Chapter

Improving Face Recognition Using Artistic Interpretations of Prominent Features: Leveraging Caricatures in Modern Surveillance Systems

Sara R. Davis and Emily M. Hand

Abstract

Advances in computer vision have been primarily motivated by a better understanding of how humans perceive and codify faces. Broadly speaking, progress made in the fields of face recognition and identification has been strongly influenced by the biological mechanisms identified by research in the field of cognitive psychology. Research in cognitive psychology has long acknowledged that human face recognition and identification rely heavily on prominent features and that caricatures are capable of modeling prominent features in a multitude of ways. The field of computer science has done little to no research in the area of application of prominent features to recognition systems. This chapter discusses existing caricature research in cognitive psychology and computer vision, current issues with the practical application of caricatures to face recognition in computer vision, and how caricatures can be used to improve existing surveillance systems.

Keywords: face recognition, caricatures, datasets

1. Introduction

The word "caricature" comes from Italian for "to exaggerate" [1, 2]. As such, caricatures are artistic renderings of a human face that exaggerate prominent features while still maintaining their resemblance to the original, veridical face. Veridical is defined as the ground truth face [3]. An example can be seen in **Figure 1**. Since the 1590s, caricatures have been considered a humorous art form, meant to either entertain or humiliate, depending on context. In the United States, caricatures rose in popularity following the American Civil War, giving rise to our modern-day interpretation of the art form. At first, these images were used to mock political leaders in an effort to humorously instill political ideology [2].

Today, many people consider caricatures to be "fun" drawings. However, the fields of psychology and neuroscience have recognized the potential application of

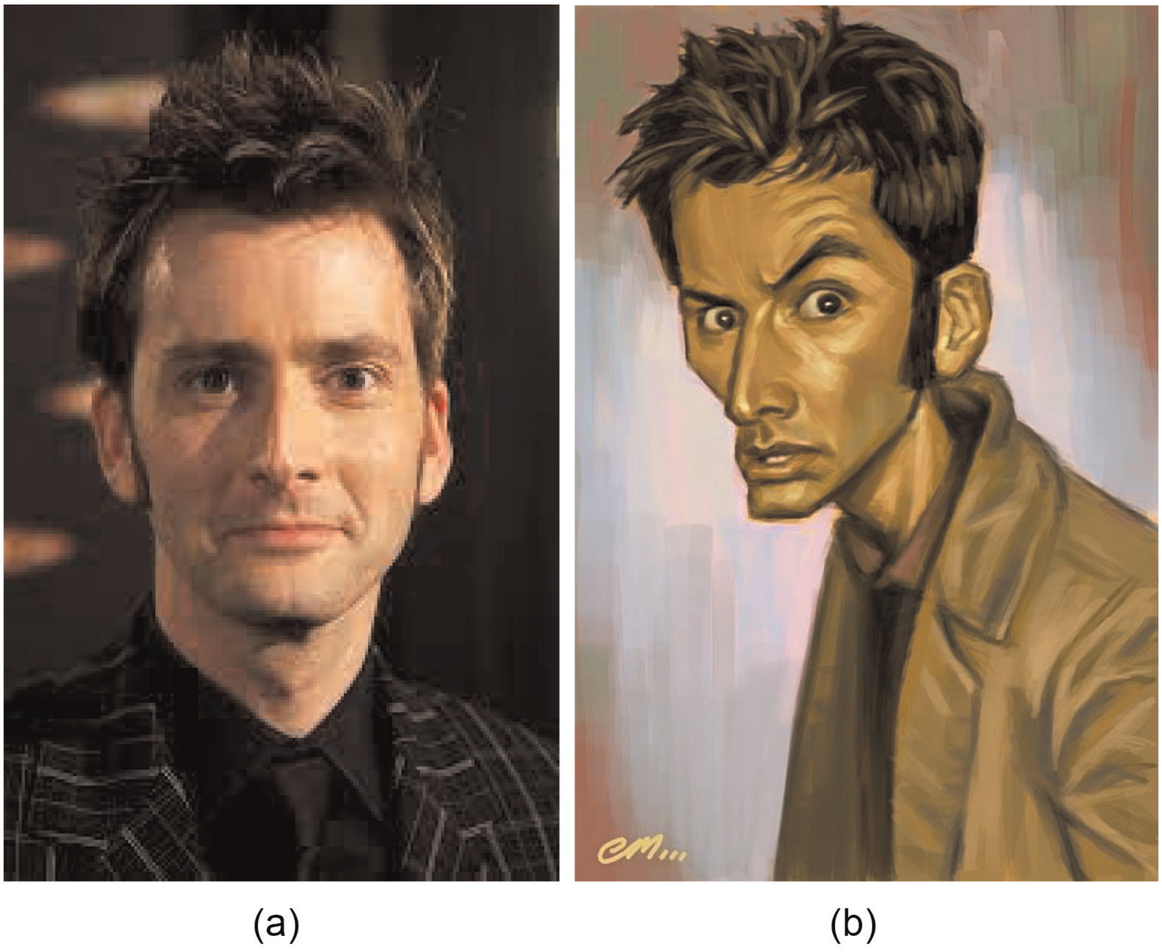

Figure 1.
A veridical image (photo) and caricature of David Tennant. (a) veridical image of David Tennant, (b) caricature image of David Tennant.

caricatures for improving automated face verification and identification systems in recent years. Not only can caricatures be identified more often and faster than veridical images by humans [4], but they can improve the accuracy of low-resolution face verification in elderly populations [3]. Studies have also shown that introducing faces using caricatures, rather than veridical images, results in higher recognition and verification overall [5–8]. These works also show that facial exaggeration in a caricature past a certain point can actually decrease recognition performance over veridical images [9]. Each of these factors makes caricatures the ideal model for exploring how humans perceive and codify faces under nonideal conditions. In this chapter, we discuss how machine learning can leverage this biologically inspired recognition mechanism to improve surveillance systems. We discuss data collection methods and possible system architectures and show that caricatures can be used to train more robust face recognition systems.

2. Caricatures in cognitive psychology

Past advances in automated face recognition and verification have been driven by face perception research advancement in the field of cognitive psychology [1]. Human facial recognition is not negatively impacted by variation in pose, lighting, or resolution, unlike automatic systems [10–12]. Additionally, research has shown that human facial recognition of familiar faces is consistently better than automated systems [4]. Thus, we propose using caricature research from the field of cognitive psychology to construct surveillance systems that are robust to changes in angle, lighting, and accidental exaggeration.

The study of facial recognition in cognitive psychology has two different schools of thought: holistic and nonholistic. Each has research to support it, though we argue that the holistic approach has the greatest probability of being applied to automated facial recognition systems, and this is supported by past research [14]. Holistic face recognition research contends that faces are stored in human memory using the relationship between all features in a face, while nonholistic research argues that a single prominent facial attribute is enough to perform face recognition [15]. The difference between the two can be thought of as holistic valuing the sum of the parts of the face to create an overall model, while nonholistic values single prominent features. Because holistic face recognition relies on facial feature relationships, the relationships can be divided into two categories: featural and configural. Featural facial feature attributes look at the general structure of the facial feature, while configural focuses on the relative placement and distance between features. An example of the difference can be seen in **Figure 2**, and is taken from ref. [13].

