictions from our approach described in Section 3.2 against two baseline methods. The first baseline uses the top 50 PCA coefficients estimated from the training set as input features [1] and applies LDA and the second baseline is random chance. The results show that our deep learning based approach for estimating geographic location from face images is superior to these baselines. The performance obtained initializing from the Hybrid model is comparable to that obtained using the ImageNet model, whereas using LDA on PCA coefficients performed worse than both of these. All of the methods significantly outperform random chance, indicating that our model did learn location dependency successfully. In Figure 4 we see the performance comparison curves. It shows that, for example, if we consider the top 10% predictions, the accuracy is almost 44% for Face2GPS using ImageNet, 47% using Hybrid, 30% using LDA, and 10% using random chance. That is, if we consider the five cities with the highest probabilities, Face2GPS using Hybrid model performs best; for 47% images from the test set, the ground truth city is one of these five cities.

Table 1: The top five most frequently predicted cities for faces from a given city.

Ground Truth	Predicted City (% of faces predicted to this city)				
Lima	Lima (48%)	New Delhi (10%)	Bengaluru (8%)	Jakarta (6%)	Istanbul (4%)
Nairobi	Nairobi (38%)	Lagos (24%)	Lima (6%)	Paris (4%)	New York (4%)
Lagos	Lagos (52%)	Nairobi (22%)	Paris (4%)	Bengaluru (4%)	Auckland (4%)
Hong Kong	Hong Kong (44%)	Kuala Lumpur (6%)	Shanghai (6%)	Beijing (4%)	Seoul (4%)
Kuala Lumpur	Kuala Lumpur (34%)	Jakarta (10%)	Singapore (6%)	Bangkok (6%)	New Delhi (4%)
Bogotá	Bogotá (34%)	Santiago (8%)	Nairobi (6%)	Caracas (6%)	Prague (4%)
Baghdad	Baghdad (34%)	Santiago (6%)	Jakarta (6%)	New Delhi (4%)	Kuala Lumpur (4%)

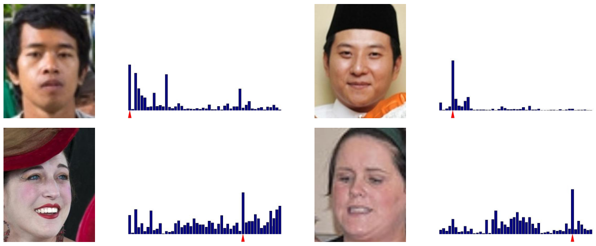


Fig. 6: Examples of faces from our dataset with the PDFs over location estimated by our method (in each histogram the bar with the red triangle under it is the ground-truth location). This demonstrates that individuals of different ethnicities (top vs. bottom) can have very different PDFs. It also shows that even individuals with more similar facial appearance (left vs. right) can have different PDFs. While it is difficult to say for sure, we suspect that these differences are largely due to the dramatic visual differences in hairstyles, headwear, and makeup and not the subtle differences in ethnicity

4.3. Error Analysis: Finding Similar Cities

We analyzed the predictions made by our proposed approach and found interesting patterns in the errors. While many can be attributed to non-geographic factors, such as strong illumination conditions or extreme facial expressions, some cannot. Many prediction errors can be attributed to similar looking people living in different cities. We speculated that such similarities are dominated by appearance similarities between the predominant ethnic groups. To investigate this, we show the most frequently predicted locations for a few of the selected cities in Table 1. As an example, if we look at the third row of this table, we see that geographic location of 52% of the faces from Lagos were correctly predicted, while 22% of them were predicted to be from Nairobi. This highlights that even though the accuracies for some cities are low, the errors are reasonable. For each row, the top cities are either geographically close, or have similar facial features.

5. DISCUSSION

The results in Section 4.3 highlight that ethnicity plays a significant role in the predictions made by our system. This motivates a natural follow-on question, "Is the proposed system just doing ethnicity detection?" To investigate this, we construct montages of the individuals that scored highest as having been photographed in a particular city. Specifically, we sort all the faces from a pair of cities and calculate their score as the difference of the probabilities of being from those two cities. We show the highest scoring 20 faces from each city for two pairs of cities in Figure 5. By comparing nearby cities with similar predominant ethnic groups, we see that there seem to be striking differences in the appearance of people in these cities that are beyond ethnic cues. For example, in Figure 5a, people in Kuala Lumpur are more likely to be wearing a specific type of headwear. On the other hand, in Figure 5b, more people from the Bengaluru montage can be seen smiling. While purely qualitative, this suggests that there is more to learn than just ethnicity. As additional evidence, consider the histograms shown in Figure 1 and Figure 6. These show that our classifier was able to identify differences between people with the same ethnicity as well as people with different ethnicities.

