Walk and Learn: Facial Attribute Representation Learning from Egocentric Video and Contextual Data

Jing Wang
Northwestern University
jing.wang@u.northwestern.edu

Yu Cheng
IBM T. J. Watson
chengyu@us.ibm.com

Rogerio Schmidt Feris
IBM T. J. Watson
rsferis@us.ibm.com

Abstract

The way people look in terms of facial attributes (ethnicity, hair color, facial hair, etc.) and the clothes or accessories they wear (sunglasses, hat, hoodies, etc.) is highly dependent on geo-location and weather condition, respectively. This work explores, for the first time, the use of this contextual information, as people with wearable cameras walk across different neighborhoods of a city, in order to learn a rich feature representation for facial attribute classification, without the costly manual annotation required by previous methods. By tracking the faces of casual walkers on more than 40 hours of egocentric video, we are able to cover tens of thousands of different identities and automatically extract nearly 5 million pairs of images connected by or from different face tracks, along with their weather and location context, under pose and lighting variations. These image pairs are then fed into a deep network that preserves similarity of images connected by the same track, in order to capture identity-related attribute features, and optimizes for location and weather prediction to capture additional facial attribute features. Finally, the network is fine-tuned with manually annotated samples. We perform an extensive experimental analysis on wearable data and two standard benchmark datasets based on web images (LFWA and CelebA). Our method outperforms by a large margin a network trained from scratch. Moreover, even without using manually annotated identity labels for pre-training as in previous methods, our approach achieves results that are better than the state of the art.

1. Introduction

Describing people based on attributes, such as gender, age, hair style and clothing style, is an important problem for many applications, including suspect search based on eyewitness descriptions [11], fashion analytics [30, 18, 5], face retrieval and verification [27, 2, 31], and person re-identification [22, 38]. In this work, we address the problem of learning rich visual representations (i.e., “good features”) for modeling person attributes without manual labels, with a focus on facial attribute prediction.

The state of the art in facial attribute classification, as demonstrated by standard evaluation benchmarks [31], has been advanced by methods that use deep convolutional neural networks (CNNs) pre-trained on massive amounts of images that have been manually annotated with identity labels. In fact, it has been shown that identity-related attributes such as gender, hair color, and age are implicitly encoded in nodes of CNNs that are trained for identity discrimination [31, 42]. Despite the excellent performance, the feature representation learned by these methods requires *costly manual annotation* of hundreds of thousands or even millions of images in the pre-training stage. Moreover, the pre-trained network fails to encode attributes that are not related to identity, such as eyewear and different types of hats.

In this paper, we address these issues by taking a different approach. Instead of relying on manually annotated images from the web, we learn a discriminative facial attribute representation from *egocentric videos* captured by a person walking across different neighborhoods of a city, while leveraging discretized geo-location and weather information readily available in wearable devices as a free source of supervision. The motivation for using location and weather data to construct facial attribute representations is illustrated in Figure 1. In New York City, for example, the likelihood of meeting an Afro-American casual walker in certain regions of Harlem is more than 90%. The same is true for Hispanics in Washington Heights, East Asians in Flushing, South Asians in India Square, East Europeans in Brighton Beach, and so on. These groups are characterized by their unique facial attributes (hair color, hair length, facial and eyes shape, etc.). Moreover, the weather conditions influence the facial appearance changes due to lighting variations and also dictate the clothing and accessories people wear. As an example, on sunny and warm days, the likelihood that a person will wear sunglasses, baseball hats, t-shirts, and shorts increases, whereas the presence of scarfs, beanies, and jackets is much more frequent in cold days.