A literature survey [16] found that most facial recognition performed by humans appears to use a holistic approach, considering each attribute in relation to the other available facial attributes. Other work has looked at face recognition in holistic and nonholistic settings to compare the possible way humans represent faces in memory [15]. They found that participants could more accurately identify unique facial attributes when the attribute was presented with the whole face as context, rather than the isolated facial attribute. For example, when participants viewed a nose on its own, they were less likely to identify that facial feature as prominent. However, if the participant viewed that same nose on the face that it belonged to, they had an easier

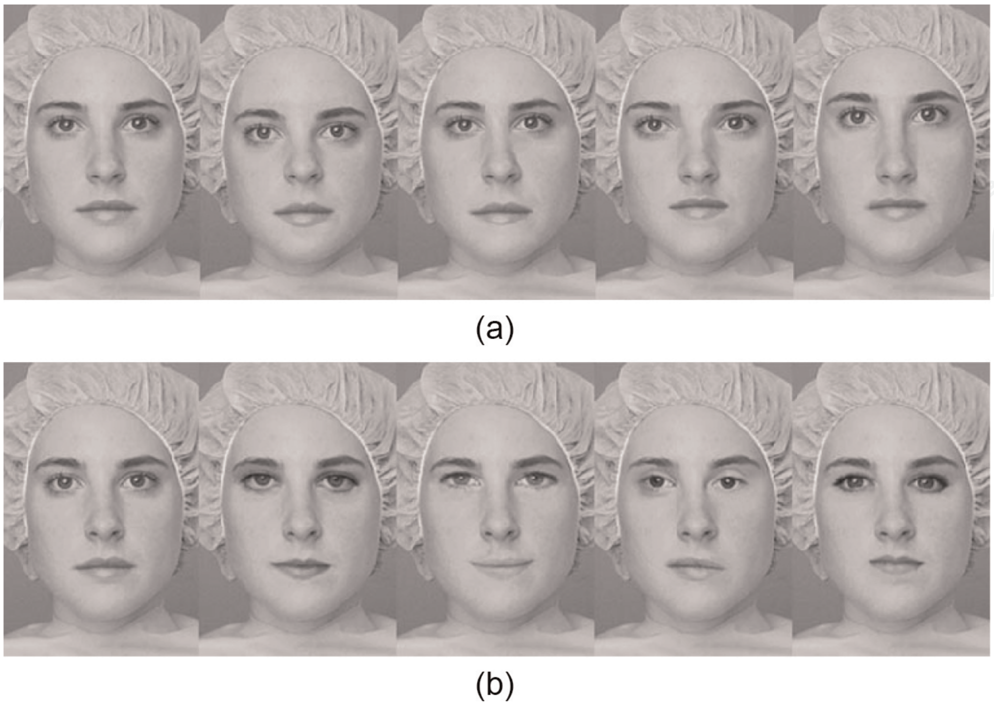

Figure 2.
An illustration of the difference between configural (top) and featural (bottom) feature relationships taken from ref. [13].

time identifying the nose as prominent. This supports the theory that humans represent faces holistically, and that they need to compare facial features to each other in order to determine which features are most prominent for that identity. Additionally, they found that if faces were inverted, recognition accuracy drastically decreased whether the part was presented with the face as context or on its own. This implies that representations in memory of human faces are strongly correlated with orientation and configural properties of facial attributes.

In order to test the importance of configural properties, another study [14] manipulated the distance between eyes, and participants were tasked with performing face recognition using distance-altered eyes. Each participant was presented with the altered eyes in (1) the original face, (2) a slight alteration of the original face, and (3) isolation. The authors found that face recognition was best using the original face, followed by the slightly altered original face, and the worst performance was when the distance-altered eyes were presented in isolation. Related to prominent feature recognition, the authors also found the configuration of the original face with the altered-eye distance resulted in lower accuracy rates in recognition of the nose and mouth features, even though the nose and mouth features had not been altered. This further supports the holistic school of thought, that is, humans learn faces holistically, and understanding the face depends both on featural and configural information.

Research has shown that participants are able to more quickly identify faces using simple line drawings of caricatured faces as compared to veridical faces [7]. This same study also suggested that caricatures can be used to better understand how humans represent faces using prominent facial features. Specifically, they found that faces and caricatures are stored in memory using the deviation of a prominent feature from the normal presentation of that feature. The authors call this norm-based coding. Another study found that the improvement in face verification rates is not specific to caricatures but is most likely caused by memory retention of facial features that deviate from the average [3]. This implies that human face encoding is closely affiliated with prominent facial features. In another study, Rhodes performed a series of experiments that tested the possible relationship between configural-based coding and norm-based coding using caricatures [8]. They found that the configural-based coding that is necessary for veridical face recognition is not necessary for caricature recognition. This implies that (1) caricature/veridical face pair recognition relies on a memory mechanism that is independent of the face/facial feature pair recognition, (2) caricature/veridical face pair recognition relies on norm-based coding, and (3) the approach to performing veridical/veridical, caricature/caricature, and caricature/veridical image recognition should be different due to the difference in coding methods.

Research has found that caricatures are accurately identified more quickly and more often than veridical images, with caricatures of familiar faces being recognized with the best accuracy [6]. Additional studies have found that caricatures of unfamiliar faces also improved verification rates by approximately 30%. Furthermore, above a certain rate of exaggeration, caricature verification is actually hindered; in other words, caricatures need to have a reasonable resemblance to the original face [9]. Past work also found that using caricatures led to better recognition of unfamiliar faces across the entire human lifespan, that it improved low-resolution face verification in older adults, and that face verification of other races also improved [3]. Each of these studies indicates that there is a link between human facial recognition, prominent features, and the general configuration of facial features.

Cognitive psychology defines facial features as either internal, such as the eyes, nose, or mouth, or external, such as hair or chin [17]. An example can be seen in

Figure 3. Past work has shown that familiar faces are more accurately identified if internal features were used, rather than external [17]. Feature type did not have an effect on identifying unfamiliar faces. The authors argued that the manner in which faces are modeled and stored in memory is different for familiar and unfamiliar faces, and thus, their treatment in facial recognition should be different. However, another study [18] found that internal and external features both activate similar face-selective regions of the brain, though internal features result in a greater response. Both of these works [17, 18] found that internal features were more important for familiar faces. Additionally, the study found that altering just the external features resulted in a decrease in identification accuracy, regardless of whether the face was familiar or unfamiliar. This indicates that the internal and external features interact with each other to create a holistic representation in memory and that internal and external features are likely of similar importance in machine learning applications to understanding faces. This is of particular importance to the application of caricatures to surveillance system construction because most face datasets are constructed of famous individuals; however, fame is not consistent across countries and cultures. For example, Fan Bingbing is a famous actress in China, but she is not nearly as well known in the United States. Since face recognition datasets are often comprised of a variety of celebrities from around the world, approaches to automated face recognition should not assume familiarity with the subjects in the dataset in order to make the approach relatable to human face recognition processes. Put simply, since the normal participant in a cognitive psychology study is unlikely to be familiar with every individual in a face recognition dataset, any automated system built for facial recognition should not assume familiarity with the subject.