6. CONCLUSION

We demonstrated that human facial appearance is strongly related to the location an image was captured. To our knowledge, this is the first work that uses a deep convolutional neural network for directly predicting image geo-localization. We compared the proposed technique with two baseline methods and found that it significantly outperforms both. We also analyzed the errors made by our system and found interesting patterns. This highlights that there is rich structure in the relationship between facial appearance and geographic location that is worthy of further study. While we only consider a single face in this work, it is likely that considering multiple faces per image will increase localization accuracy.

Acknowledgments

This work was partially supported by the NSF (CNS-1156822) and DARPA (D11AP00255). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor. See `http://cs.uky.edu/~tarik/papers/face2gps/` for image attributions.

7. REFERENCES

[1] M.T. Islam, S. Workman, Hui Wu, N. Jacobs, and R. Souvenir, "Exploring the geo-dependence of human face appearance," in *WACV*, 2014.

[2] Connor Greenwell, Scott Spurlock, Richard Souvenir, and Nathan Jacobs, "GeoFaceExplorer: Exploring the geo-dependence of facial attributes," in *ACM SIGSPIATAL Workshop (GEOCROWD)*, 2014.

[3] J. Hays and A.A. Efros, "Im2gps: estimating geographic information from a single image," in *CVPR*, 2008.

[4] A.R. Zamir and M. Shah, "Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1546–1558, Aug 2014.

[5] D. Haase, E. Rodner, and J. Denzler, "Instance-weighted transfer learning of active appearance models," in *CVPR*, 2014.

[6] S. Fu, H. He, and Z. Hou, "Learning race from face: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2483–2509, Dec 2014.

[7] H. H. K. Tin and M. M. Sein, "Race identification for face images," *ACEEE International Journal on Information Technology*, vol. 01, pp. 327–344, 2011.

[8] A. J. Calder, G. Rhodes, M. Johnson, and J. V. Haxby, *Oxford Handbook of Face Perception*, Oxford Univ. Press, London, U.K., 2011.

[9] Usman Tariq, Yuxiao Hu, and ThomasS. Huang, "Gender and race identification by man and machine," in *Pattern Recognition, Machine Intelligence and Biometrics*, PatrickS.P. Wang, Ed., pp. 313–333. Springer Berlin Heidelberg, 2011.

[10] Xiaoguang Lu, , Xiaoguang Lu, and Anil K. Jain, "Ethnicity identification from face images," in *SPIE International Symposium on Defense and Security : Biometric Technology for Human Identification*, 2004.

[11] Bor-Chun Chen, Chu-Song Chen, and WinstonH. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *ECCV*, 2014.

[12] Tao Wu and Rama Chellappa, "Age invariant face verification with relative craniofacial growth model," in *ECCV*, 2012.

[13] Yael Eisenthal, Gideon Dror, and Eytan Ruppin, "Facial attractiveness: Beauty and the machine," *Neural Computation*, vol. 18, no. 1, pp. 119–142, 2006.

[14] Amit Kagian, Gideon Dror, Tommer Leyvand, Daniel Cohen-Or, and Eytan Ruppin, "A humanlike predictor of facial attractiveness," in *NIPS*, 2006.

[15] Jacob Whitehill and Javier R Movellan, "Personalized facial attractiveness prediction," in *FG*, 2008.

[16] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning identity-preserving face space," in *ICCV*, 2013.

[17] Y. Taigman, Ming Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[18] Junlin Hu, Jiwen Lu, and Yap-Peng Tan, "Discriminative deep metric learning for face verification in the wild," in *CVPR*, 2014.

[19] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong, "Facial expression recognition via a boosted deep belief network," in *CVPR*, 2014.

[20] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless, "Geolocating static cameras," in *ICCV*, 2007.

[21] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros, "What makes paris look like paris?," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 101, 2012.

[22] N. Jacobs, N. Roman, and R. Pless, "Toward fully automatic geo-location and geo-orientation of static outdoor cameras," in *WACV*, 2008.

[23] Tsung-Yi Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *CVPR*, 2013.

[24] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys, "Large scale visual geo-localization of images in mountainous terrain," in *ECCV*, 2012.

[25] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua, "Worldwide pose estimation using 3d point clouds," in *ECCV*, 2012, pp. 15–29.

[26] Lin Wu and Xiaochun Cao, "Geo-location estimation from two shadow trajectories," in *CVPR*, 2010.

[27] N. Jacobs, K. Miskell, and R. Pless, "Webcam geo-localization using aggregate light levels," in *WACV*, 2011.

[28] Jean-François Lalonde, Srinivasa G Narasimhan, and Alexei A Efros, "What do the sun and the sky tell us about the camera?," *International Journal of Computer Vision*, vol. 88, no. 1, pp. 24–51, 2010.

[29] Scott Workman, R. Paul Mihail, and Nathan Jacobs, "A Pot of Gold: Rainbows as a Calibration Cue," in *ECCV*, 2014.

[30] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (DeepVision)*, June 2014.

[31] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, pp. 3320–3328. Curran Associates, Inc., 2014.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, 2015.

[33] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[34] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database.," *NIPS*, 2014.