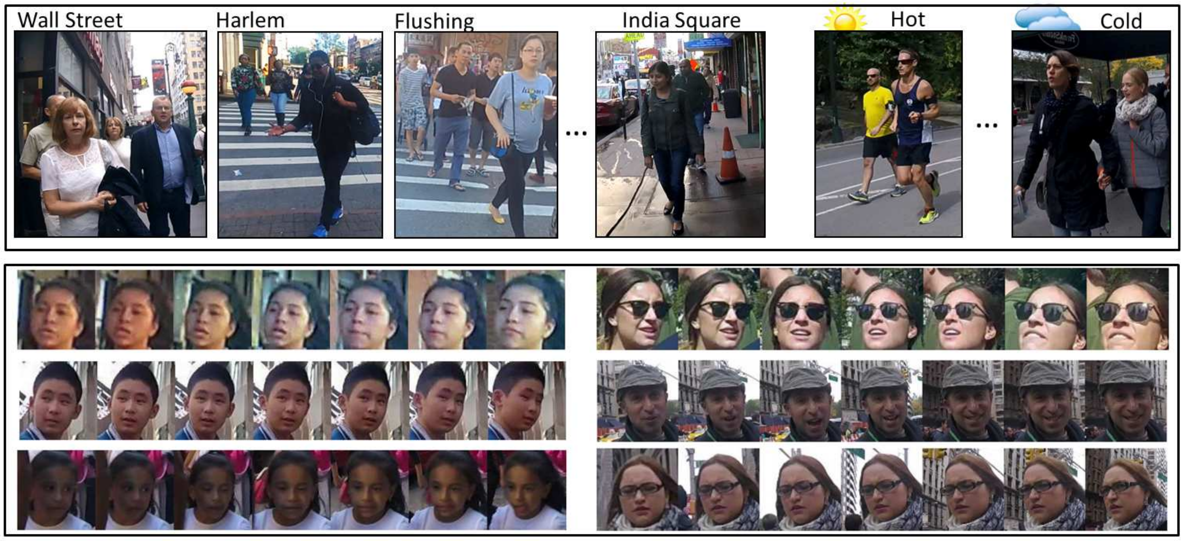


Figure 1: **Top:** Casual walkers as imaged by people with wearable cameras walking across different neighborhoods of New York City. Due to changes in demographics, the expected appearance of facial attributes is highly dependent on location. Moreover, the weather conditions change facial appearance due to different lighting, and influence the choice of outfit and the use of accessories such as sunglasses, hats, and scarfs. **Bottom:** Face images obtained via face detection and landmark tracking. Note the large variations in lighting, expression, face pose, ethnicities, and accessories. We exploit this information to build rich visual feature representations for facial attribute classification.

Our goal is to leverage data about location and weather as weak labels to construct rich facial attribute representations.

Overview of our Approach. Our proposed feature learning method relies on processing identity-unlabeled data and learning feature embeddings from a few supervised tasks. We first track the faces of casual walkers using facial landmark tracking in more than 40 hours of egocentric video, obtaining face images under a variety of conditions, as shown in Figure 1. These face images are then arranged into pairs, where information from tracking is used to label the pairs as belonging to the same individual or not. Nearly 5 million pairs are generated and fed into a network that encodes identity-related features through a Siamese structure with contrastive loss, while further embedding contextual features based on location and weather prediction. Finally, the obtained feature representation is fine-tuned with manual labels for the task of facial attribute classification.

Generally, our proposed feature representation learning for person attribute modeling has the following advantages over previous methods: First, it does not require costly manual annotation in the pre-training stage. Second, by leveraging location and weather information, it encodes facial features beyond identity, in contrast to methods pre-trained on large image repositories with identity labels [2, 31]. Third, it leverages the rich appearance of faces from a large number of casual walkers at different locations and lighting conditions, which may not be captured by images available on the web.

Our main **contributions** can be summarized as follows:

1. We introduce a new *Ego-Humans dataset* containing more than 40 hours of egocentric videos captured by people with wearable cameras walking across different regions of New York City. The data covers tens of thousands of casual walkers and includes both the weather and location context associated with the videos.

2. To the best of our knowledge, this is the first time a "walk and learn" approach that leverages discretized geo-location and weather information has been proposed for constructing deep visual representations for person attribute modeling. Our method seamlessly embeds this contextual information in a Siamese network that measures similarity of face pairs automatically extracted from tracks.

3. We show that our *self-supervised* approach can match or exceed the performance of state-of-the-art methods that rely on supervised pre-training based on hundreds of thousands or millions of annotated images with identity labels. In addition, we show that facial attributes are implicitly encoded in our network nodes as we optimize for location, weather, and face similarity prediction.

2. Related Work

Egocentric Vision. First-person vision methods have received renewed attention by the computer vision community [26, 28, 35]. Current methods and datasets have focused on problems such as video summarization [26, 45], activity recognition [10, 52], and social interaction [9]. In contrast, our work is focused on the problem of *looking at people* and modeling facial attributes from a first-person vision perspective. Compared to existing egocentric datasets [3], our Ego-Humans dataset is the first of its kind; it deals with a different task, it is larger in scale, and it also has associated geo-location and weather information, which could be relevant for many other tasks.

Facial Attribute Modeling. Kumar et al. [21] proposed a method based on describable facial attributes to assist in face verification and attribute-based face search. Siddiquie et al. [39] and Luo et al. [32] exploited the inter-dependencies of facial attributes to improve classification accuracy. Chen et al. [4] built a feature representation that relies on discrimination of images based on first names, and showed improved results in age and gender classification. Berg and Belhumeur [2] introduced part-based one-vs.-one features (POOFs) and showed that features constructed based on identity discrimination are helpful for facial feature classification. Li et al. [27] proposed a method that jointly learns discriminative binary codes and attribute prediction for face retrieval.

More recently, deep convolutional neural networks have advanced the state of the art in facial attribute classification. N. Zhang et al. [55] proposed pose-aligned networks (PANDA) for deep attribute modeling. Z. Zhang et al. [56] proposed a deep model based on facial attributes to perform pairwise face reasoning for social relation prediction. Luo et al. [31] achieved state-of-the-art performance on the LFWA and CelebA datasets using a network pre-trained on massive identity labels. Our work, instead, achieves the same or superior performance without requiring manually annotated identity labels for the pre-training step.

Geo-Tagged Image Analysis. Many methods have been proposed for geo-tagged image analysis. In particular, image geo-localization, i.e., the problem of predicting the location of a query image, has received increased attention in the past few years [13, 12, 29, 25]. Other related research includes discovering architectural elements and recognizing city attributes from large geo-tagged image repositories [8, 58] and using location context to improve image classification [44]. More closely related to our work, Islam et al. [50] investigated the geo-dependence of facial features and attributes; however they used off-the-shelf facial attribute classifiers for this analysis, whereas the goal of our work is to build feature representations so as to improve the accuracy of facial attribute classifiers.

