Facial Attribute Recognition: A Survey

Nathan Thom and Emily M. Hand*
nathanthom@nevada.unr.edu, emhand@unr.edu
University of Nevada, Reno
* = corresponding author

November 17, 2021

1 Definition

We present a survey of attribute recognition research in the computer vision community over the past decade. Most of our attention is given to facial attributes, but attributes of objects, pedestrians, and actions are considered as well.

2 Background

Facial attributes – human-describable features of faces – were introduced to the computer vision community in 2008, with their first application being image search [1]. Kumar et al. identified a problem with the image search engines of the time, realizing that simple descriptive search terms would not produce expected face image results. Attributes were then used for face recognition and verification as well as, again, for image search and retrieval [2–4] before attribute recognition itself became the focus of research. Facial attribute recognition is related to the problem of soft biometrics [5], which is focused on identification using these so-called "soft traits" rather than recognizing them for purely descriptive purposes.

Well before the introduction of facial attributes, recognition of gender and age from faces were well-established problems in the computer vision community [6–8]. One of the earliest works on gender recognition from faces utilized a neural network that predicted sex directly from image pixels [9]. This method, like many others, required the face images to be scaled, aligned and cropped in order to perform well. In [10], the authors also took a holistic view of the face, creating so-called holons – reduced feature vectors learned via an auto-encoder – to perform identity, emotion and gender recognition. Research has shown that both age and ethnicity play a big role in gender recognition. For example, gender recognition performance has been shown to degrade when models are trained on a mixture of ethnicities, rather than focused on a target ethnicity [11]. In addition, gender recognition performance significantly depends on age, with young males and older females posing challenges for the models [12, 13].

One of the earliest works on age recognition focused on cranio-facial development theory, developing models to describe the changing shape of the face as it aged [14]. Focusing on texture as well as shape, active appearance models – statistical models – were developed for age recognition from face images [15]. Success was also found in age estimation by considering a collection of images from an individual in order to determine the aging pattern for that person [16]. Age estimation from face images remains a very challenging problem in the computer vision community because each person ages differently, and so it is a profoundly individual problem [17]. Adding to the challenge of age recognition problems, they can be considered a categorical classification problem (e.g. to what age group does this face belong?) or a regression problem (e.g. what is the age of this face?), depending on the context and data available.

Attributes exist in domains other than faces, including pedestrians, objects, and actions. An attribute is simply a describable feature, and so it lends itself nicely to many problems in computer vision. Pedestrian attributes can include clothing, gender, hair color and length, as well as part visibility and pose [18]. Attributes of objects can include multiple categories: shape, color, texture, part, and material as well as global or local presence of the attributes [19, 20]. Attributes of actions include pedestrian attributes, object attributes as well as action-specific attributes such as environment and motion [21].

The problem of attribute recognition has gained a lot of attention in the research community over the past decade, mostly due to the wide applicability of attribute prediction for real-world applications. Pedestrian attributes have been used in surveillance for re-identifying individuals and searching for suspects based on description of their visual attributes [22]. The application of attribute detection to surveillance is ultimately an image search problem where a query of attributes is provided and the most relevant results are returned. Thus, most image search techniques can directly correlate with surveillance, and many have utilized facial attributes [4, 23]. Applying attribute recognition to surveillance can lead to quicker identification and save a significant amount of human hours.

Another application of attribute recognition is in human computer interaction (HCI). Many applications that involve HCI benefit from knowledge of the user. For example, proper greetings rely on gender information (e.g. Mr., Ms.). A user's expression can determine whether or not they are enjoying an application (e.g. smiling or frowning). More specifically, attributes have been used by companies such as Facebook in order to improve accessibility of their platform, providing image descriptions to those with visual impairments [24]. Other applications of facial attributes in HCI include active authentication, the process of continuously authenticating a user on a device. Attributes have already been successfully deployed for this problem [25–27]. Being describable features by definition makes facial attributes widely applicable to many real-world problems.

Focus in attribute research has been dedicated to the general discovery of attributes for building datasets and vocabularies for preexisting data. Note that these approaches operate in a broader scope than just facial attribute recognition. [20] made strides to automatically aggregate and label data from noisy internet sources. They cite work that uses gender, race, and other attributes to improve face verification and search. Expanding on this concept, the authors mine websites that have diverse images and automatically label the images based on captions. This allows for diverse vocabulary discovery and improved predictions across multiple object types. A similar approach, with the goal of learning attributes in large datasets, draws connections between semantically unrelated objects by looking at their visually describable attributes [28]. For example, zebras, beetles, and street crossings all share the stripes attribute. With the development of several large-scale labeled datasets for the problem of facial attribute recognition, the field has grown significantly.