Experiments surrounding how faces are learned over time have also been conducted. Research has shown that after a single view of a face, recognition from a different viewpoint was better using internal features rather than external features [19]. Additionally, the study found that after repeated exposure to a face, removing external features that change with high frequency, such as hair, resulted in better identification when a face was viewed from a different viewpoint. These results suggest that providing too much inconstant information to an automated face recognition system can result in a reduction of recognition capability. This means that the

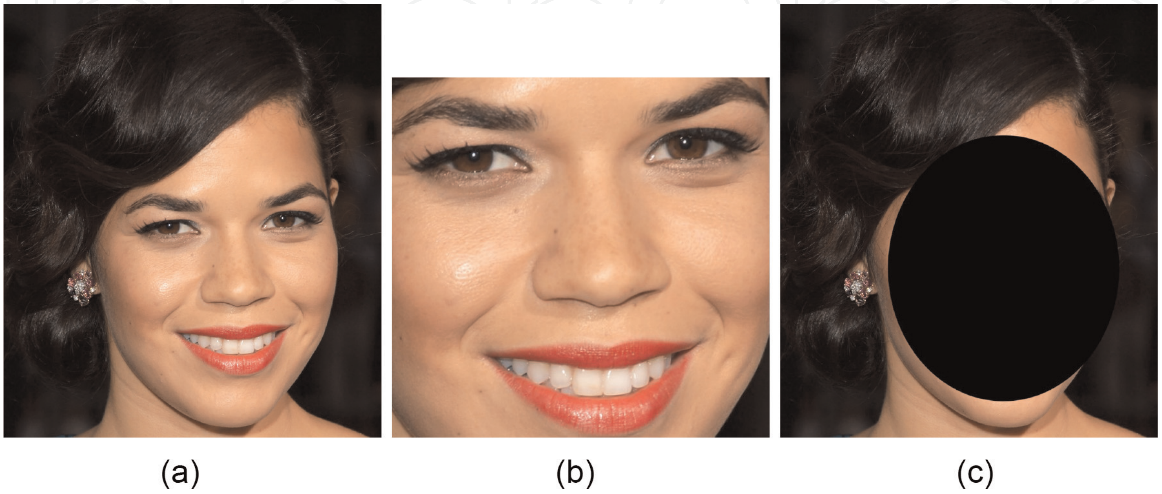

Figure 3.
Examples of the difference between internal and external features. The original veridical image is shown in (a), internal features in (b), and external features in (c). (a) cropped full face, (b) internal facial features, (c) external facial features.

method in which images of the face are cropped, image background, lighting, and even changes in hairstyle and makeup all likely have a significant effect on automated recognition. Other work [20] found that repeated exposure to the same face within different contexts—variations in pose and lighting—resulted in better facial recognition than if subjects simply viewed the same image of the face over and over again. This indicates that exposure to unique images has an effect on how faces are learned and retained in memory, which means that in order for an accurate facial representation to be built, automated systems need to utilize a variety of images for each identity and repeated exposure.

3. Caricatures in computer science

Today's automated surveillance systems rely on identity matching in some latent space [21]. We argue that the use of caricatures would better allow these systems to describe faces and prominent features, thus allowing for greater variability in pose, lighting, etc. Research shows that current automated face recognition and verification systems perform better than human recognition and verification. However, that same research suggests that automated systems only perform better on carefully curated datasets [4, 21, 22]. In other words, automated systems cannot handle images that are not taken under ideal lighting, pose, and resolution conditions. Humans, on the other hand, are capable of recognizing faces under nonideal conditions. Many surveillance systems require face alignment in order to achieve state-of-the-art results [23, 24]. This means that they work best on frontal facing images [25–27], but humans do not need a frontal facing image to perform recognition. In fact, many caricatures exaggerate face angles, and humans are still able to perform recognition with them. To improve existing automated systems, we discuss using caricatures to construct surveillance systems that are robust to changes in angle, lighting, and accidental exaggeration, and the existing research in computer science that has already leveraged these images.

Past work in automated face recognition in computer science belongs to one of two system types: traditional machine learning or deep learning. Traditional machine learning techniques to perform face recognition include using deep belief networks [28], metric learning [29, 30], and dimensionality reduction via principal component analysis (PCA) and/or linear discriminant analysis (LDA) [31]. With the rise of better hardware and GPU cycling, deep learning has become the standard approach. Typical approaches use convolutional neural networks (CNNs) [32] or autoencoders [33] and may be combined with more traditional methods to increase performance [34]. Most methods try to increase interclass margins while decreasing intraclass margins, so that distinct class clusters are created in high-dimensional feature space [35]. The recognition task is complicated by pose variance, lighting changes, and changes in an individual's appearance [36, 37], as mentioned before. Nguyen *et al.* [38] proposed a representation learning method to overcome the issues caused by recognition under nonideal conditions. They found that the cosine similarity between images can be used to improve face recognition under nonideal conditions. Past research has also shown that soft biometrics, such as the use of prominent facial features or hairstyle, can be used to improve facial recognition technology [31, 39].

Research in the area of feature learning and architecture found that facial recognition methods can be improved by utilizing ResNet CNN architectures, rather than VGG [40]. The same study discusses methods of face detection, facial alignment, and how to determine what ResNet structure is best for a selected dataset, with significant performance improvements on standardized datasets with wide variance. Unfortunately, the vast majority of work in the area of face recognition is dataset-dependent, and using proposed methods on other datasets results in an unexpected behavior [4, 21, 22], which we discuss in Section 4.

Though cognitive psychology has shown that the use of caricatures improves human recognition, work in computer science using caricatures for face recognition is rather limited. As deep learning representations for face recognition have become more accurate, face generation systems have been proposed, typically using a generative adversarial network (GAN) [42–44]. GANs are a class of deep generative models [10]. To backpropagate loss through the GAN, the input to the system must be differentiable [45]. While using a GAN can be quite successful, it can also lead to mode collapse [46] and vanishing gradient behavior [47]. Additionally, while the initial GAN results appear promising at first glance, the authors typically only report their best results and neglect to show that the vast majority of generated images are nonsensical [48]. An example of an image that is not representative of its target identity is shown in **Figure 4**. One approach to enforcing differentiability, so that better images are generated, is to use a kernel-based moment-matching scheme over a reproducing kernel Hilbert space (RKHS) [49]. This forces the real and generated images to have matched moments in the latent-feature space, and helps combat mode collapse while encouraging images that are descriptive and varied [49].

Despite these limitations, recent work generates a caricature from veridical images using GANs [41, 50–52], but does not try to understand or utilize caricatures to improve verification or recognition. Some work attempts to exploit caricatures to improve verification. However, the dataset is small and does not use modern deep learning methods [53]. Work in verification and recognition improvement using caricatures, rather than caricature generation, is relatively new and not well explored. Past work in the field [54] introduced a method to extract facial attribute features from photos but required manual labeling of facial attribute features on caricatures, which is time-consuming. Furthermore, the study computed feature importance