Representation Learning. Most high-performance computer vision methods based on deep learning rely on visual representations that are learned based on *supervised pre-training*, for example, using networks trained on millions of annotated examples such as the ImageNet dataset for general object classification [6, 19], or relying on massive amounts of identity labels for facial analysis tasks [31, 42]. Our work, instead, is focused on building rich visual representations for person attribute classification without using manual annotations in the pre-training step.

There is a long history of methods for unsupervised learning of visual representations based on deep learning [14, 54, 57]. When large collections of unlabeled still images are available, auto-encoders or methods that optimize for reconstruction of the data are popular solutions to learn features without manual labeling [23, 57, 46]. Doersch et al. [7] proposed learning supervised "pretext" tasks between patches within an image as an embedding for unsupervised object discovery. These approaches, however, have not yet proven effective in matching the performance of supervised pre-training methods.

When video data is available, additional regularization can be imposed by enforcing temporal coherence [33, 47] or through the so called slow feature analysis [49]. More recently, Srivastava et al. [40] used multilayer Long Short Term Memory (LSTM) networks to learn representations of video sequences, combining auto-encoders and prediction of future video frames.

Our work is related to other methods that learn visual representations from videos captured by wearable cameras or vehicle-mounted cameras [16, 1], where awareness of egomotion can be used as a supervisory signal for feature learning. In contrast to those methods, however, we leverage the geo-location and weather data that are readily available in wearable sensors as a source of free supervisory signal to learn rich visual representations which are suitable to facial attribute classification.

3. Ego-Humans Dataset

Data Collection. The Ego-Humans dataset was collected in New York City over a period of two months, from August 28 to October 26 (during the summer and fall seasons). The data consists of videos captured by three people with chest-mounted cameras, walking across different neighborhoods of the city. Two camera models were used: a GoPro camera (higher-quality) and a Vievu camera (lower-quality), both with 1080p resolution, capturing data at 30 frames per second. Within the two-month period, 25 days were selected for data collection, covering different regions of Manhattan and nearby areas, including the Financial District, Times Square, Central Park, Harlem, Little Italy, Brooklyn Bridge, Chinatown, Flushing, and others. In each day, one or more hours of video were recorded, at different times of the day and in different weather conditions. In total, we recorded more than 40 hours of egocentric videos, split into chunks of 15 minutes. In association with these videos, we recorded location using a GPS sensor and detailed weather information, such as temperature, precipitation, and weather condition, using an open weather API that

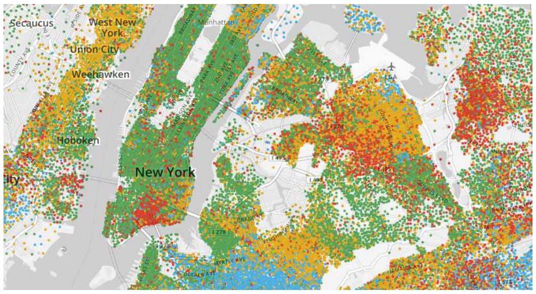


Figure 2: Grouping of GPS coordinates based on an ethnicity map defined by census data (best viewed in color).

retrieves this information based on geographic coordinates.

Discretization of Contextual Data. Rather than relying on fine-grained GPS coordinates, our learning algorithm considers a coarse set of locations as class labels. More specifically, we cluster GPS coordinates according to published census/ethnicity data [1]. In particular, we consider four ethnical groups: *White, Black, Asian, and Indian*. Figure 2 shows an ethnicity map segmented based on census data, where each cluster has its own peculiar predominance of facial attributes. We are currently expanding this set (including Hispanics, for example) as we capture more data in other locations. Regarding weather, our data includes a variety of temperatures and conditions, but for training we have used two classes: *sunny/hot and cloudy/cold*. We note that other partitions of our data could be used for other tasks. As an example, for clothing attributes, GPS clustering based on socio-economic factors could be relevant, as well as finer-grained weather conditions and temperatures.

In addition to extracting the weather and location labels, it is also important to generate face pairs (similar and dissimilar) for encoding identity features, which are helpful for discriminating several facial attributes. This procedure consists of two steps: 1) tracking casual walkers via face detection and landmark tracking and 2) image pair selection.

Tracking Casual Walkers. We used the OpenCV frontal face detector and facial landmark tracking based on the supervised descent method (SDM) [51] to track casual walkers in the videos. The detector was tuned to output only high-confidence detections, with virtually no false alarms, at the expense of more false negatives. We used the *intraface* implementation of the SDM landmark tracking [2], which works remarkably well, greatly expanding the set of captured face poses, lighting, and expressions as illustrated in Figure 1, without drifting. In total, we collected 15,000 face tracks, for a total of 160,000 face images.

Table 1: Statistics of the Ego-Humans Dataset.