In the following sections we detail the research in facial attribute recognition from images and videos. We will also present work in the general field of attribute recognition, when applicable. All the while we will be introducing datasets and discussing methods based on traditional machine learning and computer vision as well as those based on deep learning. Our survey will conclude with a discussion of open problems in the field.

3 Related Research

Attributes are not solely applicable to faces. They have been successfully applied to objects, pedestrians and actions. Here we provide a brief history of each field, from early works to state-of-the-art.

3.1 Objects

Attributes of objects include textures, colors, patterns, shapes and many other describable features. Early methods for attribute recognition were focused on aiding object recognition. These initial works recognized basic patterns, textures, and colors [29] [30] [31] [32]. As researchers became more active in the field, the focus shifted to describing objects, rather than simply naming them [19] [33]. Many researchers utilized object attributes for few and zero-shot learning as they provide a compact description of objects that a system may not have seen previously. In [34], Wang et al. focused on dependencies between objects and attributes, improving both attribute and object recognition. [35] identifies attributes (e.g. shape, color, material) of 3D objects with the goal of helping autonomous robots understand and interact with the world around them. In [36], the authors present a dataset (AirplanOID) and a method for understanding objects in fine-grain detail, using attributes. The dataset contains attributes such as *facing direction*, *is-airline*, *location*, etc. More recently, Wang et al. further explore attributes for object recognition [37]. Their approach utilizes attributes as additional information during model training, requiring no attribute labels at test time.

3.2 Actions

Attributes of actions include descriptive features such as environment, pose, objects involved, etc. that can be used to break an action down into its component parts. One of the first works in action attribute recognition modeled the human visual cortex. This was accomplished by applying motion-direction sensitive units to video inputs, thereby recognizing human body, head, hand and general animal actions [38]. [21] [39] [40] all focused on identifying action parts in still images. In [21], the authors used attributes of a scene to understand actions. Yao et al. used a combination of given action verbs (e.g. bending, squatting, riding, etc.) along with poselets and objects to predict actions from still images [39]. In [40], the authors presented a method which learns a template for a variety of actions in order to localize actions in a frame. Zhang et al. presented a multi-task learning method in which attributes and actions are learned simultaneously [41]. [42] and [43] focused on attributes for action recognition in 3D. State-of-the-art methods rely on supervised deep learning in order to recognize attributes of actions [44] [45].

3.3 Pedestrians

Attributes of pedestrians include whole-body attributes such as clothing, pose, etc. as well as facial attributes. Identifying attributes in this context can be challenging due to viewpoint and extreme pose changes. With a focus on gait analysis, [46] used K Nearest Neighbors and spectral clustering to identify attributes such as gender and age from gait information including speed, acceleration, rhythm, etc. In work done by Deng et al., support vector machines were trained on a large-scale dataset to recognize attributes of pedestrians [47] [48]. The authors collected the PETA dataset for these works, which is still a benchmark in the field [47]. In recent years, deep learning has become the standard for pedestrian attribute recognition, with the focus on convolutional and recurrent neural networks [49] [50] [22] [51] [52]. Convolutional neural networks are useful for localization of pedestrian attributes, while recurrent neural networks are successful in identifying attribute relationships. Automatic recognition of pedestrian attributes has applications in soft-biometrics, surveillance and autonomous vehicle guidance.

4 Theory and Application

We review work on facial attribute recognition from images and video and separate work into two categories: traditional methods and deep learning.

Figure 1: Sample images from the FaceTracer dataset [1].

4.1 Attribute Recognition with Traditional Methods

Prior to the advent of deep learning in all aspects of computer vision, other traditional methods, such as support vector machines were used for attribute recognition. In 2008, Kumar et al. built a face search engine that they called FaceTracer [1]. This search engine operated on user queries involving one or more of the available attributes. For example, "smiling Asian men with glasses". The search engine would then return face images that exhibited the desired traits. The search engine was built on a set of attribute classifiers, capable of identifying binary facial attributes in an image. The attribute classifiers were built on four feature sets: face region, pixel data color space (e.g. RGB, HSV), normalization method, and data aggregation method. Support vector machines (SVMs) were then trained for every region, feature type, and parameter combination. Adaboost is then run on this set of "local SVMs" to generate a set of strong classifiers. Finally, a global SVM is trained by finding the union of the strong classifiers. Along with the FaceTracer search engine, Kumar et al. introduced a dataset by the same name. At the time of publication, the dataset consisted of over 3.1 million face images, 17,000 of which were manually labeled with 10 attributes: age, gender, race, hair color, eye wear, mustache, expression, blurry, lighting, and environment. Sample images from the FaceTracer dataset are shown in figure 1.

