Who Goes There? Approaches to Mapping Facial Appearance Diversity

Zachary Bessinger
Dept. of Computer Science
University of Kentucky
zach@cs.uky.edu

Chris Stauffer
Visionary Systems and Research
stauffer@vsandr.com

Nathan Jacobs
Dept. of Computer Science
University of Kentucky
jacobs@cs.uky.edu

ABSTRACT

Geotagged imagery, from satellite, aerial, and ground-level cameras, provides a rich record of how the appearance of scenes and objects differ across the globe. Modern web-based mapping software makes it easy to see how different places around the world look, both from satellite and ground-level views. Unfortunately, interfaces for exploring how the appearance of objects depend on geographic location are quite limited. In this work, we focus on a particularly common object, the human face, and propose learning generative models that relate facial appearance and geographic location. We train these models using a novel dataset of geotagged face imagery we constructed for this task. We present qualitative and quantitative results that demonstrate that these models capture meaningful trends in appearance. We also describe a framework for constructing a web-based visualization that captures the geospatial distribution of human facial appearance.

CCS Concepts

•Information systems → Geographic information systems; •Computing methodologies → Appearance and texture representations;

Keywords

Geotagged imagery, facial appearance modeling, web-based mapping

1. INTRODUCTION

Understanding cultural and demographic trends and their spatial distribution is increasingly essential for individuals, corporations, and governments. Social scientists attempt to discover such trends, but the vast scale of this problem means that traditional approaches, which often involve manual data collection and scholarly dissemination, are insufficient. With the advent of social media, it has become quite easy to automatically collect sufficient data that reflects these trends. However, novel methods for interpreting this data are still needed.

Figure 1: We use geotagged social media images to learn how human facial appearance varies globally. This montage shows representative images for different clusters of people.

We develop location-dependent human appearance models and visualizations, based on geotagged social media images, that enable novice users to understand world populations. We propose approaches for modeling the distribution, $P(f|\ell)$, of facial appearance, f, for arbitrary geographic locations, ℓ. Two main types of approaches we could consider for learning the relationship between human appearance and geographic location are discriminative and generative. Islam et al. [1] use a discriminative approach by addressing the face localization problem using a deep convolutional neural network (CNN) to estimate the city in which a given facial image was captured. This approach is appealing because it lends itself toward straightforward quantitative evaluation, however it does not directly support our goals of enabling user-focused visualizations. Therefore we take a generative approach; we construct models that allows us to estimate an individual's appearance for a given geographic location. We further extend these models by conditioning on other attributes, such as gender, age, and face shape. The resulting models are used to support visualizations, web-based applications, and analytics.

The main contributions of this work are: 1) A massive new dataset of geotagged face images, 2) a regression model that uses location, and potentially other attributes, to predict the facial appearance distribution for any location in the world, and 3) mapping applications that allow novice and expert users to explore our dataset and the learned models. We hope this work will serve as a foundation for a more ambitious platform for understanding human appearance.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-4589-7/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2996913.2996997

2. RELATED WORK

Geotagged Image Analysis.

Two common tasks in the area of geotagged image analysis are geolocalization and image-driven mapping. Geolocalization takes an input image and estimates where it was captured. Hays et al. [2] collected six million geotagged Flickr images and use a nonparametric approach to predict the location of a query image. Works since then have proposed various approaches using multiple sources of geotagged imagery, including aerial, ground-level, and landcover [3, 4].

Image-driven mapping is the process of recognizing attributes from imagery and then conveying the learned attributes in the form of a map. Crandall et al. [5] explore visual and textual characteristics of 35 million geotagged images from Flickr. Other works in this area include weather [6] and scenicness [7] mapping. Most work has focused on scene-level appearance attributes, while relatively little has explored object-level appearance attributes. Our work is an important step in this direction.

Soft Biometrics.

Understanding human faces is a long-standing problem in computer vision and a wide body of research has been done, including detection, recognition, and pose estimation. Many issues can arise when observing facial images, including unsuitable lighting conditions and challenging camera angles. Recent advances in CNNs have led to learned feature representations that are both compact and maintain semantics of identity, while being invariant to lighting and pose [8, 9]. Parkhi et al. [10] propose VGG Face, similar to [9], however they train on a large dataset of celebrity faces and use the VGG architecture. We utilize VGG Face in our work because the authors have publicly released their models.

Geo-Facial Image Analysis.

