2

Facial Recognition Technologies 101

Technical Insights

Ali Akbari

2.1 INTRODUCTION

The best way to anticipate the risks and concerns about the trustworthiness of facial recognition technologies (FRT) is to understand the way they operate and how such decision-making algorithms differ from other conventional information technology (IT) systems. This chapter presents a gentle introduction to characteristics, building blocks, and some of the techniques used in artificial intelligence (AI) and FRT solutions that are enabled by AI. Owing to simplification and limitation, this is by no means a complete or precise representation of such technologies. However, it is enough to better understand some of the available choices, the implications that might come with them, and considerations to help minimise some of the unwanted impacts.

When talking about facial recognition technologies, usually the first thing that comes to mind is identifying a person from their photo. However, when analysing an image that includes a face, quite a few processes can be done. Apart from the initial general image preparation and enhancement steps, everything starts with a face detection process. This is the process to find the location of all of the faces within an image, which usually follows by extracting that part of the image and applying some alignments to prepare it for the next steps.

Face recognition that follows the detection step deals with assessing the identity of the person in the extracted face image and can be either an identification or a verification process. Face identification is when a 1:N, or one-to-many, search happens and the target face image is compared with a database of many known facial images. If the search is successful, the identity of the person in the image is found. For example, when doing a police check, a newly taken photo of the person might be checked against a database of criminal mugshots to find if that person had any past records. In the verification process, by performing a 1:1, or one-to-one check, we are actually trying to confirm an assumed identity by comparing a new facial image with a previously confirmed photo. A good example for this can be when a newly taken photo at a border checkpoint is compared with the photo on the passport to confirm it is the same person.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

Although it is not always categorised under the facial recognition topic, another form of facial image processing is face categorisation or analysis. Here, rather than the identity of the person in the image, other characteristics and specifications are important. Detecting some demographic information such as gender, age, or ethnicity, facial expression detection, and emotion recognition are a few examples with applications such as sentiment analysis, targeted advertisement, attention detection, or driver fatigue identification. However, this sub-category is not the focus in this text.

All of the above-mentioned processes on facial images fall under the computer vision field of research, which is about techniques and methods that enable computers to understand images and extract various information from them. This closely relates to image processing, which can, for example, modify and enhance medical images but not necessarily extract information or automatically make decisions based on them. Eventually, if we go one step further, along with computer vision and image processing, any other unstructured data processing such as speech processing or natural language processing falls under the umbrella of AI. The importance of this recognition is that facial recognition technologies inherit a lot of their characteristics from AI, and in the next section we take a closer look at some of these specifications to better understand some of the underlying complexities and challenges of FRT.

2.2 WHAT IS AI?

Although there have been many debates around the definition of AI, we do not yet have one universally accepted version. The definition by the Organisation for Economic Co-operation and Development (OECD) is among one of the more commonly referenced ones: 'Artificial Intelligence (AI) refers to computer systems that can perform tasks or make predictions, recommendations or decisions that usually require human intelligence. AI systems can perform these tasks and make these decisions based on objectives set by humans but without explicit human instructions.'[1]

2.2.1 *AI versus Conventional IT*

While the OECD has provided a good definition, in order to better understand AI systems and their characteristics it would be beneficial to compare them with conventional IT systems. This can be considered across the following three dimensions:

[1] OECD, *Artificial Intelligence in Society* (OECD Publishing, 2019), https://doi.org/10.1787/eedfee77-en.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

- *Instructions* – In order to achieve a goal, in conventional IT systems, explicit and step by step instructions are provided. However, AI systems are given *objectives* and the system comes up with the best solution to achieve it. This is one of the most important factors that makes the behaviour of AI systems not necessarily predictable because the exact solution is not dictated by the developers of the system.
- *Code* – The core of a conventional IT system is the codebase in one of the programming languages that carries the above-mentioned instructions. Although AI systems also contain codes that define the algorithms, the critical component that enables them to act intelligently is a *knowledge* base. The algorithms apply this knowledge on the inputs to the system to make decisions and perform tasks (so called outputs).
- *Maintenance* – It is very common to have periodic maintenance on conventional IT systems to fix any bugs that are found or add/improve features. Moreover, an AI system that is completely free of bugs and performing perfectly might gradually drift and start behaving poorly. This can be because of changes in the environment or the internal parameters of the models in the case of continuous learning capability (this is discussed further in Section 2.3.4). Owing to this characteristic, apart from maintenance, AI systems need *continuous monitoring* to make sure they perform as expected along their life cycle.