Collection period	08/28 – 10/26
No. of days	25
Video footage	4̃0 hours
Contextual info	GPS and Weather Data
No. of face tracks	15,000
No. of face images	160,000
No. of generated face pairs	4.9 million

Selecting Informative Pairwise Constrains: Given the face images extracted by face detection and tracking, we consider the following pairwise constraints:

- Temporal information: two faces connected by the same track can be assumed to belong to the same person. Conversely, two faces detected at the same video frame at different locations do not belong to the same person.
- Geo-location: two faces captured from totally different geographic areas are assumed to be from different people.

Based on these constraints, we generate nearly 5 million face pairs, along with their *same/not same* labels. As detailed in the next section, preserving similarity of face pairs connected by the same track improves robustness to lighting and pose variation, and learning features to discriminate different individuals is important for the final facial attribute classification task. Table 1 summarizes the information about our data.

4. Facial Attribute Representation Learning

In the previous section, we introduced our unique Ego-Humans dataset. Next, we describe how we use this data to build a rich visual representation for facial attribute classification, based on a deep network that encodes features related to facial similarity, as well as weather and location information, without requiring manual annotations in the pre-training stage.

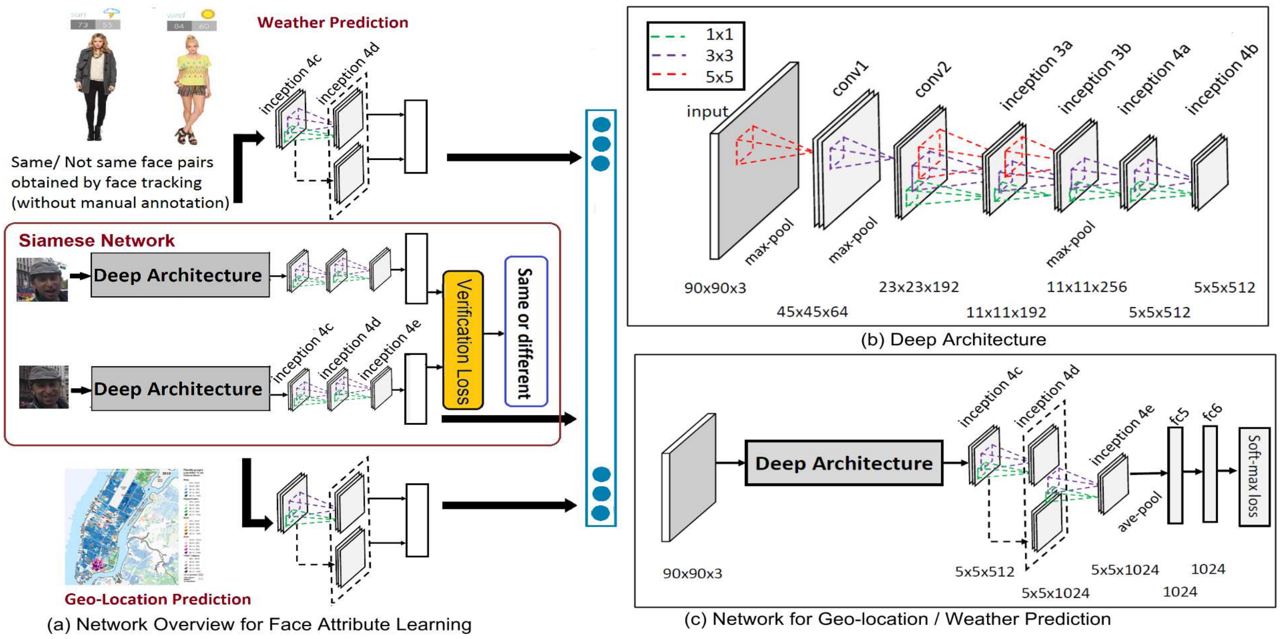
4.1. Learning Objective

Our learning framework builds upon the millions of face pairs automatically generated based on face detection and landmark tracking as described in the previous section, along with weather and location information. Our goal is to learn good features for facial attribute classification by leveraging this data. Specifically, given a face image $\mathbf{x_i} \in \mathcal{X}$ in the original pixel space, our goal is to obtain its associated facial representation $\mathbf{r_i} \in \mathcal{R}^{\mathbf{N}}$, so that a facial attribute classifier can be constructed on top of $\mathbf{r_i}$ (e.g., via network fine-tuning with a small set of manual labels).

In our learning algorithm, we have a training set $\mathcal{U}$ of N_u face pairs, $\mathcal{U} = \{(\mathbf{x}_i, \mathbf{x}_j); y_{i,j}\}$, where $y_{i,j} \in \{1, -1\}$ indicates whether $(\mathbf{x_i}, \mathbf{x_j})$ are images of the same person or not. In addition, we also have another two training sets $\mathcal{L}^w$

[1] http://projects.nytimes.com/census/2010/explorer

[2] http://www.humansensing.cs.cmu.edu/intraface/



(a) Network Overview for Face Attribute Learning

(b) Deep Architecture

(c) Network for Geo-location / Weather Prediction

Figure 3: (a) Overview of the proposed network for facial attribute learning. (b) The base deep architecture model. (c) The network for location/weather prediction

of N_w images, $\mathcal{L}^w = \{\mathbf{x}_k; c_k\}_{k=1}^{N_w}$, where $c_k \in \{1, ..., C_w\}$ indicates the label of weather; and $\mathcal{L}^g = \{\mathbf{x}_l; c_l\}_{g=1}^{N_g}$, where $c_l \in \{1, ..., C_g\}$ indicates the label of geo-location. We use discretized values for weather and location labels as described in the previous section.