Geo-facial image analysis is similar to understanding soft biometrics, with the primary difference being that it incorporates location as an observed variable. The GeoFaces dataset [11] is the largest face dataset explicitly designed to explore the relationship between location and facial appearance. Greenwell et al. [12] develop a pipeline to process geotagged imagery from Flickr and map several detected attributes. Our work differs in that we create maps of facial appearance, not the distribution of facial attributes such as age and gender. Islam et al. [13] provides a broad overview of problems in geo-facial image analysis. Our work is similar, however we introduce a larger dataset, use improved low-level processing methods, and focus on learning generative models of human facial appearance. In contrast to their use of appearance based features (e.g., PCA applied to raw pixel intensities) for mapping, we use semantic features related to identity.

3. THE WGT DATASET

We have curated a large dataset of geotagged face images, referred to as the WhoGoesThere? (WGT) dataset. Our data source is the Yahoo Flickr Creative Commons 100M (YFCC100M) [14], which contains 100 million images, of which 49 million are geotagged, and their associated metadata. The metadata consists of 34 attributes including the date uploaded, user and machine tags.

3.1 Face Detection and Filtering

For each geotagged image, we detect a bounding box for the face, extract the face patch, and align it using the detected landmark points. Facial bounding boxes and their landmark points are extracted using the method of Kazemi et al. [15] which is known to have an extremely low false positive rate. This process resulted in $2,106,468$ geotagged face patches, each containing 68 fiducial landmarks.

Once the fiducial landmarks have been detected for a face, we extract four different face patches. We extract both wide (256×256) and tight (153×153) crops of each face. We then align each crop using a similarity transformation based on reference eye centers and a perspective transformation, conditioned on the detected gender to male and female reference models. Gender-specific alignment is done to capture differences in male and female facial structure.

3.2 Feature Extraction

We extract three features for each of the face patches: PCA (appearance), VGG Face [10] FC8 (identity), and several additional attributes. We randomly sample $200,000$ faces from our dataset and learn a PCA basis using the similarity aligned, tight cropped patches as input. The remaining images are reserved for experiments and evaluation. In our experiments, we find that $200,000$ is a sufficient number of training images to learn a basis that captures facial appearance.

Identity features are extracted in a similar manner. We use the VGG Face network and extract features from the network's FC8 layer which correspond to the semantic labels. We find that these identity features are more invariant to lighting and pose than the PCA appearance features.

To support our interest in a data-driven approach to learning demographics at a worldwide scale, we provide age and gender estimates using the CNNs of Levi et al. [16] and reverse geocodings. The WGT dataset, including the extracted face patches, appearance and identity image features, detected fiducial points, and age/gender estimates are available at: `http://wgt.csr.uky.edu`.

4. PREDICTING FACIAL APPEARANCE

The appearance of a face is dependent on many factors, including the individual's age, gender, face shape, and pose. In addition to these proximate factors, appearance is also dependent, albeit indirectly, on the geographic location where the image was captured. We propose using regression to predict facial appearance from varying combinations of these factors to better understand the relationship between facial appearance and location. In this section, we focus on unimodal regression methods and minimize the L_2 loss function for all models. In Section 5 we consider multimodal models.

We propose models of how age, gender, facial shape, and location affect the expected value of appearance given these factors. Using an 80/20 training and testing dataset split, we learn two models: a linear regression model and a random forest (RF) model whose objective is predicting the top 2048 PCA coefficients for different subsets of predictor variables. Since each model minimizes the same L_2 loss function, we find that the RMSE for all models is close in the range of [10.245, 10.328]. The location RF model has the lowest RMSE and the age linear model has the highest.

Figure 2 shows faces generated from our models. The first column shows a source patch we extracted features from and

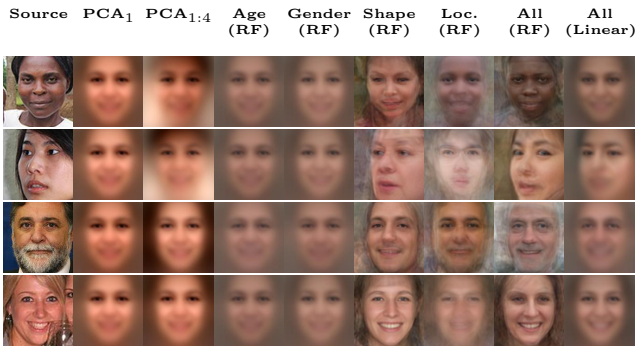

Figure 2: Results of learning models conditioned individually on age, gender, location, and face shape and then conditioned on all four of these attributes. The predicted components are then used to reconstruct the original image.

the remaining columns show reconstructions from our models. In the ideal case, each reconstruction would produce the source patch. The second and third columns show reconstructions from the top one and four PCA components, highlighting illumination at various angles. The following five columns show reconstructions for each RF model conditioned on a single objective attribute and then all attributes. The last column compares our RF model to our linear model. These reconstructions show that shape is more informative than age and gender alone, but by conditioning on all attributes our RF model is able to reconstruct a plausible face relative to the source patch.