2.2.2 *Contributors in AI Systems*

A common challenge with FRT and more broadly AI systems is to understand their behaviour, explain how the system works or a decision was made, or define the scope of responsibilities and accountability. Looking from this angle, it is also worth reminding ourselves of another characteristic of AI systems, which is the possibility of many players contributing to building and applying such solutions.

For example, let us consider a face recognition solution being used for police checks. The algorithm might be from one of the latest breakthroughs developed by a research centre or university and publicly published in a paper. Then a technology provider may implement this algorithm in their commercial tools to create an excellent face matching engine. However, in order to properly train the models in this engine, they leverage the data being collected and prepared by a third company that may or may not have commercial interest in it. This face matching engine by itself only accepts two input images and outputs a similarity score that cannot be used directly by police. Hence a fourth company comes into play by integrating this face matching engine in a larger biometrics management solution in which all required databases, functionalities, and user interfaces exactly match the police check requirements. Before putting this solution into operation, the fifth player is the police department, which, in collaboration with the fourth company, runs tests and decides the suitable parameters and

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

configuration that this solution should use when implemented. Finally, the end users who will take a photo during operation of the system may affect success as the sixth player by providing the image with the best conditions.

In such a complex scenario, with so many contributors to the success or failure of an FRT solution, investigating the behaviour of the system or one specific decision is not as easy as in the case of other simpler software solutions.

2.3 AI LIFE CYCLE AND SUCCESS FACTOR CONSIDERATIONS

Considering the foregoing, the life cycle of AI systems also differs slightly from the common software development life cycle. Figure 2.1 is a simple view of these life cycle steps.

2.3.1 *Design*

Following the inception of an idea or identification of a need, it all starts with the design. Many critical decisions are made at this stage that can be based on various hypotheses and potentially reviewed and corrected in the later steps. Such decisions may include but are not limited to the operations requirements, relevant data to be collected, expected data characteristics, availability of training data or approaches

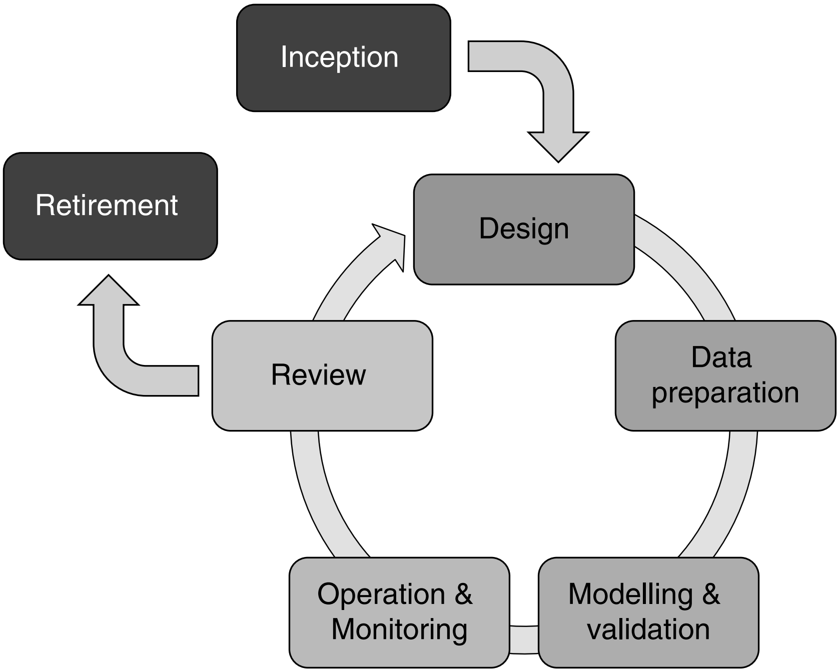


FIGURE 2.1 AI system life cycle

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

to create them, suitable algorithms and techniques, and acceptance criteria before going into operation. For example, an FRT-based access control system developer might assume that their solution is going to be always used indoors and in a controlled imaging environment, and decide only simple preprocesses are required based on this consideration. A system developed based on this design may perform very poorly if used for outdoor access control and in a crowded environment with varying light and shade conditions.

