Analyzing Human Appearance as a Cue for Dating Images

Tawfiq Salem, Scott Workman, Menghua Zhai, Nathan Jacobs
Department of Computer Science, University of Kentucky
{salem, scott, ted, jacobs}@cs.uky.edu

Abstract

Given an image, we propose to use the appearance of people in the scene to estimate when the picture was taken. There are a wide variety of cues that can be used to address this problem. Most previous work has focused on low-level image features, such as color and vignetting. Recent work on image dating has used more semantic cues, such as the appearance of automobiles and buildings. We extend this line of research by focusing on human appearance. Our approach, based on a deep convolutional neural network, allows us to more deeply explore the relationship between human appearance and time. We find that clothing, hair styles, and glasses can all be informative features. To support our analysis, we have collected a new dataset containing images of people from many high school yearbooks, covering the years 1912–2014. While not a complete solution to the problem of image dating, our results show that human appearance is strongly related to time and that semantic information can be a useful cue.

1. Introduction

The time when an image was captured has a dramatic, albeit usually indirect, impact on the appearance of the image. Time impacts a wide variety of more immediate causes, such as: the lighting conditions, the weather, the season, the age of individuals, biases in photographic viewpoints (e.g., the recently introduced "selfie stick"), trends in camera technology (e.g., the decline in the use of film and the rise in popularity of fisheye lens cameras), and trends in the appearance of objects. This last element has received growing interest lately, with exciting work exploring the time-dependence of architectural styles [13] and automobile appearance [14]. We propose to continue this research direction by investigating the time-dependence of human appearance, including facial appearance and clothing styles.

While our goal is to explore the relationship between human appearance and time, we take a discriminative approach. Specifically, we propose algorithms which use human appearance to estimate the date when an image was captured (Figure 1). The task of automatically assigning a date to an image has received significant attention recently [5, 13, 14, 17]. Coming full circle, we use the models we train discriminatively to investigate the relationship between human appearance and when an image was captured.

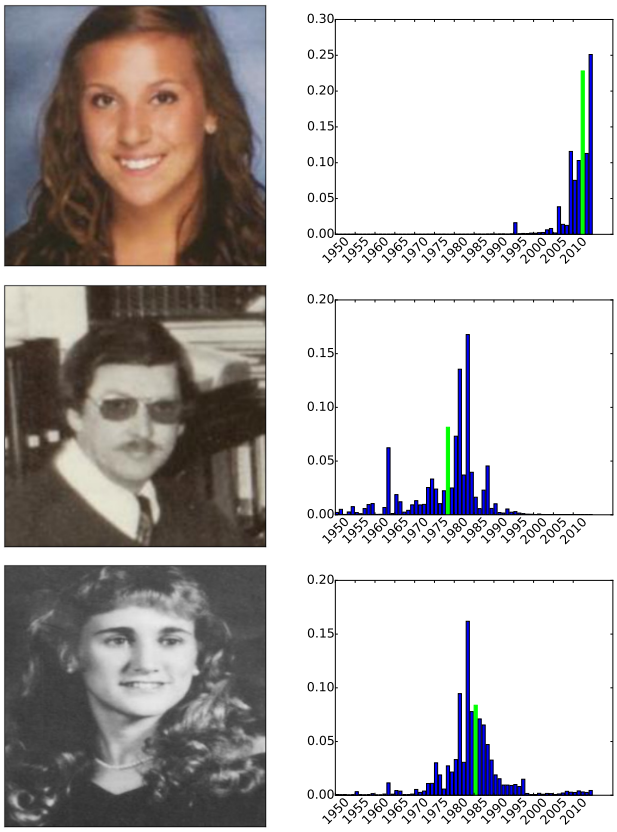
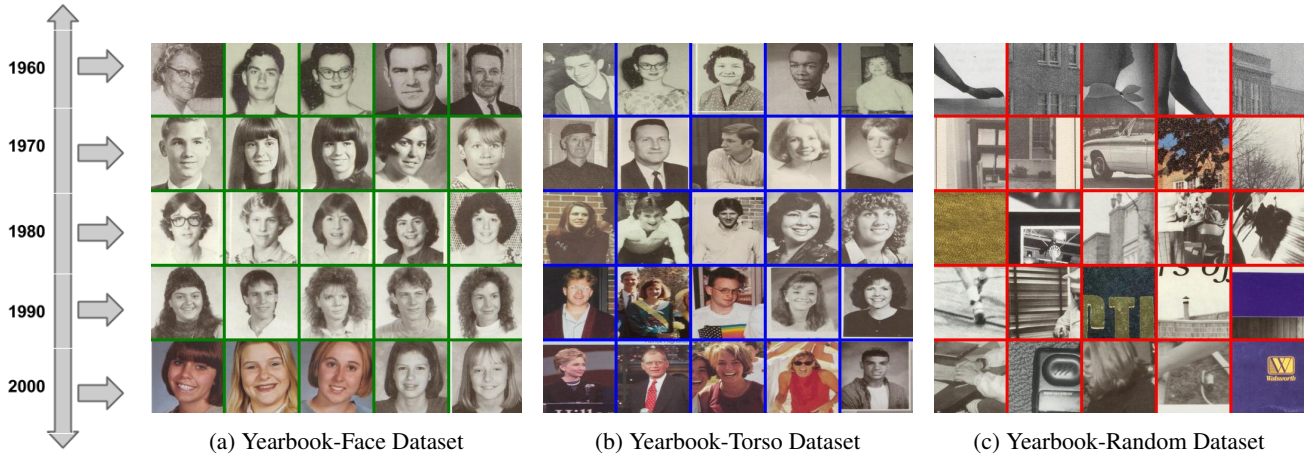