A year later, the same group shifted their research focus toward face verification using attributes [2]. Face verification aims to address the following question: given two images, do they belong to the same person? The authors developed two different methods to generate descriptions of faces, using attribute and simile classifiers. SVMs are used as attribute classifiers, trained on a collection of low-level features, similar to the previous work [1]. They introduced additional low-level features such as edge magnitudes and gradient directions. As a part of this work, the authors introduced a new dataset, PubFig, for face verification. Sample images from the PubFig dataset are shown in figure 2. Additional data was collected in order to train facial attribute classifiers. 1000 images were labeled for each of 65 binary attributes using Amazon Mechanical Turk. Simile classifiers were used as well to identify the similarity of a face to a set of reference faces. For each reference individual, a classifier is trained on each region to distinguish that region from the same region on other faces. These simile classifiers allowed for comparisons between faces without requiring additional labels. The final face verification system utilized a hybrid of attribute and simile classifiers and achieved state-of-the-art accuracy. After the release of PubFig, the authors tested their attribute classifiers on all images in the dataset, providing 65 attribute scores for each image along with the image data. Some methods utilized these scores as labels in order to train attribute classifiers [25]. In 2011, the same group again used facial attributes for improved face verification and image search [3], extending the set of attributes to 73.

Several groups realized the potential of facial attributes to improve image search and retrieval with natural queries [1,4,23,53]. With a focus on surveillance, Vaquero et al. utilized pedestrian

Figure 2: Sample images from the PubFig dataset [2].

attributes for search and retrieval in low-quality video [53]. Others have explored different ways to perform multi-attribute search queries [4, 23]. [4] improved over previous ranking methods that required individual models for each search term. Instead they used correlations amongst attributes to provide additional information to the search query. Their Multi-Attribute Retrieval and Ranking (MARR) method benefited from the strong relationship amongst attributes. The authors labeled a subset of the Labeled Faces in the Wild (LFW) dataset [54] (9992 images) with 27 binary attributes (a subset from the work of [2]). A year later, [23] focused on developing a meaningful way to combine different attribute scores. They construct a normalized score space based on Extreme Value Theory. The authors aimed to convert raw SVM output scores to a normalized score that would be more consistent with human labeling as well as with the scores of other attributes. After converting the scores, they were able to fuse them to allow for multi-attribute queries in a shared score space. The authors use the method of [3] to extract attribute scores from face images.

All of the publicly available datasets up to this point considered facial attributes to have binary values, that is, the attribute is either present or it is not. This can be a very challenging way to view the problem when many attributes are subjective or exist on a gradient (e.g. some hair is more blond than others). Parikh et al. aimed to address this issue in [55], focusing on so-called "relative attributes." The authors utilize the concept of relative attributes to generate a ranking function for each attribute allowing for a new type of zero-shot learning in which they can describe unseen objects relative to previously seen ones. They propose a learning-to-rank formulation that learns a desired ordering of the training images. This learning framework resulted in a model that could better capture the strength of a particular attribute compared to a binary classification model. The authors utilized a subset of the PubFig dataset in order to learn their ranking functions.

Labeling attributes is a time-consuming process, with each image needing multiple (over 70 in the case of [3]) labels. This became a limiting factor in facial attribute research very quickly. In [56], the authors introduce a likeness measure as a way to utilize describable features without requiring an extensive labeling process. The goal of this work is to improve face verification performance. For each pair of subjects, the authors create a classifier that is capable of distinguishing between the two subjects. This process results in many likeness classifiers, or "Tom v. Pete" classifiers, as they call them. Face images are classified using this collection of "Tom v. Pete" classifiers giving a set of scores that indicate the person's likeness to a particular subject in a pair classifier. This set of scores is then used as a subject's feature vector which is in turn used for face verification. This work resulted in human-describable features of faces, in the form of their likeness to other individuals. That is, "this person looks more like subject 1 than subject 2, and more like subject 3 than subject 2" etc. [56] built on the concept of automatically generated attributes as seen in the simile classifiers of [2].

Figure 3: Sample images from the CelebA dataset [57].

Figure 4: Sample images from the LFW (also LFWA) dataset [54, 57].

4.2 Attribute Recognition with Deep Learning

In 2015, Liu et al. introduced two large-scale benchmark datasets for the problem of facial attribute recognition in unconstrained images – CelebFaces Attributes (CelebA) and Labeled Faces in the Wild Attributes (LFWA) [57]. CelebA contains over 200,000 images each labeled with forty binary attributes, which are a subset from those used in [3] . The CelebA dataset contains a wide range of images including full body and close cropped faces. The dataset includes these original images as well as cropped and aligned face images. Sample cropped and aligned images from CelebA are shown in figure 3. Along with CelebA, attribute labels were added to the popular face verification benchmark LFW, creating LFWA. LFWA contains roughly 13,000 images, labeled with the same forty binary attributes from CelebA. Some sample images from LFW can be seen in figure 4. CelebA and LFWA were the first (and only to date) large-scale datasets introduced for the problem of facial attribute recognition from images. Prior to CelebA and LFWA, no dataset labeled with attributes was large enough to effectively train deep neural networks. With the introduction of this dataset, many deep learning methods were used for facial attribute recognition [26, 57–60].