5. MULTI-MODAL DISTRIBUTIONS

In locations with diverse populations, a conditional average face, or any individual exemplar image, is likely insufficient to accurately reflect the diversity of facial appearance. To overcome this, we propose learning a conditional multi-modal distribution. The key idea is to cluster faces in image feature space, assign each face to a cluster, and then learn to predict cluster membership of a given geographic location. Once this model is trained, we input location and obtain a distribution over the expected face types. We are essentially fitting a mixture model, $P(f|\ell) = \sum P(f|c)P(c|\ell)$, where f is a facial feature, c is a cluster, and ℓ is a location.

5.1 Clustering Faces

Our goal is to cluster faces into groups of similar facial appearance. At one extreme, we could group all faces into one cluster, which is essentially the approach used in the previous section. This approach does not allow us to model the multi-modal nature of human appearance. At the other extreme, we could attempt to make each cluster only contain images from a single individual. This approach would make learning a conditional distribution difficult because there would likely be few samples per label. Experimentally, we found that clustering into $k = 250$ groups was a good compromise. Initially, we also found that using appearance features (PCA) resulted in clusters mostly grouped faces by pose and lighting conditions. We discovered that clustering the identity features (VGG Face FC8) resulted in more semantically meaningful clusters.

The identity features for each face patch are clustered using iterative k-means clustering on a subset of our dataset.

Figure 3: (left) An exemplar face for a given class, c_i. (right) The conditional distribution of that class for each location, $P(c = c_i|location)$.

We use a stratified sampling approach to minimize the impact of dataset bias (few images captured in Africa). We discretize the world into a 10×10 grid and randomly sample no more than 500 faces from each bin to form our stratified training set.

The final step of the clustering process is to construct an exemplar face. We select the $5,000$ faces nearest to their cluster center for each cluster. We then compute the average landmarks for the $5,000$ closest faces and preserve top 800 faces whose landmarks are nearest to the average landmarks to ensure the face is mostly front-facing. We then take these 800 faces and apply Collection Flow [17] to structurally refine the exemplar face. The resulting image is assigned as the exemplar image for each individual cluster. A subset of exemplar faces are shown in Figure 3.

5.2 Conditional Distribution of Clusters

With our faces now clustered in feature space, we use a neural network to represent the distribution over cluster assignments. Our network takes geographic location and land-cover class as input and outputs the conditional probability of the cluster assignment. We found that including the land-cover class improved the accuracy of our model significantly and made training converge more reliably. The network is feed-forward with three hidden layers, with 100, 100, and 50 nodes respectively. All activations are hyperbolic tangent and L_2 regularization ($\lambda = 1e^{-5}$) is used. We train the network using stochastic gradient descent (batch size $= 10,000$) with a cross-entropy loss function.

Our neural network outputs $P(c|\ell)$, however by using Bayes rule we can visualize $P(\ell|c = c_i)$. This distribution reflects where you would be most likely to find a face belonging to the given cluster center, c_i, as shown in Figure 3. Specifically, for each map we sample from $P(\ell|c = c_i)$ for a particular cluster, c_i, at a dense grid of geographic locations. The darker the location on the map the more likely it is that a face seen at that location will be from the cluster. These maps highlight that our model has learned how different ethnic groups are distributed around the world.

6. EVALUATION

6.1 Quantifying Appearance Diversity

In this section, we design a metric useful for quantifying appearance diversity. Using the FC8 image identity features, we quantitatively measure diversity of a population by their fraction of variability. We begin by querying a set of high-population countries scattered throughout the world. Since Africa has a relatively sparse number of images, we select several countries from Central Africa to compare diversity. For each country, we compute the covariance of the identity

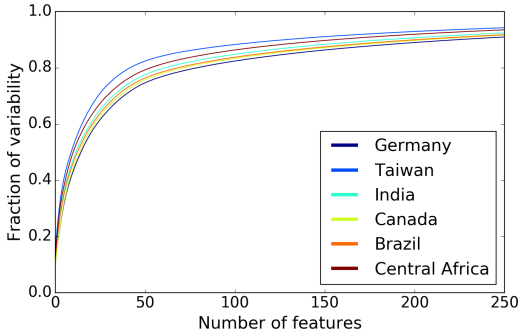

Figure 4: Quantifying appearance diversity using the fraction of variability explained by the top k PCA components of the FC8 identity features. For a given number of components, larger values imply less diversity, because more of the variability is explained by the top k components.