The learned feature $\mathbf{r_i}$ should capture identity-related attributes (embedding in $\mathcal{U}$) and also preserve the high-level factors in $\mathcal{L}^w$ and $\mathcal{L}^g$. Towards this goal, the deep network is initially trained over $\mathcal{U}$ by minimizing the verification loss $\mathrm{d_e}(\cdot)$ (to be described next) for face verification using a Siamese network structure. To learn high-level features from $\mathcal{L}^w$ and $\mathcal{L}^g$, we train weather and location networks independently by minimizing their own softmax loss functions. The two contextual networks are initialized by the weights from the verification-trained model on the bottom layers and fine-tuned with individual contextual labels. The feature $\mathbf{r_i}$ is the concatenation of the learned feature vectors of the top layer from each network and is further applied to train the facial attribute model.

4.2. Deep Network Structure

To learn the embedding in $\mathcal{U}$ we design a Siamese network. A Siamese network consists of two *base networks* which share the same parameters. The Siamese structure is depicted in Figure 3(a). For our experiments, we take images with a size of $90 \times 90 \times 3$ as input. The size of the face is constrained by the image quality and the resolution from the videos. The base network uses the GoogLeNet style architecture in [37]. This deep architecture contains two convolutional layers and six layers of inception modules [43] as shown in Figure 3(b). Due to the small input size, our architecture removes 5×5 filters from the inception models from layer inception 4a to inception 4d. The network contains around 4 million parameters. In the Siamese network, we connect the deep architecture with three inception modules and one fully connected layer.

The weather and location models share the same base architecture as the Siamese network, but do not share parameters at the top layers. In particular, as illustrated in Figure 3(c), we feed the fully-connected layer with inception modules 4c and 4d. This allows us to capture more localized features in the weather and location models, while encoding more global similarity in the identity verification model. In the three models (identity verification, weather, and location), the output feature vectors of the top fully connected layer are all 1024-dimensional vectors and are further concatenated to form the final facial attribute feature representation. We implemented the network using the Caffe deep learning toolbox [17]. The complete network structure is shown in Figure 3(a).

Loss Function: The Siamese network used to generate identity-related attribute features uses contrastive loss to preserve visual similarity of faces connected by the same track and dissimilarity to other tracks. The contrastive loss $\mathrm{d_e}(\cdot)$ is defined as:

$$\begin{aligned}\mathrm{d_e}(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}, \mathrm{y_{i,j}}) = \mathbb{1}(\mathrm{y_{i,j}} = 1)\mathrm{d}(\mathbf{x_i}, \mathbf{x_j}) + \\ \mathbb{1}(y_{i,j} = -1)\max(\delta - \mathrm{d}(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}), 0)\end{aligned} \quad (1)$$

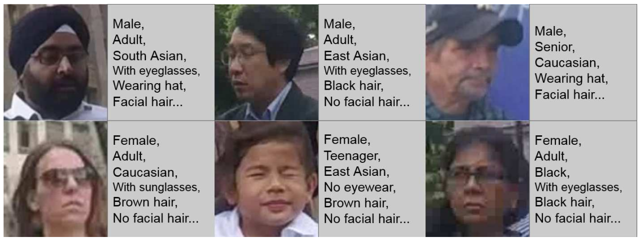


Figure 4: Example of annotated wearable data with predicted attributes

methods	Adult	Bald	Black	Black Hair	Brown Hair	Caucasian	East Asian	Eyeglasses	Facial Hair
random weights	84	90	86	71	71	81	84	81	83
id(Ego-Humans)	90	92	93	73	73	87	87	88	90

	Gender	Hat	No Eyewear	Senior	South Asian	Sunglasses	Teenager	Yellow hair	**Average**
random weights	82	82	86	85	83	89	85	85	82
id(Ego-Humans)	88	91	92	89	88	91	87	87	**87**

Table 2: Attribute prediction results of training the base net from scratch and with models after pre-training based on identity verification using the Ego-Humans dataset and the CASIA dataset.

where $\mathbb{1}(\cdot)$ is the indicator function. This contrastive loss penalizes the distance between $\mathbf{x_i}$ and $\mathbf{x_j}$ in positive mode, and pushes apart pairs in negative mode up to a minimum margin distance specified by the constant δ. We use the l_2 norm for the distance measure. The parameters of the network are updated using stochastic gradient descent (SGD) [48] by standard error back-propagation [24, 36]. The weather and location prediction models use the softmax loss as mentioned earlier.

Fine-Tuning for Attribute Learning. After we obtain our pre-trained model based on the optimization described previously, the next step is to use standard fine-tuning with images manually labeled with facial attribute labels. Additional output layers are added for fine-tuning and the cross-entropy loss is used for attribute classification.