Figure 1: Our work explores the relationship between human appearance and time. Given a photo containing a human, we train discriminative models to estimate the year the image was captured. Using these models, we investigate which aspects of human appearance are most dependent on time.

To support this effort, we constructed a large dataset con-



(a) Yearbook-Face Dataset (b) Yearbook-Torso Dataset (c) Yearbook-Random Dataset

Figure 2: Sample images from our Yearbook datasets.

taining images of people over a 100 year span by processing a collection of digitized high school yearbooks. The result is a dataset of approximately 600 000 images of people in the time period ranging from 1912 to 2014. We use this dataset to both train our model and to evaluate its performance.

Main Contributions The main contributions of this work are:

- introducing a dataset containing timestamped images of people captured over a 100 year span,
- proposing an approach, based on deep convolutional neural networks, to estimate when an image was captured directly from raw pixel intensities, and
- providing a detailed evaluation, both quantitative and qualitative, of the learned models for a variety of different settings.

2. Related Work

We provide an overview of work in three related areas: studying human appearance, spatial and temporal modeling, and learning with convolutional neural networks.

Studying Human Appearance The study of human appearance is fundamental in the field of computer vision. Traditional tasks in this domain include face recognition and verification [21], estimating pose [7], predicting age and gender [15, 26], and interpreting fashion [19, 22]. Islam et al. [9] found that ethnicity and other appearance attributes, such as facial expressions and hair styles, are related to geographic location. Inspired by this study, our work focuses on the relationship between human appearance and time. Ginosar et al. [6], which was developed concurrently and independently from our work, considers the same relationship. Like our work, they use yearbook imagery. However, they use weakly-supervised data-driven techniques to analyze appearance trends. In our work we focus on the task of dating imagery using human appearance and show that our discriminative models capture semantics, such as clothing and hair styles, that are typical of different eras.

Spatial and Temporal Modeling A significant amount of work has tried to characterize the relationship between the appearance of objects and how it changes with respect to location and time. For example, analyzing fashion trends [19], characterizing city identity [2, 29], estimating geo-informative features [3, 9, 24], and automatic image geolocalization [8, 25]. Recently dating historical imagery has received a lot of attention [5, 13, 14, 17]. Palermo et al. [17] introduce a method for dating historical color images using hand-designed color features. Lee et al. [13] find visual patterns in the architecture of buildings, relate them to certain time periods, and show how they can be used to date buildings. We develop methods for dating imagery which take advantage of human appearance.

Learning with Convolutional Neural Networks Convolutional neural networks have recently become the most popular machine learning algorithm in computer vision due to their ability to learn custom feature hierarchies for a given task. Such networks are designed to take advantage of a two dimensional input, employing a series of convolutional layers for extracting features at different spatial locations. They have achieved state-of-the-art results for many vision tasks, including object recognition [20], scene classification [28], and 3D image understanding [23]. We build on this success and explore their application to dating imagery.