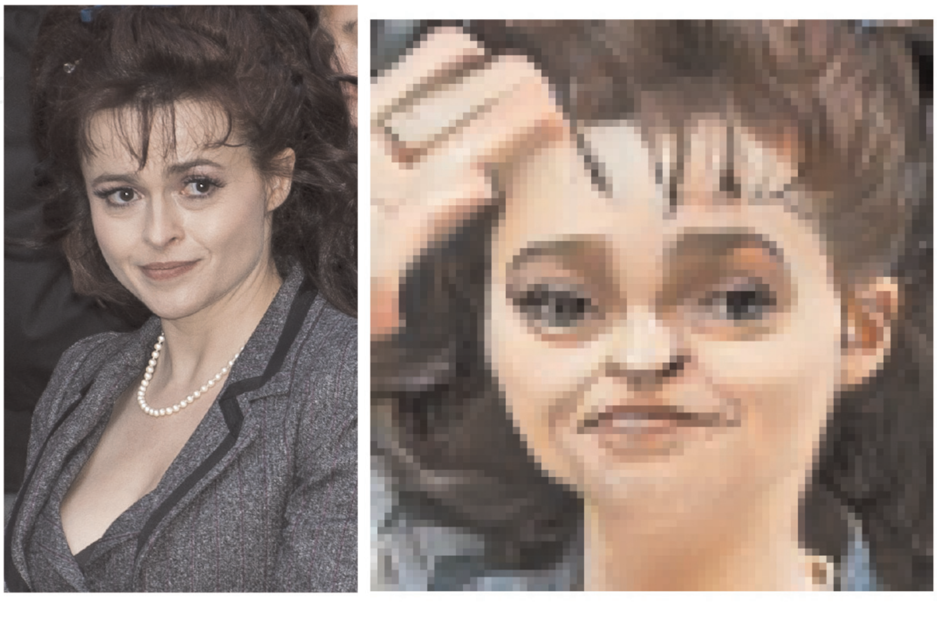

Figure 4.
An example of a poorly GAN-generated caricature produced by WarpGan [41], one of the current state-of-the-art caricature generation systems. Note that this caricature (right) is not identifiable as Helena Bonham Carter (left).

using genetic algorithms, which are extremely slow compared to deep learning.
In the field of cognitive psychology, [55] showed that facial recognition improves when PCA is applied to all of an identity's images and then averaged. This indicates that the (1) human memory holds the average of a person's face after multiple exposures; and (2) PCA is one method that might be applied when creating an automated face recognition system. The most comprehensive published work in automated caricature verification is WebCaricature [56], which provides an end-to-end framework for face verification and identification using caricatures, though we discuss in Section 4 the use of flawed data in their study.

Though caricatures have not been widely used in automated face recognition systems, facial attribute recognition is a well-researched task in the field [57–62]. Recent research has focused on performing attribute recognition, and introducing new datasets and deep learning frameworks [63–66]. The current state-of-the-art facial attribute prediction methods include "Walk and Learn," which pretrains a network on face verification data and then fine-tunes it on attribute recognition [65], as opposed to pretraining on object data [67]. Work has also shown that dataset imbalance, which we discuss in Section 4, can be ameliorated by using a multi-task network with the mixed objective loss [66]. Attribute relationships have also been used within deep neural networks to improve prediction [64]. Unfortunately, current work is focused on facial attribute identification and prediction. To date, there has not been any work in using *prominent* facial features to perform recognition, despite the fact that research in cognitive psychology has shown that human recognition relies on prominent facial features (Section 2). Thus, we argue that existing surveillance systems could be improved by creating systems capable of using prominent facial features so that models are better trained to focus on the same features that humans use to identify faces.

4. Data collection methods

From the perspective of this chapter, we care about the application of caricature data to improve surveillance systems. Generally speaking, surveillance systems have at least one frame with an un-exaggerated snapshot of identity, similar to a photo. Therefore, caricature research typically constructs datasets by collating a caricature set and a matching identity real photo set. Past research has focused on curating datasets with as many images as possible using web scraping [49, 50, 56]. For real images, there are existing methods to remove duplicate images and images of low quality. Unfortunately the same is not true for caricatures. Since the field is relatively new, systems have not yet been built to recognize image duplicates, and even if one was constructed, it would not handle the issue of under-exaggeration or representation fidelity. Of the previously cited works, none ensure that the image is of acceptable quality, images in the caricature group are actually caricatures and not some other form of art, and that images are actually of the target identity [49, 50, 56]. In some cases, datasets inaccurately incorporate character representations of an identity, rather than the actual identity; for example, the WebCaricature dataset [56] inaccurately labels general images of Harry Potter, a cultural icon, as Harry Potter rather than gathering images of Daniel Radcliffe. This introduces a high degree of variability, as the character "Harry Potter" is not always depicted as Daniel Radcliffe, just as Daniel Radcliffe is not always seen portraying Harry Potter (**Figure 5**). Each of these conditions is critical to creating a dataset that creates an accurate caricature

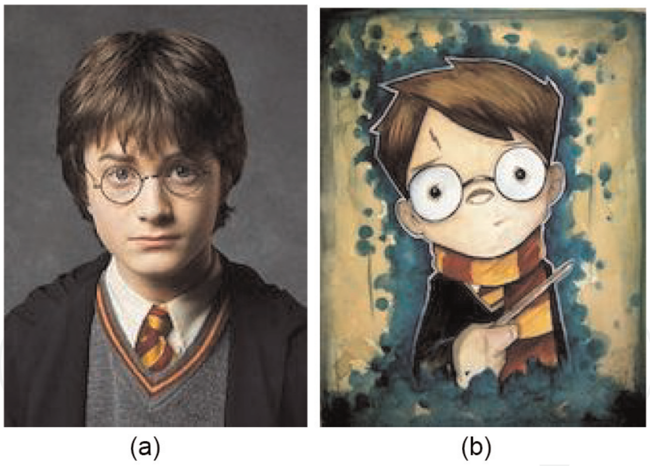

Figure 5.
An instance of the character representation of a fictional character (Harry Potter) not matching the affiliated actor (Daniel Radcliffe). (a) veridical image of Daniel Radcliffe. (b) pop culture representation of Harry Potter.

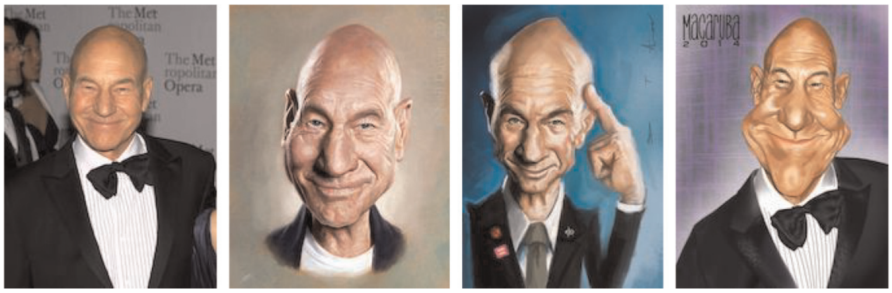

Figure 6.
Examples of variation in caricature representation of the same person (Patrick Stewart) taken from [56]. The left-most image is the photo (veridical face) and subsequent images are caricatures. Note that while there is wide variation in the representation of the veridical image, the identity of each of the caricatures is still obvious.

representation of supplied identities. In a deep learning system, data quality directly affects train and test performance [67, 68]. This means that ensuring that data is representative of each identity is exceedingly important; otherwise, the recognition task becomes unnecessarily more difficult and possibly more biased.

The caricature recognition task is additionally complicated by the fact that caricatures are artistic renderings. This means that artists may choose to exaggerate some facial features over others (**Figure 6**). Furthermore, as an artistic medium, classifying an image as a "caricature" as opposed to some other art form like painting can be difficult, as shown in **Figure 7**. These two issues highlight an important core issue in data collection for caricature trained systems: (1) caricatures should be caricatures and not some other art form, and (2) the caricature should actually resemble the target identity. Since we propose using caricatures to construct surveillance systems that are robust to changes in angle, lighting, and accidental exaggeration, caricatures should still be fairly exaggerated, and using caricatures with variation in style and degree of exaggeration will improve a surveillance system's ability to accommodate large exaggerations that typically hurt performance in the existing state-of-the-art systems. Because caricature drawing is an artistic medium, the same person can be portrayed with a wide array of variations in facial features that are over or under-exaggerated. An example of this can be seen in **Figure 6**.