5. Experiments

5.1. Ablation Studies on Wearable Data

In this section, we first analyze the effectiveness of each component of our network on our wearable dataset. We have manually annotated 2714 images from 25 egocentric videos randomly selected from the data described in Section 3. The faces in this dataset have large variations in pose and resolution. Each annotated image contains seventeen facial attributes covering global attributes (e.g., gender, ethnicity, age) and local features (e.g., eyewear, hair color, hat). All attributes are further categorized into binary class tasks. For this dataset, we randomly select 80% of the data as the training set and keep the rest for testing.

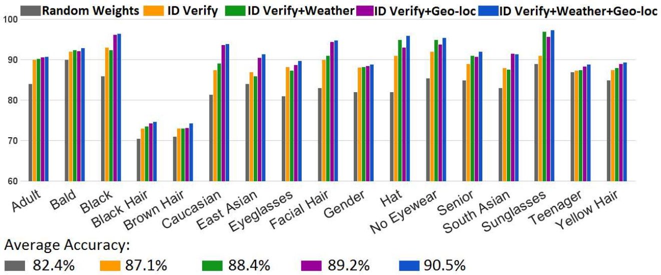


Figure 5: Results of various baseline methods on the annotated wearable dataset. The embedding of location and weather net features help boost the performance, especially on ethnicity and non-identity related attributes.

Analysis of the Verification Model. We first consider our base network (without the location and weather models). We evaluate the performance of training (fine-tuning) this network with the few available manual labeled examples, considering the following cases:

1) training from scratch: The network is initialized with random weights and the global learning rate is set as 0.001.

2) id(Ego-Humans): Training with our pre-trained model based on identity verification with 5M image pairs automatically extracted from our Ego-Humans dataset. After pre-initializing the network with the weights learned from the verification models, we set the fine-tuning global learning rate with 0.0001, but with a learning rate in the top two layers of 10 times the global learning rate.

Both cases run through the whole wearable training data with 100 epochs in attribute learning. The results in Table 2 demonstrate that the pre-trained model outperforms training from scratch by a large margin. This is not surprising, given the relatively small training size of the dataset and the large variations in pose and lighting. This demonstrates the richness of our verification model obtained from unlabeled egocentric videos.

Analysis of the Geo-Location and Weather Models. Now we evaluate the benefit of features learned from geo-location and weather models. We perform experiments on the annotated wearable dataset by fine-tuning the network, considering the identity verification model only (id-verify), the inclusion of geo-location (id-verify + geo-loc), weather (id-verify + weather) and geo-location and weather models concatenated (id-verify + weather + geo-loc). The results are shown in Figure 5. The performance indicates average improvements of 2%, 2% and 4% when concatenating the base-net with the features fine-tuned from the geo-location,

dataset	**methods**	5-o-shadow	Arch. Eyebrows	Attractive	Bags Un. Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bussy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	H.Cheekbones	Male
LFWA	FaceTracer	70	67	71	65	77	72	68	73	76	88	73	62	67	67	70	90	69	78	88	77	84
	PANDA-w	64	63	70	63	82	79	64	71	78	87	70	65	63	65	64	84	65	77	86	75	86
	PANDA-l	84	79	81	80	84	84	73	79	87	94	74	74	79	69	75	89	75	81	93	86	92
	LNets+ANet	84	82	83	83	88	88	75	81	90	97	74	77	82	73	78	95	78	84	95	88	94
	Ours	76	82	82	91	82	93	75	92	93	97	86	83	78	79	81	94	80	91	96	96	93
CelebA	FaceTracer	85	76	78	76	89	88	64	74	70	80	81	60	80	86	88	98	93	90	85	84	91
	PANDA-w	82	73	77	71	92	89	61	70	74	81	77	69	76	82	85	94	86	88	84	80	93
	PANDA-l	88	78	81	79	96	92	67	75	85	93	86	77	86	86	88	98	93	94	90	86	97
	LNets+ANet	91	79	81	79	98	95	68	78	88	95	84	80	90	91	92	99	95	97	90	87	98
	Ours	84	87	84	87	92	96	78	91	84	92	91	81	93	89	93	97	92	95	96	95	96

		Mouth S. O.	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Reced. Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wear. Earings	Wear. Hat	Wear. Lipstick	Wear. Necklace	Wear. Necktie	Young		**Average**
LFWA	FaceTracer	77	83	73	69	66	70	74	63	70	71	78	67	62	88	75	87	81	71	80		74
	PANDA-w	74	77	68	63	64	64	68	61	64	68	77	68	63	85	78	83	79	70	76		71
	PANDA-l	78	87	73	75	72	84	76	84	73	76	89	73	75	92	82	93	86	79	82		81
	LNets+ANet	82	92	81	79	74	84	80	85	78	77	91	76	76	94	88	95	88	79	86		84
	Ours	94	83	79	75	84	87	93	86	81	77	97	76	89	96	86	97	95	80	89		**87**
CelebA	FaceTracer	87	91	82	90	64	83	68	76	84	94	89	63	73	73	89	89	68	86	80		81
	PANDA-w	82	83	79	87	62	84	65	82	81	90	89	67	76	72	91	88	67	88	77		79
	PANDA-l	93	93	84	93	65	91	71	85	87	93	92	69	77	78	96	93	67	91	84		85
	LNets+ANet	92	95	81	95	66	91	72	89	90	96	92	73	80	82	99	93	71	93	87		87
	Ours	97	90	79	90	79	85	77	84	96	92	98	75	85	91	96	92	77	84	86		**88**

Table 3: Performance comparison with state of the art methods on 40 binary facial attributes

weather, and geo-location + weather, respectively. The geo-location model provides more complementary information to the verification network on ethnicities like East Asian and South Asian. And the weather model adds in weights for non-identity-related but weather-related attributes like sunglasses and hat. Figure 4 illustrates some examples of attribute prediction in our data.