Along with CelebA and LFWA, Liu et al. introduced a method for attribute recognition that involves two networks: LNet and ANet. LNet is a localization network that localizes the face with weak attribute supervision, and ANet uses the localized face to predict facial attributes. LNet was built on the widely popular AlexNet [61] architecture trained on the ImageNet object recognition dataset [62]. After being pre-trained on the large-scale Imagenet dataset, LNet was fine-tuned on the original (full body, unaligned) CelebA images using weak attribute supervision. With the weak supervision from CelebA's facial attributes, LNet was able to accurately localize the face in a given image. Once the face was localized by LNet, ANet was trained from the cropped face image. ANet was also built on the AlexNet architecture, pre-trained on ImageNet,

and fine-tuned on CelebA. This two-network scheme produced impressive results on CelebA and LFWA.

In 2016, Wang et al. introduced a method dubbed "Walk and Learn" in which they utilized face tracks as additional supervision for facial attribute recognition [59]. The authors collected additional data by attaching wearable cameras to their bodies and walking around different areas of New York City. They used face tracking to identify individuals in every frame and used face verification to pre-train their network. Their deep network was then fine-tuned on the CelebA dataset, producing improved results on some challenging attributes over [57].

Just a year after the release of CelebA and LFWA, many researchers began to notice some very serious label imbalance issues. In particular, [60] focused on adjusting the label imbalance during network training. The authors introduced a Mixed Objective Optimization Network (MOON) capable of learning all attributes at once while at the same time adjusting for label imbalance. We note that prior to this work, individual models were learned for each attribute, including all of the methods previously discussed [1–3, 57, 59]. This was incredibly inefficient and did not take advantage of a shared representation for all attributes. MOON addressed both of these issues by utilizing the popular VGG-16 network architecture [63] and training from random initialization on CelebA. MOON was the first to combine attribute learning into one network, address dataset imbalance, and train on CelebA from scratch rather than fine-tune. MOON addressed label imbalance by calculating source and target distributions for each attribute and applying a weight to the backpropagation within a euclidean loss in order to adjust for the distribution discrepancies. The source distribution for an attribute was the distribution of positive and negative instances of the attribute in CelebA, and the target distribution could be set to any desired distribution, though the authors experimented with an even target distribution. MOON produced impressive results on CelebA and highlighted the severe imbalance issues associated with it.

[64] also tackled the problem of multi-task learning for facial attribute recognition, utilizing a Restricted Boltzmann Machine (RBM) rather than a CNN. Their model is trained with both the aligned face images from CelebA and facial landmark points as inputs. The authors extend RBMs to handle multiple tasks and multiple inputs naming it the Multi-Task Multimodal RBM (MTM-RBM). The MTM-RBM compares favorably with [57]. To date this is the only method for facial attribute recognition that utilizes an RBM model.

With the introduction of deep learning in the facial attribute domain, many began to wonder how robust these models truely are. In [65] they aim to address this question by introducing an adversary. They develop a Fast Flipping Attribute (FFA) method that generates adversarial examples that cause classification errors. The FFA method identifies directions which can generate adversarial examples by inverting the classification score and calculating the gradient with respect to the inverted score. Searching along those gradient directions results in images that produce classification errors. The authors found that some attributes (e.g. *wavy hair* and *wearing necklace*) were more robust to adversarial attacks than others (e.g. *big nose* and *young*).

Several groups began to address the problem of facial alignment in attribute recognition. In [66], the authors introduce the Alignment-Free Facial Attribute Classification Technique (AFFACT), which performs data augmentation allowing a deep convolutional neural network to recognize attributes without first aligning the face images. The AFFACT method performs augmentation of the dataset through scaling, rotation, shifting, and blurring. Training ResNet [67] architectures, the authors applied AFFACT data augmentation to CelebA and were able to achieve state-of-the-art performance. [68] also aimed to address the problem of attribute recognition from unaligned face images by utilizing a cascade network capable of identifying different regions of the face and recognizing attributes without alignment. Their face region localization network is capable of detecting face regions based on weakly supervised attribute data. Rather than performing data augmentation like [66], this work focused on part-based

approach to attribute prediction.

Only two years after the introduction of CelebA and LFWA, performance on the benchmark datasets began to plateau. In [58], the authors aimed to improve attribute recognition accuracy by taking advantage of relationships amongst attributes both implicitly and explicitly. The authors introduce a new deep CNN architecture for attribute recognition: Multitask CNN (MCNN). MCNN had fewer than 16 million parameters compared to the 138 million parameters in the VGG-16 model used for MOON. MCNN took advantage of attribute relationships by learning a shared representation at the lower levels of the network and branching off into spatial attribute groupings at the higher levels of the network. Finally, attribute relationships were learned at the score level with an auxiliary network (AUX) that was attached to the trained MCNN. The combined network, MCNN-AUX utilizes attribute relationships in three different ways and produced state-of-the-art results on CelebA and LFWA.