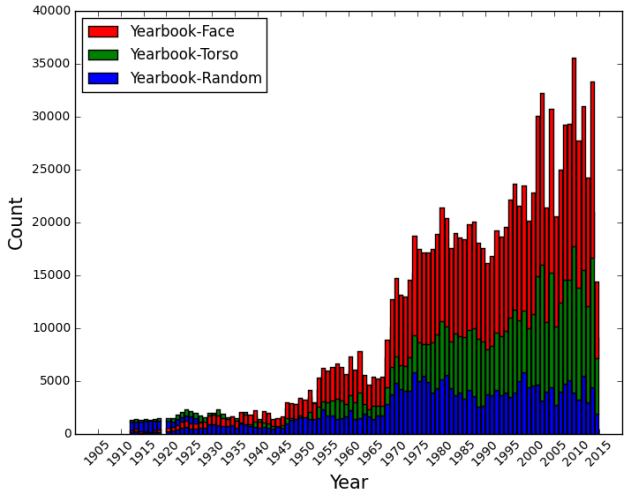


Figure 3: Visualizing the distribution over time of images in our Yearbook datasets.

3. A Dataset of Timestamped People Images

We constructed a large dataset of timestamped images of people from digitized school yearbooks, which were made available by the Daniel Boone Regional Library.[1] The yearbooks contain images of the students, faculties, and staff members from eight high schools in Boone and Callaway County, Missouri for many years, from 1912 to 2014. Each yearbook was digitally scanned page-by-page and uploaded to a photo sharing website. In total, the collection contains 372 yearbooks, each clearly labeled with the year it was created, and 62 939 digitized pages. We detected people in each yearbook page and extracted three different types of patches (see below for details) to build the three datasets of timestamped images we use for evaluation. This section describes the construction process for these datasets.

3.1. Yearbook-Face Dataset

To construct the Yearbook-Face dataset, we detected faces [1, 4] using the Dlib C++ Library [11] and extracted the corresponding patches (dilating the detection bounding box by 25%). This resulted in 571 686 face patches, each with a known year. We show example images from the dataset in Figure 2a. In Figure 3, we visualize the distribution of the timestamped face patches with respect to year. More images are found in recent years; we conjecture that this is due to two factors: rising student populations and missing yearbooks from early years.

[1] http://www.dbrl.org/reference/community-school-yearbook-archive

3.2. Yearbook-Torso Dataset

Starting with the detections from the previous section, we created a dataset of torso patches, which we refer to as Yearbook-Torso. To construct this, we dilated the detection bounding boxes by 35% to the left and right, 25% in the up-direction, and 80% in the down-direction. We discarded dilated patches that fall outside the image boundaries. This process resulted in 565 069 timestamped torso images. We show several sample torso images in Figure 2b and the distribution of images with respect to year in Figure 3.

3.3. Yearbook-Random Dataset

We also constructed a dataset, Yearbook-Random, that contains random patches sampled from the yearbook pages. To construct this dataset, we filtered out pages with more than five faces and randomly sampled ten patches, of size 250×250, from each. This resulted in 264 840 timestamped patches. Figure 2c shows example patches and Figure 3 shows their distribution over time. This dataset will serve as a baseline for comparison, to highlight the extent to which we are able to learn in a way that is related to human appearance, not just the appearance of the yearbook page.

4. From Human Appearance to Year

We take a discriminative approach to learn the relationship between human appearance and the year. Specifically, we learn to predict the year an image was captured using a deep convolutional neural network (CNN). We begin with a brief overview of the use of CNNs for similar prediction tasks and then describe our methods in detail.

4.1. Background

Our approach uses the CNN architecture proposed by Alex Krizhevsky et al. [12] as a foundation. This architecture has eight layers with trainable parameters: five convolutional layers followed by three fully connected layers, each connected in a feed-forward manner. Rectified linear units (ReLU) are used as the non-linear activation function between layers. Max pooling and local response normalization layers are interspersed amongst the convolutional layers. The first two fully connected layers use dropout, a strategy used to prevent overfitting.