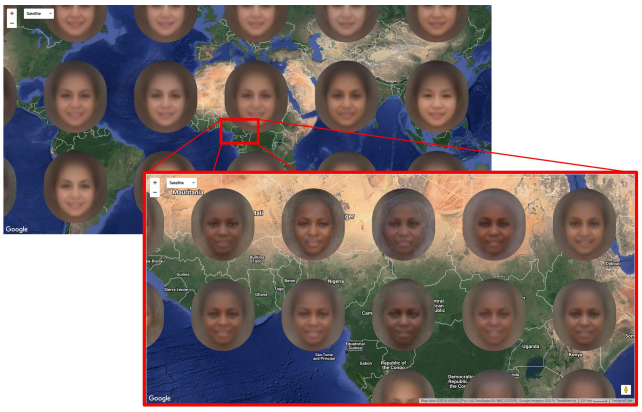

Figure 5: Multiscale visualization. Zooming in reveals finer details about world populations.

features and apply SVD to the covariance matrix. The fraction of variability is defined as: $\lambda_n = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{N} \lambda_i}$, where λ are the set of eigenvalues calculated by SVD, n is the number of top eigenvalues considered significant, and $N = |\lambda|$. This metric allows us to examine multivariate variability. In our case, this implies that the more appearance diverse a region is, the lower the fraction of variability will be. Conversely, less diverse regions will have a higher fraction of variability. Figure 4 shows the fraction of variability for the selected countries. If we compare Taiwan with Germany using the top 50 eigenvalues, their fractions of variability are 0.744 and 0.822, respectively. These values tell us that Germany is 9.49% more appearance diverse relative to Taiwan when considering 50 dimensions of diversity.

6.2 Interactive Visualizations

In this section, we describe a method for multiscale visualization of human appearance. We begin by discretizing the world into a set of spatial bins. To scale with the addition of new images, we apply an on-line method by maintaining the sufficient statistics of a Gaussian distribution for each bin. We maintain the count of images in each bin, c, and the running sum of the images, $I_{sum} = \sum_k^c I_k$. These two values allow us to generate the mean and covariance in an efficient manner for any given bounding box, enabling us to visualize the average facial distribution of any queried region.

Figure 5 shows a screenshot from the multiscale web application. The top image is at a higher zoom level and the bottom image is a lower zoom level showing finer-grained appearance. The user is also able to toggle the age and gender representation. We are actively developing new visualizations as our research in this area progresses.

7. CONCLUSIONS

Overall, we have curated a large-scale dataset of geotagged faces and shown many ways that we can both qualitatively and quantitatively model worldwide appearance and diversity. We have demonstrated a variety of visualizations and shown several pragmatic applications of our work. Our dataset and web-based visualizations are publicly released in hopes that this work will serve as a multidisciplinary foundation towards furthering our understanding of human appearance diversity.

8. REFERENCES

[1] Islam, M.T., Workman, S., Jacobs, N.: Face2gps: Estimating geographic location from facial features. In: ICIP. (2015)
[2] Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: CVPR. (2008)
[3] Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: CVPR. (2013)
[4] Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: ICCV. (2015)
[5] Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: WWW. (2009)
[6] Jacobs, N., Burgin, W., Fridrich, N., Abrams, A., Miskell, K., Braswell, B.H., Richardson, A.D., Pless, R.: The Global Network of Outdoor Webcams: Properties and Applications. In: ACM GIS. (2009)
[7] Xie, L., Newsam, S.: Im2map: deriving maps from georeferenced community contributed photo collections. In: Proceedings of the 3rd ACM SIGMM international workshop on Social media. (2011)
[8] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR. (2014)
[9] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. (2015)
[10] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. (2015)
[11] Islam, M.T., Workman, S., Wu, H., Souvenir, R., Jacobs, N.: Exploring the Geo-Dependence of Human Face Appearance. In: WACV. (2014)
[12] Greenwell, C., Spurlock, S., Souvenir, R., Jacobs, N.: GeoFaceExplorer: Exploring the Geo-Dependence of Facial Attributes. In: ACM SIGSPATAL GEOCROWD. (2014)
[13] Islam, M.T., Greenwell, C., Souvenir, R., Jacobs, N.: Large-Scale Geo-Facial Image Analysis. EURASIP Journal on Image and Video Processing (JIVP) **2015**(1) (2015)
[14] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: The new data and new challenges in multimedia research. arXiv preprint arXiv:1503.01817 (2015)
[15] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR. (2014)
[16] Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: CVPR Workshop. (2015)
[17] Kemelmacher-Shlizerman, I., Seitz, S.M.: Collection flow. In: CVPR. (2012)