Unfortunately, the construction of a good caricature dataset is slow and labor intensive. Each caricature needs to be assessed for quality, and currently, the methods

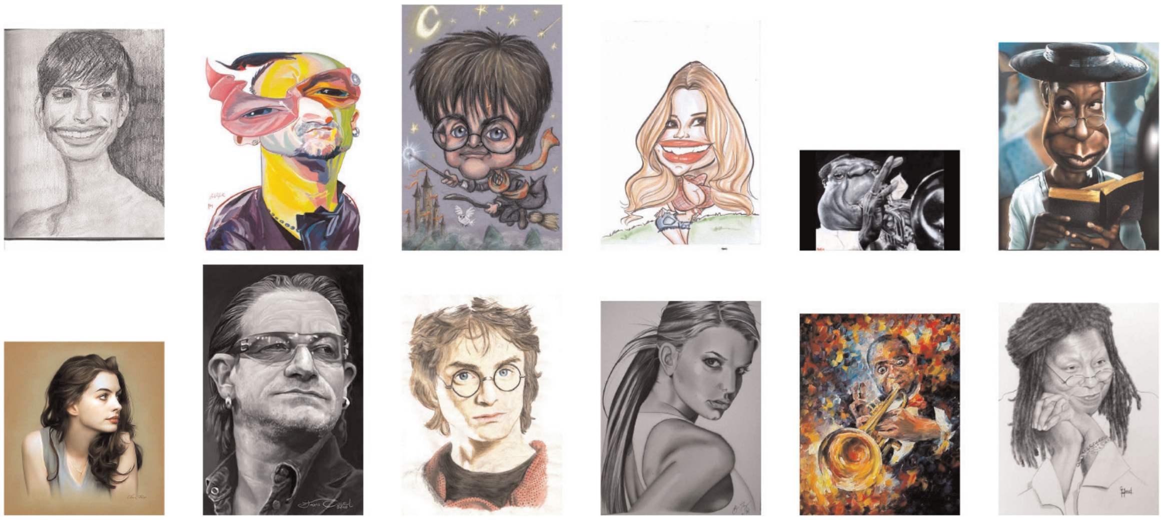

Figure 7.
Images from WebCaricature [56] that are not of acceptable quality to be included in a computer vision dataset. The first row contains images where the identity is not immediately obvious without knowing who the person is. The second row contains images where the image is not a caricature, but rather a painting, drawing, or cartoon.

to do that are manual. That means checking every caricature of resemblance to the target identity and art style. Additionally, many state-of-the-art surveillance systems are reliant on facial landmarks, which can already be inaccurate in normal photos [69, 70]; this inaccuracy is exacerbated in caricatures, particularly caricatures with a high degree of exaggeration across internal facial features.

It is also critical that datasets are constructed in a way that limits bias as much as possible, as any dataset bias will be trained into a surveillance system. For example, in 2015, Google released an image labeler that had been poorly trained, so that it mislabeled human faces as gorillas [71]. Company representatives later acknowledged that this example of racism was caused by data the system was trained on.

Since many caricatures reflect cultural norms by trying to exaggerate consistent prominent features, racist interpretations are more likely. For example, because "Bruce Lee" is an Asian-American man, many racist caricatures overly exaggerate the degree of eye closure and inferred mouth pout. An exaggeration is considered racist when it does nothing to improve the machine representation of the target identity while enforcing stereotypes that exist in popular culture (**Figure 8**). Additionally, since most surveillance systems see a variety of genders, races, and angles, it is important that the dataset used to train the surveillance system is as representative as possible. The ACLU has pointed out that existing surveillance systems are more prone to misidentifying women and people of color [72]. Additionally, many police departments use mugshots to create their databases, which perpetuates the issue of racism since people of color are up to four times as likely to be arrested for the same crime perpetrated by Caucasian suspects [72]. This means that most police surveillance systems use data sets that are overwhelmingly comprised of citizens of color, making it easier to identify them than Caucasian citizens [72]. Additional work by Buolamwini and Gebru in 2018 found that datasets curated by sources other than law enforcement are composed of overwhelmingly white male subjects. This data imbalance leads to high accuracy in identifying white male subjects, but high rates of misidentification of women and people of color, and especially women of color [73]. The US Department of Commerce later reported findings consistent with Buolamwini and Gebru [74]. In a surveillance system, particularly surveillance systems used by law enforcement,

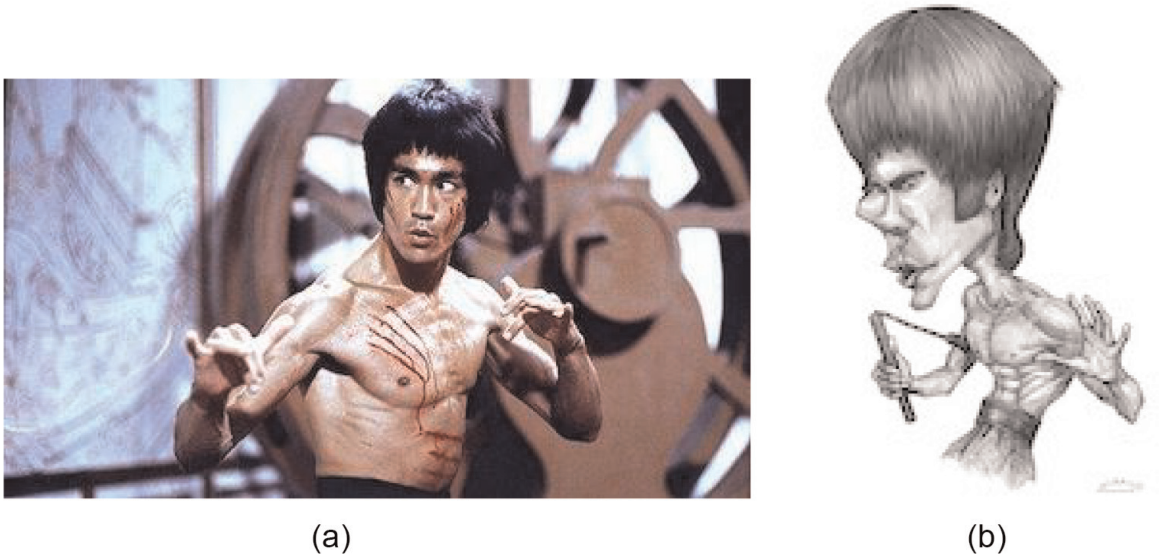

Figure 8.
An instance of caricaturists exaggerating racist components of Bruce Lee's features that do nothing to enhance the identifiability of the image. (a) veridical image of Bruce Lee. (b) racist representation of Bruce Lee.