Aiming to utilize localization cues to improve facial attribute prediction, [69] combined the problem of facial attribute recognition with that of semantic segmentation. Semantic segmentation requires predicting a label for every pixel in an image, producing a class map over the entire image. The authors aggregated face segments, provided as a part of the Helen Dataset [70], to create seven segments: background, hair, face skin, eyes, eyebrows, mouth and nose. They utilize a gating mechanism to focus the attribute recognition network on regions of interest for a particular attribute. For example, they focus mouth-related attributes (e.g. smiling, mouth open) on the mouth segment provided by the segmentation method.

Along a similar vein, [71] uses generative adversarial networks (GANs) to generate abstraction images that are then used to improve facial attribute recognition through a multi-stream network acting on the abstraction and original images. The abstraction images produce a kind of facial segmentation with textual information, localizing parts and providing additional supervision to the facial attribute recognition task. The multi-stream abstraction image formulation for attribute recognition outperformed the recent work of segmentation for improved facial attribute recognition [69].

As a follow-up to MOON [60], [72] introduced a method called "Selective Learning" to perform balancing of multi-label datasets during training of a deep neural network in order to address the label imbalance in CelebA. The authors introduce a new CNN – Attribute CNN (AttCNN), which has roughly 6 million parameters, compared to the 16 from MCNN [58]. In MOON, the labels were balanced by considering an overall dataset source distribution for each attribute. In [72], the authors note that this does not fully address the problem as each batch that is used to train the CNN may be more or less balanced than the overall training set. The author's solution was to perform label balancing at the batch level. Every attribute in each batch was balanced according to a desired target distribution by sampling from the over-represented class and weighting the underrepresented class. The Selective Learning method produced comparable results to MOON on CelebA and LFWA. The authors also introduced a new evaluation dataset: the University of Maryland Attribute Evaluation Dataset (UMD-AED). UMD-AED consists of roughly 3,000 images sparsely labeled with facial attributes. Some sample images from UMD-AED are shown in figure 5. Each of the forty attributes in CelebA has fifty positive and fifty negative instances in UMD-AED, allowing for balanced testing of facial attribute recognition methods. Selective Learning and AttCNN significantly outperformed MOON on UMD-AED.

Most research in facial attribute recognition focused on unconstrained images. Hand et al. shifted the focus to video in [73] using weakly labeled video to train attribute prediction models. The authors labeled four frames in every video of YouTube Faces [74] – a video dataset collected for face verification – with the forty binary facial attributes from CelebA. They introduced several methods for utilizing weakly labeled frames to improve attribute prediction in video: Motion Attention and Temporal Coherence. Their motion attention mechanism focused attribute models on areas of motion in the video, reducing overfitting, and the temporal

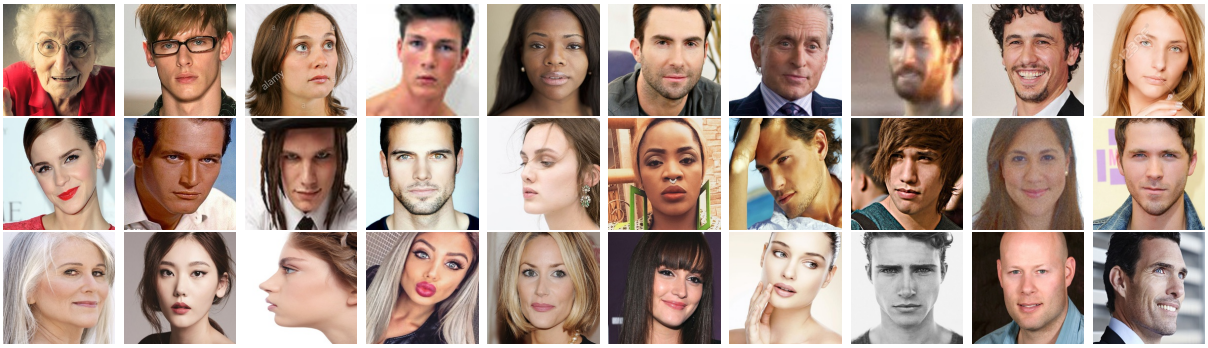

Figure 5: Sample images from the UMD-AED dataset [72].

coherence constraint encouraged nearby frames to have similar network responses, relying on the fact that nearby frames in a video will likely have similar – but perhaps not the same – attributes. Combining motion attention and temporal coherence, the authors were able to train a deep CNN on unlabeled video frames from YouTube Faces, outperforming traditional fine-tuning methods. [73] was the first, and only to date, attempt to utilize video for facial attribute recognition.