5.2. Comparison with the State of the Art

In this section we evaluate the effectiveness of our network with quantitative results on two standard facial attribute datasets, CelebA and LFWA, constructed based on face datasets CelebFaces [41] and LFW [15], respectively. Both datasets have forty binary facial attributes, as listed in Table 3. We use the exact same partition of data as in [31]: 160k images of CelebA are used to fine-tune the network. In the remaining 40k CelebA images and the LFWA dataset, 50% of the images are used to extract the learned top-layer fc features from the network and to train a linear SVM classifier for attribute classification, and the other 50% are used for testing.

We evaluate the performance of our network on the two datasets with four state-of-the-art methods: FaceTracer [20], two versions of PANDA [55] network, PANDA-w and PANDA-l, based on the setting described in [31]; and LNet+ANet [31]. The same data was used for all approaches. FaceTracer utilizes hand-crafted features (HOG + color histogram) on face functional regions to train an SVM classifier. LNet+ANet uses a massive set of images with manually labeled identities for pre-training and cascades two networks to automatically detect the face region and consequently learn the facial attributes from the detected part.Apart from LNet+ANet, all the methods obtain cropped faces externally either from given landmark points (FaceTracer and PANDA-l) or based on off-the-shelf detection (PANDA-w and ours).

As shown in Table 3, our approach significantly outperforms the four other methods on LFWA and reaches comparable performance with LNet+ANet on CelebA on average score, without using manual labeling in the pre-training stage. Our approach achieves better results than the prior methods on most of the forty attributes.

5.3. Visual Attribute Discovery

The quantitative results for the above three datasets show that the pre-trained models on identity verification, geo-location, and weather classification boost the prediction of facial attributes despite the fact they are not explicitly

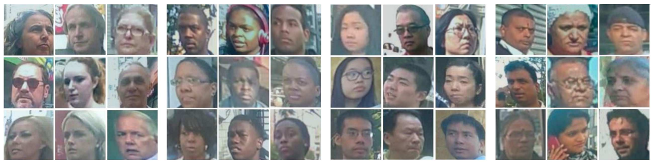


Figure 6: Top selected images in each class with max activations on layer inception 4e after pre-training loc-net. The discretized geo-location classes according to census from left to right are: White, Black, Asian, and Indian.

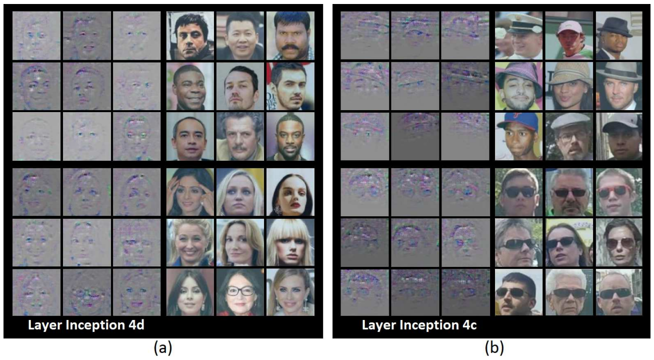


Figure 7: Visualized feature of top ranked neurons in models after pre-training identity verification (a) and pre-training weather classification (b). Best viewed in electronic form.

trained for attribute classification. To better understand the attribute-related contextual information the pre-trained nets have learned, we show some qualitative examples of the top activated neurons in the pre-training phase.

Given a layer in the pre-trained net, for example inception 4d in loc-net, we get top class-related neurons with high activations across all images within a class. For each selected neuron, we further select the most related image with highest value across the whole dataset. Figure 6 shows the corresponding selected images of the top nine neurons in each class. We can see that the selected images in different geo-locations are from different races, which means the neurons in loc-net are learning strong priors about the concept of ethnicity in the location classification.

To better visualize attributes discovered by neurons, we construct the deconvnet framework following the ideas in [53] and project selected top neurons back to the input pixel space. Figure 7 presents the visualized features of the top nine neurons of specific layers after pre-training identity verification and weather classification separately. Recovering the whole face contour with clear discriminative parts such as the eyes and mouth, the visualized results of the selected neurons in the verification model from layer inception 4d in Fig 7(a) reveal that the neurons capture global identity-related face features. Therefore, facial attributes that are intrinsic to the identity, such as "gender", can be discovered by the network. The illustrated neurons in the weather model are from layer inception 4c and 4d. The visualizations of the selected neurons partially recover the upper face and focus on similar local components. By capturing local attributes such as "sunglasses" or "hat", the visualization explicitly demonstrates that the pre-trained weather model provides complementary features on identity-non-related attributes to the model from identity verification.