This CNN architecture was originally developed for object recognition and trained on 1.2 million images from the ImageNet ILSVRC-2012 challenge [18]. The final fully connected layer has 1 000 output dimensions corresponding to the 1 000 possible object classes. During training, a softmax loss function (softmax function followed by a multinomial logistic loss) is used to optimize the network parameters with stochastic gradient descent.

To adapt this architecture to a new classification task, the only change necessary is to modify the final fully connected

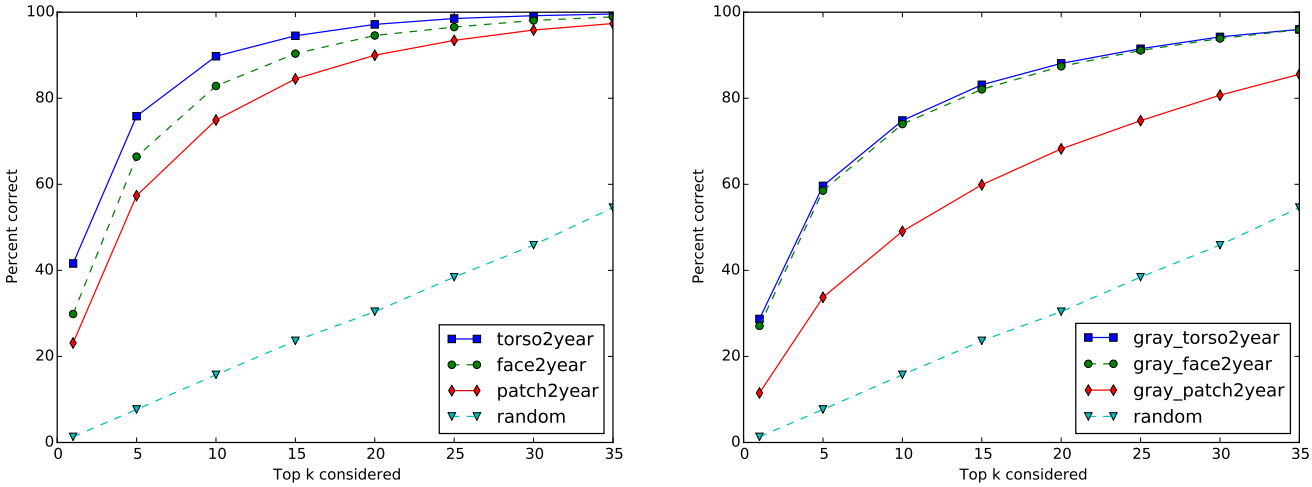


Figure 4: Accuracy of the networks over the colored (left) and grayscale (right) images.

layer to have the correct number of output classes for the new task. Then the network's weights can be *fine-tuned* on new training data by randomly initializing the weights of the modified final layer and using the weights from an existing model to initialize all other layers.

Following the approach outlined above, we train three different color-based networks *face2year*, *torso2year*, and *patch2year*, one for each dataset defined in Section 3. We generate a training and testing split using the following strategy. Starting from the full set of images, we filter images outside the time period of 1950 to 2014, and randomly pick 1 400 images from every year, splitting each into 80% for training, 10% for validation, and 10% for testing.

To modify the architecture described in the previous section, we update the final fully connected layer to have 65 outputs corresponding to the 65 years in the period of 1950 to 2014. The network weights are initialized and fine-tuned from the weights of a network originally trained for object classification [18].

We follow this strategy and create a set of three additional grayscale-based networks, *gray_face2year*, *gray_torso2year*, and *gray_patch2year* by replacing the color input image with a grayscale input image during training and testing.

4.2. Implementation Details

Our networks are implemented using the Caffe [10] deep learning framework. We use the CaffeNet reference network architecture, a variant of AlexNet, and initialize using pre-trained networks from the Caffe Model Zoo.[2] Our networks were trained on an NVIDIA Tesla K40 GPU for 24 hours each. The full network definition, network weights, and the output from our methods will be made available online for all networks http://cs.uky.edu/~salem/face2year/.

[2] http://caffe.berkeleyvision.org/model_zoo.html

Table 1: The accuracy of different networks.