We present the average accuracy over all attributes in CelebA for all state-of-the-art methods in table 1. We can see that since the introduction of the dataset in [57], only a four percent gain in accuracy has been achieved on average. This emphasizes that there are many challenges that have yet to be addressed in the field of facial attribute recognition.

The field of facial attribute recognition is still a very young one, having been introduced just over a decade ago. Since it's introduction, huge strides have been made, with current systems capable of recognizing facial attributes in unconstrained images and video. There are many open research directions that will lead to significant improvements in the state-of-the-art in facial attribute prediction. With many applications relying heavily on the recognition of human-describable features, the field of facial attribute recognition will be of great interest for many years to come.

Method	**Accuracy**
Liu et al. (LNet+ANet) [57]	87.30%
Ehrlich et al. (MTM-RBM) [64]	87.00%
Wang et al. (Walk and Learn) [59]	88.00%
Rudd et al. (MOON) [60]	90.94%
Gunther et al. (AFFACT) [66]	91.97%
Hand et al. (MCNN-AUX) [58]	91.30%
Kalayeh et al. [69]	91.80%
Ding et al. [68]	91.23%
Hand et al. (AttCNN) [72]	91.05%
He et al. [71]	91.81%

Table 1: Average attribute classification accuracy across all forty attributes in CelebA for current state-of-the-art methods.

5 Open Problems

5.1 Addressing Bias

With the popularization of deep learning methods comes the inevitable question of bias. Most of the state-of-the-art deep learning models have tens of millions of parameters. Overfitting is a very common problem with such complex and deep architectures. For many problems, including attribute recognition, overfitting leads to learning dataset biases and imbalances. Due to the large number of parameters in deep learning models, the cause of bias is also difficult to locate.

Even outside of deep learning, bias can be an issue for models that draw correlations between attributes that frequently exist together. [4] proposes a method for image search and ranking for multi attribute queries. Their method is to find inter-dependencies of queried attributes with attributes not mentioned. The danger of this approach is that without diverse data, potentially inaccurate or even inappropriate results can be returned. Hand et al. found this to be true in [58], when utilizing implicit and explicit attribute relationships to improve prediction. They found that many relationships were not indicative of the real world, but rather were overfit to the CelebA dataset (e.g. *heavy makeup* and *arched eyebrows*).

Most of the methods for attribute prediction presented here simply consider accuracy as an evaluation metric. When we are dealing with severe class imbalance, precision and recall are more appropriate measures of good performance. Additionally, it is essential to account for age, race, and gender in attribute recognition research in order to truly learn a representation for an attribute across different groups. Ultimately, most bias stems from a lack of representation in the data, or bias in the labeling of the data. Addressing these issues will improve attribute recognition research.

5.2 Noisy Data

There is a recent plateau in facial attribute recognition performance, which could be due to poor labeling of data. Many research groups rely on Amazon's Mechanical Turk, as well as other paid services to provide manual labels on large data sets [2, 3, 23, 56]. While crowd-sourcing such tasks can be very useful and result in large quantities of reasonably labeled data, there are some tasks which may be consistently labeled incorrectly – such as subjective tasks (e.g. *hair color*, *attractive*). Additionally, certain groups of people may find it challenging to label certain tasks. For example, men can have difficulty identifying makeup and lipstick in images of faces. Mislabeled data is a very difficult challenge to overcome for most learning methods.

Mislabeled data is also known as noise. An open problem in facial attribute recognition – and many problems in computer vision and machine learning – is that of noise identification and removal. The difficulty here lies in removing noise while maintaining outliers – challenging, but correctly labeled, samples. It is essential to keep outliers in the dataset, as they represent samples that are outside of the average for a particular problem.

Identifying noise versus outliers is an especially challenging problem in the domain of facial attribute recognition. Many attributes are extremely subjective, and exist on a gradient. Some examples of subjective attributes include: *attractive*, *5 o'clock shadow*, *arched eyebrows*, *young*, etc. Due to this subjectivity, there is a need to consider the problem of facial attribute recognition as one of regression, or real-value prediction. In this way, attributes can exist on a scale, and subjectivity will not pose as much of a problem as it does in the binary case. Parikh et al. introduced the concept of relative attributes in [55], and it has yet to be built upon. This is likely due to the significant amount of time that would be required to properly label a large dataset with relative attributes. However, it is undeniable that attributes as real-valued variables results in a more natural, rich description of faces. Relative attributes will help to alleviate the problem of poorly labeled data as well, since attributes will exist on a scale rather than in strict categories.