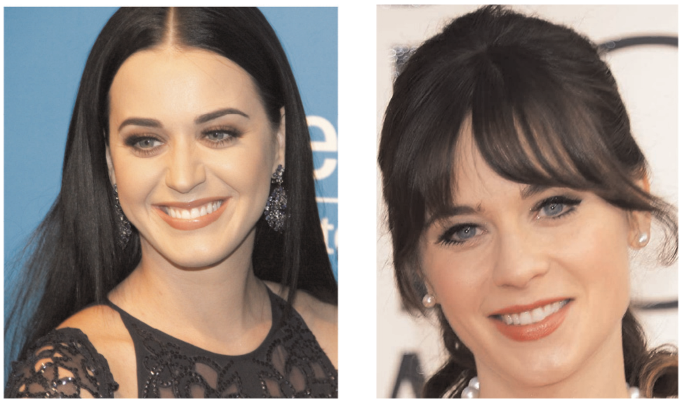

Figure 9.
In some cases, veridical image and caricature image identity may be difficult to distinguish, as shown by Katy Perry (left) and Zooey Deschanel (right) in this figure. Thus, it is imperative to introduce a significant amount of quality data to allow a surveillance system to differentiate between similar faces.

misidentification can have life-long impacts on a suspect's quality of life and likelihood to be reincarcerated.

Previous works have focused primarily on gathering as much data as possible [49, 50, 56], and while there is certainly a data problem in the machine learning field, the issue of bad data is far greater when constructing a system meant to surveil. In order for systems to learn accurate representations, those systems must be trained on accurate representations. That is, if we want a system that recognizes that two similar-looking people, such as Katy Perry and Zooey Deschanel, are different (see **Figure 9**), we need to supply a significant amount of representative data to do that, and that applies to every race, gender, and age. However, datasets gathered using a web scraper en masse tend to use celebrities, and celebrities in western culture are typically young, Caucasian, and attractive. If data were to be gathered with complete disregard for dataset balance, the face representation constructed by that system would likely perform well at identifying young, attractive, Caucasian people, and struggle with images of anyone that does not fit that description [67], and this is supported by past research [75]. While techniques like data balancing exist [64, 76], those techniques are not typically capable of fully handling the bias present in a dataset. Thus, datasets should be constructed with as much balance between gender, age, ethnicities, and image type (caricature vs veridical) as possible to ensure that the trained system is as fair as possible, *especially in systems meant to have any applicability to law enforcement*. This can be difficult to do, depending on the dataset content and availability of applicable data on the internet.

Since caricatures are a unique data source, gathering relevant, representative data is made even more difficult. Currently, the largest publicly available dataset for caricature verification and recognition is WebCaricature [56]. We have already outlined why the mindset of "quantity over quality" is detrimental to creating a fair recognition system. The WebCaricature dataset [56] illustrates this point well. WebCaricature consists of 6,042 caricatures and 5,974 veridical images over 252 identities. At a cursory glance, this dataset seems like a great resource just due to its sheer size. However, we find that there are many quality issues with the dataset itself, examples of which can be seen in **Figure 7**. First, the dataset does not bother checking that images fairly represent the target example. In other words, the target identity of each caricature is not immediately clear. Because these caricatures are not representative, they should not be included in the dataset. Second, both the caricature portion and photo portion of the dataset contain images that are not of their respective type. For example, there are multiple caricatured images that are a drawing, cartoons, or veridical images incorrectly labeled as a caricature. Third, there are many instances where the dataset contains duplicate images or images that are not of the target identity. Fourth, the authors did not collect a dataset that was balanced in terms of gender, ethnicity, or age, making it (and any system trained on it) inherently biased. Fifth, and finally, there are many included identities that have dozens of veridical images and only a single caricature image. This introduces a bias toward photo representations into the dataset and any system that uses it. After careful analysis, it becomes clear that the WebCaricature dataset's focus on quantity has led to a marked decrease in quality that would unduly bias any surveillance system that uses it.

Thus, we propose that the following list of questions be used to construct future caricature datasets:

1. Are there a significant number of images (caricature AND veridical) of this identity to create an accurate representation of this person?

2. Are images being used as caricatures actually caricatures, and do the caricatures accurately represent the identity without incorporating racist or cultural elements?

3. Factoring in all identities, is there as much balance across identity race, gender, and age as possible?

4. Factoring in all images, is there as much balance across identity race, gender, and age as possible?

We concede that most publicly sourced datasets from Western culture will have an easier time collecting images of white individuals. That means that maintaining race balance will restrict the number of images of Caucasian subjects that can be collected since a roughly equal balance is necessary to ensure fairness. In terms of dataset size, this means sacrificing quantity for quality in the interest of creating a fair surveillance system.

5. Using caricatures for prominent feature recognition in surveillance systems

As discussed in Section 3, existing research does not address prominent facial feature recognition, despite the fact that cognitive psychology has been trying to better identify them for decades. The most common approach to designing the system architecture for face recognition is to first detect any faces in the image and then to landmark that image. The landmarks are then given as input to some sort of machine learning algorithms, such as a deep belief network [28], convolutional neural network [56], or genetic algorithm [53]. We believe that this same generalized process can continue to be used, so long as the field addresses the gap between prominent feature research in cognitive psychology and computer vision. This can be accomplished by using caricatures to better model prominent feature exaggeration.

Future work using caricatures should seek to address this critical gap in facial recognition by doing the following:

1. Improve prominent feature recognition and labeling using caricatures.

2. Utilize prominent feature recognition to better train multi-task surveillance systems that leverage unique facial attributes.

3. Generate high-fidelity caricatures with a strong resemblance to the veridical image.

We address each of these points below, with suggestions for courses of future research.

Future research should use landmarking on caricature and veridical images to measure feature deviation from the average. Controlling for well-known conditions that affect feature size, such as gender and ethnicity, measuring configural and size properties of each facial feature can provide insight into what an image's prominent features are. For example, Helena Bonham Carter's eyes are large when compared to other celebrities, and they are also large in comparison to the rest of her facial features. Landmarks can be used to quantitatively analyze the difference in the size of relative features and the deviation from average in order to identify prominent features. It is worth noting that research that quantitatively analyzes feature size and shape *must* control for gender and ethnicity in order to create better models; past medical research has shown that nose shape, for example, is highly correlated with race [77]. By controlling for variables, such as race and gender, systems trained to recognize prominent features can be better attuned to small differences between features in different subjects. We warn that systems that do not implement this control into their experiments are likely to miss fine-grained feature differences, and may perpetuate bias, which is an obvious downside to the use of prominent facial features if they are not used carefully. Providing landmark data to a simple machine learning model, such as a support vector machine (SVM), or to a deep convolutional neural network should provide a baseline method for prominent facial feature recognition. This baseline should be simple to implement and is a first step in improving prominent feature recognition and labeling. Preliminary results using simple models may look fairly underwhelming because it is unlikely that they will optimally handle a large amount of landmarking data coming; however, results that are better than chance will indicate that prominent feature usage is worth pursuing.

The developed prominent feature recognition method can be used to train surveillance systems that are capable of leveraging prominent facial features. We argue that by using prominent facial feature labels and this prominent feature methodology, deep learning models used in surveillance can improve their performance. Additionally, prominent feature recognition methods can be used as an additional task in existing surveillance systems, which should ultimately make a more robust, less overfit model [31, 78, 79]. This second step is critical to better mimicking human face perception and should improve most recognition and surveillance systems. Again, if prominent features are used to train a surveillance system, race and gender need to be controlled as part of the proposed multitask network so that the system is not unintentionally biased.

After a model is in place to identify prominent features, the identified features from a veridical image can be used to better train existing GAN models used for caricature generation. This will create caricatures with higher fidelity to the veridical image. These generated images, can, in turn, be used to ameliorate the data quantity problem that most surveillance systems have.