6. Discussion

In the next few years, the amount of egocentric video associated with contextual information will grow significantly. As an example, many police officers around the world are already using body-worn cameras in patrol operations. This growth may be even greater as wearable devices become mainstream among ordinary people. We believe our work offers novel ways to learn rich facial representations from the ever-growing pool of unlabeled egocentric videos with contextual data. Although we have considered only the case of people walking across different neighborhoods of a city, our method could be applied at different geo-scales (e.g. worldwide) to capture larger variations.

One could argue that the face pairs generated by our approach inherit some bias. To the contrary, we have shown that in practice this is not an issue. In fact, we observed that faces across the same track exhibit large variations in pose and lighting, helping our approach to be more robust against these factors.

We would like to point out that our approach only requires location and weather data at the training stage. Although this contextual information could be useful at test time to improve accuracy, it may not always be available.

Finally, we are currently applying our approach to learn representations for fine-grained clothing attribute classification [5], as weather and location clearly influence clothing choices. By learning with diverse contextual information, the framework could be also applied to other high-level analysis tasks such as urban perception [34].

7. Conclusions

In this paper we have proposed a novel deep learning framework for learning facial attributes. Different from previous approaches, our method can capture good representations/features for facial attributes by exploiting videos and contextual data (geo-location and weather) captured by a wearable sensor as the person walks. The proposed framework can leverage the rich appearance of faces from tens of thousands of casual walkers at different locations and lighting conditions without requiring the cost of manual labels. We demonstrate our approach in several real-world datasets, showing substantial improvement over other baselines.

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *Proc. ICCV*, 2015.

[2] T. Berg and P. N. Belhumeur. POOF: Part-Based One-vs-One Features for fine-grained categorization, face verification, and attribute estimation. In *Proc. CVPR*, 2013.

[3] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 25(5):744–760, 2015.

[4] H. Chen, A. Gallagher, and B. Girod. What's in a name: First names as facial attributes. In *Proc. CVPR*, 2013.

[5] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proc. CVPR*, 2015.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. CVPR*, 2009.

[7] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proc. ICCV*, 2015.

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 31(4):101:1–101:9, July 2012.

[9] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proc. CVPR*, 2012.

[10] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Proc. ECCV*, 2012.

[11] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *Proc. ICMR*, 2014.

[12] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Proc. CVPR*, 2013.

[13] J. Hays, A. Efros, et al. Im2gps: estimating geographic information from a single image. In *Proc. CVPR*, 2008.

[14] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[16] D. Jayaraman and K. Grauman. Learning image representations equivariant to ego-motion. In *Proc. ICCV*, 2015.

[17] 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 *Proc. ACM Multimedia*, 2014.

[18] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *Proc. ICCV*, 2015.

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

[20] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *Proc. ECCV*, 2008.

[21] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1962–1977, 2011.

[22] R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014.

[23] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *Proc. ICML*, 2012.

[24] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[25] S. Lee, H. Zhang, and D. J. Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *WACV*, 2015.

[26] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015.

[27] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, and X. Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *Proc. CVPR*, 2015.

[28] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proc. CVPR*, 2015.

[29] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *Proc. CVPR*, 2013.

[30] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proc. CVPR*, 2012.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015.

[32] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *Proc. ICCV*, 2013.

[33] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *Proc. ICML*, 2009.

[34] V. Ordonez and T. L. Berg. Learning high-level judgments of urban perception. In *Proc. ECCV*, 2014.

[35] G. Rogez, J. S. Supancic, III, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *Proc. CVPR*, 2015.

[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3):1, 1988.

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, 2015.

[38] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *Proc. CVPR*, 2015.

[39] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Proc. CVPR*, 2011.

[40] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proc. ICML*, 2015.

[41] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014.

[42] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proc. CVPR*, 2015.

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

[44] K. Tang, P. Manohar, F. Li, F. Rob, and D. B. Lubomir. Improving image classification with location context. In *Proc. ICCV*, 2015.

[45] P. Varini, G. Serra, and R. Cucchiara. Egocentric video summarization of cultural tour based on user preferences. In *Proc. ACM Multimedia*, 2015.

[46] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proc. ICML*, 2008.

[47] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proc. ICCV*, 2015.

[48] D. R. Wilson and T. R. Martinez. The general inefficiency of batch training for gradient descent learning. *Neural Networks*, 16(10):1429–1451, 2003.

[49] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.*, 14(4):715–770, Apr. 2002.

[50] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs. DeepFocal: A Method for Direct Focal Length Estimation. In *ICIP*, 2015.

[51] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. CVPR*, 2013.

[52] S. P. Yu Cheng, Quanfu Fan and A. N. Choudhary. Temporal sequence modeling for video event detection. In *Proc. CVPR*, 2014.

[53] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, 2014.

[54] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Proc. CVPR*, 2010.

[55] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proc. CVPR*, 2014.

[56] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang. Learning social relation traits from face images. In *Proc. ICCV*, 2015.

[57] J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun. Stacked what-where auto-encoders. *CoRR*, abs/1506.02351, 2015.

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