5.3 Attributes in Video

A relatively untouched and important problem in attribute recognition is attribute prediction in videos. There is a wealth of information that comes from video. Take for example the problem of classifying facial attributes depending on pose. In a video, various frames will contain the same attributes with a large variety of poses, expressions, angles, etc. Models trained on video data could be very robust in their ability to identify facial attributes. In addition to the large amount of data that can be used in videos to solve current issues in attribute recognition, there are also new categories of attributes to be identified.

One such category is static versus dynamic attributes. Static attributes are the visual attributes that do not change with pose, expression, lighting, etc. These are features such as hair color, gender, age, facial hair, etc. Conversely, dynamic attributes include features such as smiling, narrow eyes, mouth open, etc. Accurate recognition of these attributes requires pose prediction, 3D processing, and semantic understanding of the data.

There are of course challenges associated with video processing. The main challenges of attribute recognition in videos, compared to images, are low resolution, poor lighting, severe poses, motion blur, and varying frame rate. Low resolution is a problem simply because lower quality images contain less data. Videos commonly contain content with bad lighting, which can cause unwanted effects on a person's face leading to poor attribute recognition performance. Video also suffers from severe poses and motion blur. Recognizing visual features in still images usually only takes into account images of people in relatively normal positions or with common expressions. Videos contain many blurred frames, especially as frame rate increases.

References

[1] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *European Conference on Computer Vision*, pp. 340–353, Springer, 2008.

[2] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *International Conference on Computer Vision*, pp. 365–372, IEEE, 2009.

[3] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

[4] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *CVPR 2011*, pp. 801–808, June 2011.

[5] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?," in *Biometric Technology for Human Identification*, vol. 5404, pp. 561–573, International Society for Optics and Photonics, 2004.

[6] C. B. Ng, Y. H. Tay, and B.-M. Goi, "Recognizing human gender in computer vision: a survey," in *Pacific Rim International Conference on Artificial Intelligence*, pp. 335–346, Springer, 2012.

[7] N. Ramanathan, R. Chellappa, and S. Biswas, "Computational methods for modeling facial aging: A survey," *Journal of Visual Languages & Computing*, vol. 20, no. 3, pp. 131–144, 2009.

[8] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.

[9] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, "Sexnet: A neural network identifies sex from human faces.," in *NIPS*, vol. 1, p. 2, 1990.

[10] G. W. Cottrell and J. Metcalfe, "Empath: Face, emotion, and gender recognition using holons," in *Advances in neural information processing systems*, pp. 564–571, 1991.

[11] W. Gao and H. Ai, "Face gender classification on consumer images in a multiethnic environment," in *International Conference on Biometrics*, pp. 169–178, Springer, 2009.

[12] C. BenAbdelkader and P. Griffin, "A local region-based approach to gender classi. cation from face images," in *Computer vision and pattern recognition-workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, p. 52, IEEE, 2005.

[13] G. Guo, C. R. Dyer, Y. Fu, and T. S. Huang, "Is gender recognition affected by age," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pp. 2032–2039, IEEE, 2009.

[14] Y. H. Kwon and N. da Vitoria Lobo, "Age classification from facial images," *Computer vision and image understanding*, vol. 74, no. 1, pp. 1–21, 1999.

[15] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442–455, 2002.

[16] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proceedings of the 14th ACM international conference on Multimedia*, pp. 307–316, ACM, 2006.

[17] N. Ramanathan and R. Chellappa, "Face verification across age progression," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3349–3361, 2006.

[18] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1644, 2014.

[19] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1778–1785, IEEE, 2009.

[20] T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," vol. 6311, pp. 663–676, 12 2010.

[21] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3337–3344, IEEE, 2011.

[22] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. S. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," *CoRR*, vol. abs/1611.05377, 2016.

[23] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2933–2940, IEEE, 2012.

[24] C. Newton, "Facebook begins using artificial intelligence to describe photos to blind users," April 2016. [Online; posted 5-April-2016].

[25] P. Samangouei, V. M. Patel, and R. Chellappa, "Attribute-based continuous user authentication on mobile devices," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pp. 1–8, IEEE, 2015.

[26] P. Samangouei and R. Chellappa, "Convolutional neural networks for attribute-based active authentication on mobile devices," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pp. 1–8, IEEE, 2016.

[27] P. Samangouei, E. Hand, V. M. Patel, and R. Chellappa, "Active authentication using facial attributes," *Mobile Biometrics*, vol. 3, p. 131, 2017.

[28] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," in *Trends and Topics in Computer Vision* (K. N. Kutulakos, ed.), (Berlin, Heidelberg), pp. 1–14, Springer Berlin Heidelberg, 2012.

[29] N. Jojic and Y. Caspi, "Capturing image structure with probabilistic index maps," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, June 2004.

[30] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1265–1278, Aug 2005.