Network	Top 1	Top 5	Top 10
face2year	29.9%	66.4%	82.8%
torso2year	**41.6%**	**75.8%**	**89.8%**
patch2year	23.1%	57.4%	74.9%
gray_face2year	27.1%	58.5%	74.0%
gray_torso2year	**28.7%**	**59.6%**	**74.8%**
gray_patch2year	11.5%	33.7%	49.1%

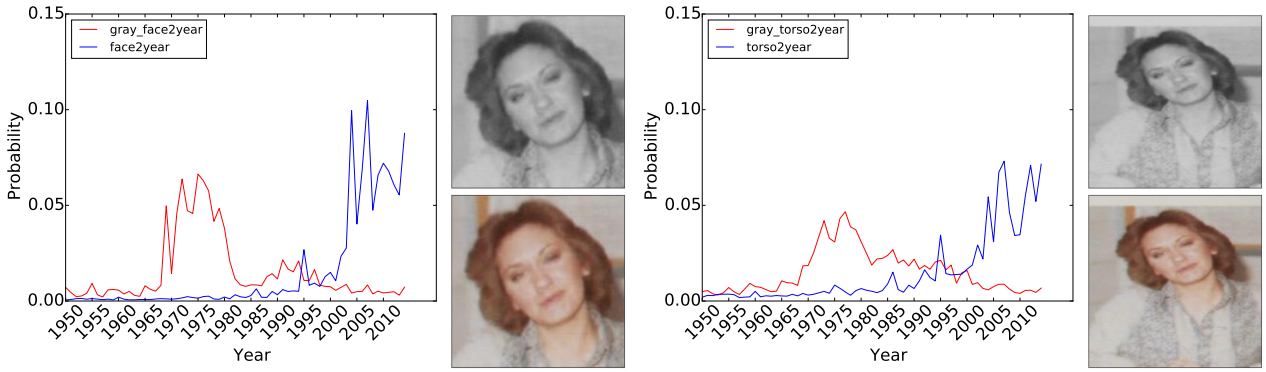
5. Evaluation

We evaluated the quantitative and qualitative properties of our six networks. We found that the color-based networks achieve higher accuracy than the grayscale-based networks. However, the grayscale-based networks appear to do a better job of capturing semantics of human appearance.

5.1. Quantitative Evaluation

Using the testing splits defined above, we evaluated the accuracy of the predictions made by our various networks. The "Top 1" column of Table 1 shows the percentage of correct predictions for each network. The "Top 5" column shows the percentage for which the correct answer was one of the five most likely years, the "Top 10" column is defined similarly. Figure 4 shows how the accuracy changes as the threshold, k, for "Top k" is varied.

Our results show that *torso2year* performs the best, followed by *face2year* then *patch2year*. The accuracy using the color images (left) is substantially higher than



(a) (b)

Figure 5: Analysis of the impact of color. (a) Without color, the network focuses on semantic properties, resulting in a prediction closer to the ground truth of 1980. (b) The same experiment repeated for the corresponding torso image.

for grayscale images (right). In addition, these results show that the patch-based prediction for the color images (*patch2year*) is much higher, relatively, than for the grayscale images (*gray-patch2year*). We conjecture that there are strong cues in the colors of the images, due to differences in the photographic process for capturing and printing the older yearbooks, that the color-based networks are exploiting. However, for the grayscale-based networks, *gray-patch2year* gives significantly lower accuracy.

Figure 5 shows the predictions of various networks for a color image from a 1980 yearbook. Reasonable predictions are given by the grayscale-based networks, but the color-based networks are very confident that the image was from the 1990s or 2000s. We suspect the color-based networks make a poor prediction because color images are not commonly found in this collection until the mid-1990s. Therefore, this unusually early color image is predicted to be from much later than it truly was. Given these results, and our interest in semantics of appearance, we focus on the grayscale-based networks for the remainder of this work.

5.2. Discovering Mid-level Visual Elements

To investigate what our networks are learning, we use deep pattern mining [16] to find the visual elements of image patches that are both representative and discriminative. We extracted features by pushing the images through the trained model of *gray_torso2year* and taking the output of the "fc6" layer which is of dimension $4\,096$. Then, we use the Apriori algorithm to find the set of patterns, P, that have the following two conditions:

$$\text{support}(P) > \text{support}_{min},$$
$$\text{confidence}(P \rightarrow pos) > \text{confidence}_{min}.$$

For this experiment, we used 0.1% as the minimum support and 70% as the minimum confidence. We processed

Figure 6: Discriminative mid-level visual elements found using deep pattern mining.