2.3.2 *Data Preparation*

The data preparation can be one of the most time-consuming and critical steps of the work. As discussed in Sections 2.3.3 and 2.6, this can also be an important factor in success, failure, or unwanted behaviour of the system. This stage covers all the data collection or creation, quality assessment, cleaning, feature engineering, and labelling steps. When it comes to the data for building and training AI models, especially in a complex and sensitive problem such as face recognition, there is always the difficult trade-off between volume, quality, and cost. More data helps to build stronger models, but curating lots of high-quality data is very costly. Owing to the time costs and other limitations in the creation of such datasets, sometimes the developers are forced to rely on lower quality publicly available or crowd-sourced datasets, or pay professional data curation companies to help them with this step. For a few examples of the datasets commonly used in FRT development, you can refer to Labeled Face in the Wild,[2] Megaface,[3] or Ms-celeb-1m.[4] However, developers should note that not only it is a very difficult task to have a thorough quality check on such huge datasets, but also each has its own characteristics and limitations that are not necessarily beneficial for any type of FRT development activity. Inadequate use of such datasets might lead to unwanted bias in FRT solutions that only gets noticed after repeatedly causing problems.

2.3.3 *Modelling and Validation*

When the data is prepared, actual development of the system can get started. The core of this stage, which is one of the most iterative steps in the AI life

[2] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, 'Labeled faces in the wild: A database for studying face recognition in unconstrained environments' (2007), Technical Report 07–49, University of Massachusetts, Amherst.

[3] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, 'The megaface benchmark: 1 million faces for recognition at scale', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA (27–30 June 2016), pp. 4873–4882, doi: 10.1109/CVPR.2016.527.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, 'Ms-celeb-1m: A dataset and benchmark for large-scale face recognition' in B. Leibe, J. Matas, and M. Welling (eds.), *European Conference on Computer Vision* (Springer, 2016), pp. 87–102.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

cycle, is to find the most suitable algorithms and configurations, and train some models by applying the algorithms to previously prepared training data. This is followed with running enough test and validation processes to become confident of the suitability of the models for the intended application. Usually, many iterations are required to get to the desirable performance levels and to confidently sign off a model to operate in the production environment. Incorrect selection of the algorithms or performance metrics and validation criteria can easily cause misleading results. For example, when checking a suspect's photo against a database of previous criminal records, we may want to consider different acceptance levels for false positive versus false negative rates; hence, a straight accuracy measure is not enough to pass or fail a model. Similarly, for a sensitive application, we might want to check such measures separately for various cohorts across demographic dimensions such as gender and ethnicity, to minimise any chances of bias. An accurate technical understanding of performance measurement metrics and meaning is critical in the correct selection and application of FRT. Unfortunately, a lack of adequate AI literacy among some of the business operators of FRT technologies can cause the choice of solutions that are not suitable for their application. For example, a technology that works well for a 1:1 verification and access control to a digital device does not necessarily perform as well as 1:N search within a criminal database.

2.3.4 *Operation and Monitoring*

Following the build and passing all readiness tests successfully, the AI system is deployed and put into operation. AI systems, as any other software, need considerations such as infrastructure and architecture to address the required security, availability, speed performance, and so on. Additionally, as briefly discussed earlier, operators should make sure that the conditions of the application are suitable and match what the models were intended and built for. What should not be forgotten is that AI systems, especially in high-risk applications, are not 'set and forget' technologies. If an AI system performs very well when initially implemented, that does not necessarily mean it will continue to keep performing at the same level. If continuous learning is used, the models keep dynamically changing and adapting themselves, which of course means the new behaviour needs to be monitored and confirmed. However, even if the models are static and not changing, a drift can still happen, which changes the performance of the models. This can be due to changes in the concept and the environment in which the model is performing. For example, specific facial expressions in different cultures might appear differently. Hence, an FRT system that is built successfully to detect various facial expressions in a specific country might start behaving poorly when too many people from a different cultural background start interacting with it. A monitoring process alongside the main solution makes sure such unexpected changes are detected

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

in time to be addressed properly. For instance, a very simple monitoring process for the scenario described here can be to observe the ratio of various expressions that are detected on a regular basis. If a persistent shift in detecting some specific expressions happens, it can be a signal to start an investigation. A good approach is to build the pairing monitoring processes in parallel with the design and development of the main models.