We also note that the use of caricatures to improve existing system architectures may prove difficult at first, especially if collated datasets are relatively small, there may just be a data abundance problem. Therefore, it is imperative that large, well-constructed datasets be created prior to any landmarking or architecture improvement. Additionally, initial research in caricature usage will likely prove to be slow, since all data will need to be manually landmarked until a proper landmarking model is devised for caricatures. Aside from data abundance and the time necessary to manually create the dataset and landmarking systems, it is also likely that initial systems, no matter how well controlled, may end up slightly biased and better at identifying identities of specific races, genders, and ages. This is simply due to the fact that it is easiest to find existing caricature data through web scraping, and web scraping will inherently lead to more images of celebrities, which will be culturally skewed in favor of one race over another. In order to control this, data augmentation methods appropriate to caricature utilization should be looked at. Finally, we note that the general subjectivity of caricature acceptability, as discussed in Section 4, may lead to a level of variation across curated datasets and unintentional bias.

Given that most computer vision advancements have been made by a better understanding of human perception [1], we argue that utilizing cognitive psychology's findings about caricatures to our advantage will result in more robust computer vision systems, and in turn, more robust surveillance systems. By developing a method to identify prominent features, surveillance systems can better leverage the same mechanisms that human face perception uses to improve recognition.

6. Conclusion

Human face recognition relies on the use of prominent facial features. We outline past research in cognitive psychology that should be leveraged to improve surveillance systems. In particular, we argue that the use of prominent facial features is critical to better modeling human face perception. In addition, Section 2 discusses the importance of using a holistic face model and internal facial features to construct robust recognition systems. In Section 3, we discuss past research in computer science that leverages the use of caricatures. We note that research in this area is hindered by the lack of study of prominent facial features, and that, in most cases, existing research in computer science that uses caricatures is rather limited and of low quality. Next, we discuss the importance of collecting a dataset that is not only large but also balanced (Section 4). We argue that dataset balance in terms of gender, race, age, and image type is critical to limiting bias within surveillance systems trained on these datasets and discuss the past research that supports our stance. Additionally, we provide a series of guidelines for caricature dataset generation, so that future caricature datasets are of acceptable quality for use in surveillance system training. Finally, we outline the ways in which caricatures can be used to improve facial recognition systems. In particular, we argue that improved prominent feature labeling and recognition is critical, so that these features can be used to better train multitask surveillance systems.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant IIS-1909707.

Author details

Sara R. Davis and Emily M. Hand*
University of Nevada, Reno, Reno, NV, USA

*Address all correspondence to: emhand@unr.edu

References

[1] Scheirer WJ, Anthony SE, Nakayama K, Cox DD. Perceptual annotation: Measuring human vision to improve computer vision. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;**36**(8): 1679-1686

[2] Wright T. A History of Caricature and Grotesque in Literature and Art. Virtue Brothers; 1865

[3] Dawel A, Wong TY, McMorrow J, Ivanovici C, He X, Barnes N, et al. Caricaturing as a general method to improve poor face recognition: Evidence from low-resolution images, other-race faces, and older adults. Journal of Experimental Psychology Applied. 2019; **25**(2):256-279

[4] Sun YK, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014. pp. 1891-1898

[5] Michael B. Lewis. Are caricatures special? evidence of peak shift in face recognition. European Journal of Cognitive Psychology. 1999;**11**(1): 105-117

[6] Mauro R, Kubovy M. Caricature and face recognition. Memory & Cognition. 1992;**20**(4):433-440

[7] Rhodes G, Brennan S, Carey S. Identification and ratings of caricatures: Implications for mental representations of faces. Cognitive Psychology. 1987; **19**(4):473-497

[8] Rhodes G, Tremewan T. Understanding face recognition: Caricauture effects, inversion, and the homogeneity problem. Visual Cognition. 1994;**1**(2–3):275-311

[9] Alex H, Hancock PJB, Kittler J, Langton SRH. Improving discrimination and face matching with caricature. Applied Cognitive Psychology. 2013; **27**(6):725-734

[10] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks, 2014

[11] Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, 2015

[12] Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. 2015

[13] Maurer D, Le Grand R, Mondloch C. The many faces of configural processing. Trends in Cognitive Sciences. 2002;**6**: 255-260

[14] James W, Sengco JA. Features and their configuration in face recognition. Memory & Cognition. 1997;**25**:583-592

[15] Tanaka J, Farah M. Parts and wholes in face recognition. The Quarterly journal of experimental psychology. A, Human experimental psychology. 1993; **46**:225-245

[16] Tanaka JW, Simonyi D. The “parts and wholes” of face recognition: A review of the literature. Quarterly Journal of Experimental Psychology. 2016;**69**(10):1876-1889

[17] Ellis H, Shepherd J, Davies G. Identification of familiar and unfamiliar

faces from internal and external features: Some implications for theories of face recognition. Perception. 1979;**8**:431-439

[18] Andrews T, Davies-Thompson J, Kingstone A, Young A. Internal and external features of the face are represented holistically in face-selective regions of visual cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience. 2010;**30**: 3544-3452

[19] Christopher A, Liu CH, Young AW. The importance of internal facial features in learning new faces. Quarterly Journal of Experimental Psychology. 2015;**68**(2):249-260

[20] Murphy J, Ipser A, Gaigg S, Cook R. Exemplar variance supports robust learning of facial identity. Journal of Experimental Psychology. Human Perception and Performance. 2015;**41**:4

[21] Novak R, Bahri Y, Abolafia DA, Pennington J, Sohl-Dickstein J. Sensitivity and Generalization in Neural Networks: An Empirical Study 2018

[22] Wang M, Deng W. Deep face recognition: A survey. Neurocomputing, 2021;**429**:215-244

[23] Zhao J, Zhou Y, Li Z, Wang W, Chang K-W. Learning gender-neutral word embeddings. CoRR, abs/ 1809.01496. 2018

[24] Abate AF, Nappi M, Riccio D, Sabatino G. 2d and 3d face recognition: A survey. Pattern Recognition Letters. 2007;**28**:1885-1906

[25] Jourabloo A, Liu X. Large-pose face alignment via cnn-based dense 3d model fitting. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016

[26] Zhu X, Lei Z, Liu X, Shi H, Li SZ. Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 146-155

[27] Bowyer KW, Chang K, Flynn P. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. Computer Vision and Image Understanding. 2006;**101**:1-15

[28] Huang GB, Lee H, Learned-Miller EG. Learning hierarchical representations for face verification with convolutional deep belief networks. CVPR; 2012. pp. 2518-2525

[29] Cai X, Wang C, Xiao B, Xue C, Zhou J. Deep nonlinear metric learning with independent subspace analysis for face verification. In: Proceedings of the 20th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery; 2012. pp. 749-752

[30] Guillaumin M, Verbeek J, Schmid C. Is that you? metric learning approaches for face identification. In: 2009 IEEE 12th International Conference on Computer Vision. 2009. pp. 498-505

[31] Hao Zhang J. Ross Beveridge, Bruce A. Draper, and P. Jonathon Phillips. On the effectiveness of soft biometrics for increasing face verification rates. Computer Vision and Image Understanding. 2015;**137**:50-62

[32] Taylor GW, Fergus R, LeCun Y, Bregler C. Convolutional learning of spatio-temporal features. In: Daniilidis K, Maragos P, Paragios N, editors. Computer Vision – ECCV 2010. Berlin, Heidelberg: Springer; 2010. pp. 140-153