Figure 7: Visualizations that highlight regions of the image (unoccluded) that have the largest impact on the predicted distribution over years. Occluded image regions have little impact on the prediction.

100 images from each year in the period 1950–2014 and found that many of the discriminative clusters have patches capturing semantic details of human appearance. Figure 6 show several such examples.

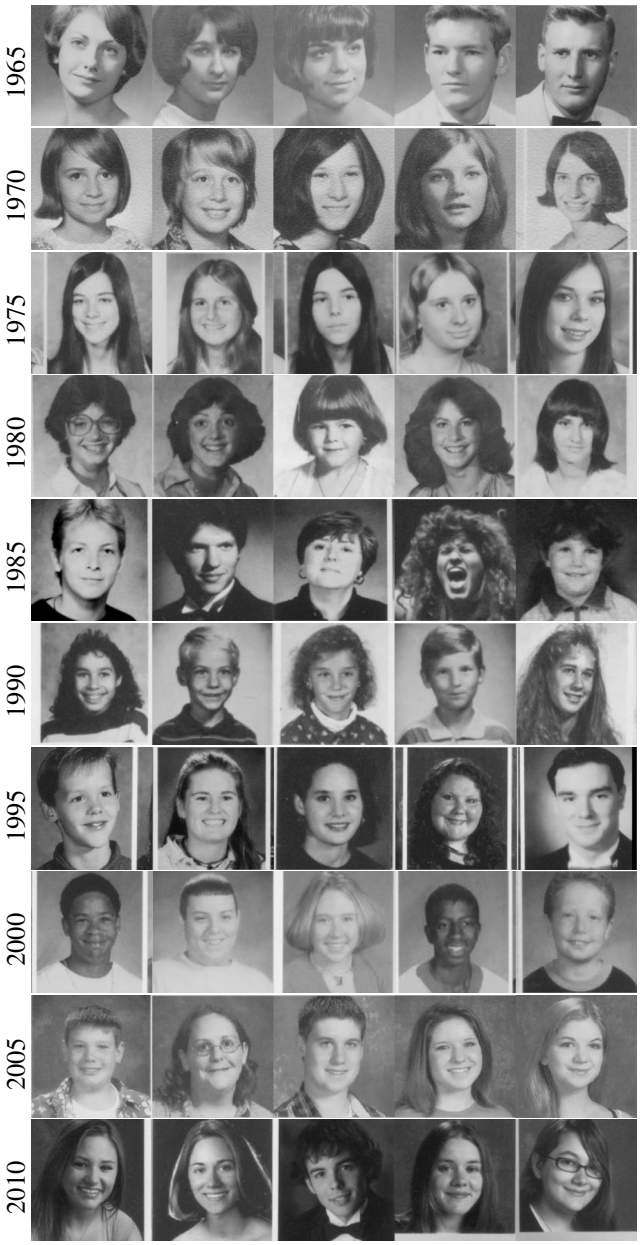
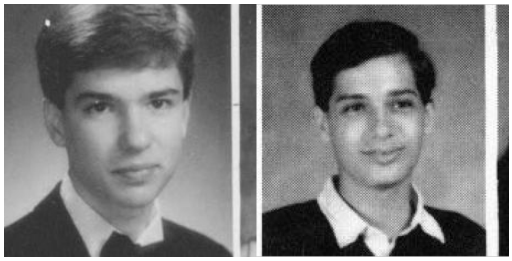


Figure 8: Each row contains the five faces with the highest probability of being from the specified year.



(a) From '90s but looks like from '60s.

(b) From '60s and looks like from '60s.

(c) From '90s and looks like from '90s.

Figure 9: A timeless sense of style? (a) Two individuals from the '90s predicted to have the highest probability of being from the '60s. (b,c) For comparison, individuals predicted to have the highest probability of being from their respective decades.