[31] Y. Liu, Y. Tsin, and W.-C. Lin, "The promise and perils of near-regular texture," *Int. J. Comput. Vision*, vol. 62, pp. 145–159, Apr. 2005.

[32] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), pp. 433–440, Curran Associates, Inc., 2008.

[33] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *CVPR 2011*, pp. 1681–1688, June 2011.

[34] X. Wang and Q. Ji, "A unified probabilistic approach modeling relationships between attributes and objects," in *2013 IEEE International Conference on Computer Vision*, pp. 2120–2127, Dec 2013.

[35] Y. Sun, L. Bo, and D. Fox, "Attribute based object identification," in *2013 IEEE International Conference on Robotics and Automation*, pp. 2096–2103, May 2013.

[36] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed, "Understanding objects in detail with fine-grained attributes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3622–3629, June 2014.

[37] X. Wang and Q. Ji, "Object recognition with hidden attributes," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, p. 3498–3504, AAAI Press, 2016.

[38] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, Oct 2007.

[39] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *2011 International Conference on Computer Vision*, pp. 1331–1338, Nov 2011.

[40] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," pp. 652–659, 06 2013.

[41] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute regularization based human action recognition," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 1600–1609, Oct 2013.

[42] D. Tahmoush, "Applying action attribute class validation to improve human activity recognition," 2015.

[43] X. Cai, W. Zhou, and H. Li, "Attribute mining for scalable 3d human action recognition," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015.

[44] K. Chen, G. Ding, and J. Han, "Attribute-based supervised deep learning model for action recognition," *Frontiers of Computer Science*, vol. abs/1707.09468, pp. 219–229, Apr 2017.

[45] F. S. Khan, J. van de Weijer, R. M. Anwer, A. D. Bagdanov, M. Felsberg, and J. Laaksonen, "Scale coding bag of deep features for human attribute and action recognition," *Machine Vision and Applications*, vol. 29, pp. 55–71, Jan 2018.

[46] M. H. Zaki and T. Sayed, "Using automated walking gait analysis for the identification of pedestrian attributes," *Transportation Research Part C: Emerging Technologies*, vol. 48, pp. 16 – 36, 2014.

[47] Y. DENG, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pp. 789–792, 2014.

[48] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Learning to recognize pedestrian attribute," *CoRR*, vol. abs/1501.00901, 2015.

[49] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," 2015.

[50] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang, "Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization," *CoRR*, vol. abs/1611.05603, pp. 224 – 229, 2016.

[51] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3177–3183, International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[52] Z. Chen, A. Li, and Y. Wang, "Video-based pedestrian attribute recognition," *CoRR*, vol. abs/1901.05742, 2019.

[53] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Applications of Computer Vision (WACV), 2009 Workshop on*, pp. 1–8, IEEE, 2009.

[54] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[55] D. Parikh and K. Grauman, "Relative attributes," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 503–510, IEEE, 2011.

[56] T. Berg and P. N. Belhumeur, "Tom-vs-pete classifiers and identity-preserving alignment for face verification.," in *BMVC*, vol. 2, p. 7, Citeseer, 2012.

[57] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

[58] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification.," in *AAAI*, pp. 4068–4074, 2017.

[59] J. Wang, Y. Cheng, and R. Schmidt Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2295–2304, 2016.

[60] E. M. Rudd, M. Günther, and T. E. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision*, pp. 19–35, Springer, 2016.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, Ieee, 2009.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[64] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer, "Facial attributes classification using multi-task representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 47–55, 2016.

[65] A. Rozsa, M. Günther, E. M. Rudd, and T. E. Boult, "Are facial attributes adversarially robust?," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 3121–3127, IEEE, 2016.

[66] M. Günther, A. Rozsa, and T. E. Boult, "Affact: Alignment-free facial attribute classification technique," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 90–99, IEEE, 2017.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[68] H. Ding, H. Zhou, S. K. Zhou, and R. Chellappa, "A deep cascade network for unaligned face attribute classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[69] M. M. Kalayeh, B. Gong, and M. Shah, "Improving facial attribute prediction using semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 4227–4235, IEEE, 2017.

[70] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European conference on computer vision*, pp. 679–692, Springer, 2012.

[71] K. He, Y. Fu, W. Zhang, C. Wang, Y.-G. Jiang, F. Huang, and X. Xue, "Harnessing synthesized abstraction images to improve facial attribute recognition," in *IJCAI*, pp. 733–740, 2018.

[72] E. M. Hand, C. Castillo, and R. Chellappa, "Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction," AAAI, 2018.

[73] E. M. Hand, C. D. Castillo, and R. Chellappa, "Predicting facial attributes in video using temporal coherence and motion-attention," in *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, pp. 84–92, IEEE, 2018.

[74] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 529–534, IEEE, 2011.