[33] Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and

composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, ICML ’08. New York, NY, USA: Association for Computing Machinery; 2008. pp. 1096-1103

[34] Dong Y, Lei Z, Stan ZL. Towards pose robust face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013

[35] Richard O, Hart PE, Stork DG. Pattern Classification. 2nd ed. New York: Wiley; 2001

[36] Cao Z, Yin Q, Tang X, Sun J. Face recognition with learning-based descriptor. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE; 2010. pp. 2707-2714

[37] Hu J, Lu J, Tan Y-P. Discriminative deep metric learning for face verification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014. pp. 1875-1882

[38] Nguyen HV, Bai L. Cosine similarity metric learning for face verification. In: Kimmel R, Klette R, Sugimoto A, editors. Computer Vision – ACCV 2010. Berlin, Heidelberg: Springer; 2011. pp. 709-720

[39] Thom N, Hand EM. Facial Attribute Recognition: A Survey. 2020

[40] Hsiao S-H, Jang J-SR. Improving resnet-based feature extractor for face recognition via re-ranking and approximate nearest neighbor. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2019. pp. 1-8

[41] Shi Y, Deb D, Jain AK. Warpgan: Automatic caricature generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 10762-10771

[42] Gauthier J. Conditional generative adversarial nets for convolutional face generation. In: Convolutional Neural Networks for Visual Recognition. 2014. p. 2

[43] Li M, Zuo W, Zhang D. Convolutional network for attribute-driven and identity-preserving human face generation. arXiv preprint arXiv: 1608.06434, 2016

[44] Lu Y, Tai Y-W, Tang C-K. Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 282-297

[45] Wang K, Wan X. Sentigan: Generating sentimental texts via mixture adversarial networks. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. 2018. pp. 4446-4452

[46] Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. In: 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017

[47] Arjovsky M, Bottou L. Towards Principled Methods for Training Generative Adversarial Networks 2017

[48] Taphorn A. Gan and Their Chances and Risks in Face Generation and Manipulation. 2020

[49] Zhang Y, Gan X, Fan K, Chen X, Henao R, Shen D, Carin L. Adversarial Feature Matching for Text Generation. 2017

[50] Jang W, Ju G, Jung Y, Yang J, Tong X, Lee S. Stylecarigan: Caricature

generation via stylegan feature map modulation. arXiv preprint arXiv: 2107.04331 2021

[51] Chiang P-Y, Liao W-H, Li T-Y. Automatic caricature generation by analyzing facial features. In: Proceeding of 2004 Asia Conference on Computer Vision (ACCV2004). Korea; 2004

[52] Zipeng Ye, Ran Yi, Minjing Yu, Juyong Zhang, Yu-Kun Lai, and Yong-jin Liu. 3d-carigan: An end-to-end solution to 3d caricature generation from face photos. IEEE Trans Vis Comput GraphIEEE Trans Vis Comput Graph, abs/2003.06841. 2021

[53] Brendan F, Bucak SS, Jain AK, Akgul T. Towards automated caricature recognition. In: 2012 5th IAPR International Conference on Biometrics (ICB). 2012. pp. 139-146

[54] Abacı B, Akgül T. Matching caricatures to photographs. Signal Image and Video Processing. 2015;**9**:1-9

[55] Mike Burton A, Jenkins R, Hancock PJB, White D. Robust representations for face recognition: The power of averages. Cognitive Psychology. 2005;**51**:256-284

[56] Huo J, Li W, Shi Y, Yang G, Yin H. Webcaricature: A benchmark for caricature recognition. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK: BMVA Press; 2018. p. 223

[57] Berg T, Belhumeur PN. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. Computer Vision and Pattern Recognition. 2013: 955-962

[58] Berg T, Belhumeur PN. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE Computer Society; 2013. pp. 955-962

[59] Kumar N, Belhumeur PN, Nayar SK. Facetracer: A search engine for large collections of images with faces. In David A. Forsyth DA, Torr PHS, Zisserman A, editors, Computer Vision - ECCV 2008, 10th European Conference on Computer, Vision, Marseille, Proceedings, Part IV, volume 5305 of Lecture. Notes in Computer Science. France: Springer; 2008. pp. 340-353

[60] Kumar N, Berg AC, Belhumeur PN, Nayar SK. Attribute and simile classifiers for face verification. In IEEE 12th International Conference on Computer Vision, ICCV 2009. Kyoto, Japan: IEEEComputer Society; 2009. pp. 365-372

[61] Kumar N, Berg AC, Belhumeur PN, Nayar SK. Describable visual attributes for face verification and image search. In: PAMI. 2011

[62] Layne R, Hospedales TM, Gong S, Mary Q. Person re-identification by attributes. In Bowden R, Collomosse JP, Mikolajczyk K, editors. British Machine Vision Conference, BMVC 2012, Surrey, UK: BMVA Press; 2012. pp. 1-11

[63] Dharr S, Ordonez V, Berg TL. High level describable attributes for predicting aesthetics and interestingness. In The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011. Colorado Springs, CO, USA: IEEE Computer Society; 2011. pp. 1657-1664

[64] Hand EM, Chellappa R. Attributes for improved attributes: A multi-task network utilizing implicit and explicit

relationships for facial attribute classification. In Singh S, Markovitch S, editors. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI Press; 2017. pp. 4068-4074

[65] Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile: IEEE Computer Society; 2015. pp. 3730-3738

[66] Rudd EM, Gunther M, Boult TE. Moon: A mixed objective optimization network for the recognition of facial attributes. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V, volume 9909 of Lecture Notes in Computer Science, Amsterdam, The Netherlands: Springer; 2016. pp. 19-35

[67] Cortes C, Jackel LD, Chiang W-P. Limits on learning machine accuracy imposed by data quality. In: Advances in Neural Information Processing Systems. 1994. p. 7

[68] Jain B, Patel H, Nagalapatti L, Gupta N, Mehta S, Guttula S, Mujumdar N, et al. Overview and importance of data quality for machine learning tasks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020. pp. 3561-3562

[69] Cummaudo M, Guerzoni M, Marasciuolo L, Gibelli D, Cigada A, Obertovà Z, et al. Pitfalls at the root of facial assessment on photographs: A quantitative study of accuracy in positioning facial landmarks. International Journal of Legal Medicine. 2013;**127**(3):699-706

[70] Lin J, Xiao L, Wu T. Face recognition for video surveillance with aligned facial landmarks learning. Technology and Health Care. 2018;**26**(S1):169-178

[71] Google apologises for photos app's racist blunder, July 2015

[72] Crockford K. How is Face Recognition Surveillance Technology Racist?: News & Commentary, Jun 2020

[73] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: PMLR. 2018

[74] Patrick Gother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) - nist

[75] Lingenfelter B, Hand EM. Improving evaluation of facial attribute prediction models. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). Jodhpur, India: IEEE; 2021. pp. 1-7

[76] Gustavo EAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter. 2004;**6**:20-29

[77] Suhk JH, Park JS, Nguyen AH. Nasal analysis and anatomy: Anthropometric proportional assessment in asians-aesthetic balance from forehead to chin, part i 2015

[78] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. Advances in Neural Information Processing Systems. 2007;**2007**:41-48

[79] Ranjan R, Patel VM, Chellappa R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. CoRR. 2016;abs/1603.01249