5.3. Image-Dependent Sensitivity Analysis

To better understand what our networks have learned, we perform a form of sensitivity analysis, inspired by [27], using the *gray_face2year* network. Given a query image, we slide an 11×11 window (size of the receptive field of the first convolutional layer neuron) across the input image, setting all pixel intensities under the window to be the mean value and then passing the resulting image through the network. For each location, we record the Euclidean distance between the output vector of the filtered image and that of the original image. Intuitively, regions that significantly impact the prediction will cause larger output changes when they are blocked. Figure 7 shows examples of the output of this analysis. Regions that did not impact the prediction are colored black and regions that do are left unoccluded. We find that this analysis often highlights shirt collars, eye glasses, and hair.

5.4. Finding Representative People

Here we explore further what the *gray_face2year* network is learning by looking at individuals that lead to extreme predictions. Figure 8 shows, for every fifth year, the five individuals with the most confident predictions to be from that year. For example, the top row contains three women and two men for which the network has the greatest confidence that they are from 1965. One step further, we ex-

plore people that are "out of time"; they appear to be from an era from which they are not. Figure 9 shows three sets of people: 1) people from the 1960s that look most like people from the 1960s, 2) people from the 1990s that look most like people from the 1960s, and 3) people from the 1990s that look most like people from the 1990s. These predictions were obtained by summing yearly probabilities from *gray_face2year*. While it is difficult to know for sure, the network seems to have identified differences in shirt collars and hair styles that are typical of the respective eras.

6. Conclusion

We introduced a large dataset of timestamped images of people and used it to evaluate the performance of a CNN-based strategy for estimating when an image was captured. We found that when applied to color or grayscale images the networks were able to predict the year from images of faces and torsos with better performance than when provided with random patches. While some of this is likely due to the more consistent input layout (centered faces), through several experiments we show that the networks learn semantic aspects of appearance, both clothing styles and hair styles that are typical of different eras. This is especially true when the input imagery is grayscale, since the network cannot rely on color alone to make predictions.

While our interest was in understanding trends in human appearance, it is impossible using current techniques to completely isolate these changes from changes in camera technology. We are actively exploring approaches to overcome this and to further extend this methodology to general Internet images and to other object classes, such as household products and vehicles.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3

[2] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4), 2012. 2

[3] Q. Fang, J. Sang, and C. Xu. Discovering geo-informative attributes for location recognition and exploration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1s), 2014. 2

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 2010. 3

[5] B. Fernando, D. Muselet, R. Khan, and T. Tuytelaars. Color features for dating historical color images. In *ICIP*, 2014. 1, 2

[6] S. Ginosar, K. Rakelly, S. Sachs, B. Yin, and A. A. Efros. A century of portraits: A visual historical record of american high school yearbooks. In *IEEE ICCV Workshop on Extreme Imaging*, 2015. 2

[7] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. 2

[8] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 2

[9] M. T. Islam, S. Workman, H. Wu, N. Jacobs, and R. Souvenir. Exploring the geo-dependence of human face appearance. In *WACV*, 2014. 2

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014. 4

[11] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10, 2009. 3

[12] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[13] S. Lee, N. Maisonneuve, D. Crandall, A. A. Efros, and J. Sivic. Linking past to present: Discovering style in two centuries of architecture. In *ICCP*, 2015. 1, 2

[14] Y. J. Lee, A. Efros, M. Hebert, et al. Style-aware mid-level representation for discovering visual connections in space and time. In *ICCV*, 2013. 1, 2

[15] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE Workshop on Analysis and Modeling of Faces and Gestures*, 2015. 2

[16] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Mid-level deep pattern mining. In *CVPR*, 2015. 5

[17] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. In *ECCV*, 2012. 1, 2

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 2015. 3, 4

[19] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, 2015. 2

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2

[22] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. In *WACV*, 2015. 2

[23] X. Wang, D. F. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015. 2

[24] S. Workman and N. Jacobs. On the location dependence of convolutional neural network features. In *IEEE/ISPRS Workshop on Looking from above: When Earth observation meets vision*, 2015. 2

[25] S. Workman, R. Souvenir, and N. Jacobs. Wide-area image geolocalization with aerial reference imagery. In *ICCV*, 2015. 2

[26] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *ACCV*, 2015. 2

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015. 6

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 2

[29] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *ECCV*, 2014. 2