2.3.5 *Review*

Review can happen periodically, similar to with conventional software, or based on triggers coming from the monitoring process. It can be considered as a combination of simplified evaluation and design steps that identifies the gaps between the existing circumstances of the AI system and the most recent requirements. As a result of such an assessment, the AI models may go through another round of redesign and retraining or be completely retired because of changes in circumstances.

2.4 UNDER THE HOOD OF AI

At a very simplistic level and in a classic view, an AI system consists of a form of representation of knowledge, an inference engine, and an optional learn or retrain mechanism, as illustrated in the Figure 2.2.

Knowledge in an AI system may be encoded and represented in different forms including and not limited to rules, graphs, statistical distributions, mathematical equations and their parameters, or a combination of these. The knowledge base represents facts, information, skills or experiences from human knowledge or existing relationships, associations, or other relevant information in the environment that can help in achieving the main objective of the AI system. For example, in an FRT system the knowledge might define what shapes, colours, or patterns can indicate the location of a human face in the input image. Or it can suggest what areas and measurements on

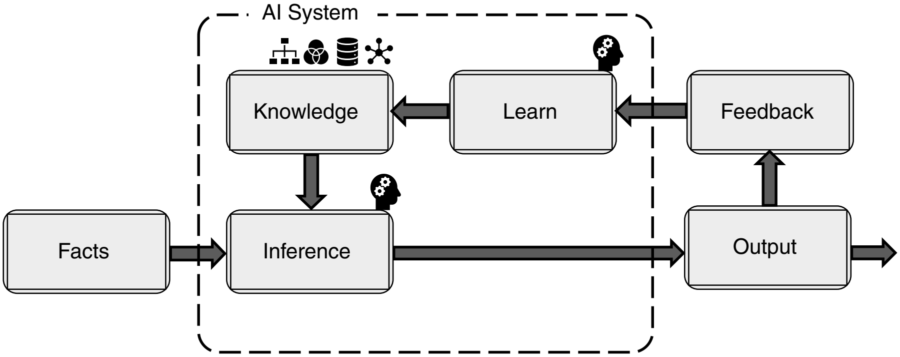


FIGURE 2.2 AI system key components

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

the face would be the most discriminating factors between two different human faces. However, it is not always as explicit and explainable as these examples.

Inference engine consists of the algorithms, mechanisms, and processes that allow the AI system to apply knowledge to the input facts and observations and to come up with the solutions for achieving its objective, making a prediction or a decision. The type of inference engine depends on the knowledge representation model to be able to apply that specific type of model, and usually they come as a pair. However, these two components are not always necessarily separable. For example, in AI systems based on artificial neural networks (ANNs), the knowledge is stored as the trained parameters and weights of the network. In such cases we can consider that the inference engine and knowledge base are combined as an ANN algorithm together with its parameters after training.

Learn or retrain, as already mentioned, is an optional component of the AI system. Many AI systems after being fully trained and put into operation remain static and do not receive any feedback from the environment. However, when the 'learn' component exists, after making a decision or prediction, the AI system receives feedback that indicates the correct output. The learning mechanism compares the predicted output with the feedback and, in case of any deviation or error, it tries to readjust the knowledge to gradually minimise the overall error rate of the system. For example, every time that your mobile phone Face ID fails to identify your face and you immediately unlock the phone using your passcode, it can be used as a feedback signal to improve your face model on the phone by using the most recently captured image. While this is a great feature for improving AI models, it also has the risk of changing their behaviour in an unexpected or unwanted manner. In the example just given, if with each failure your mobile phone keeps expanding the scope of acceptable facial features that unlock your phone, it may end up accepting other people whose faces are only similar to yours.

2.4.1 *The Source of Knowledge*

We have just mentioned how the knowledge base might be updated and improved based on the feedback received during the operation. But what is the source of the knowledge and how that knowledge base is created in the first place? Generally speaking, during the initial build of an AI system the knowledge base can be created either manually by the experts or automatically using suitable data. You might have previously seen illustrations similar to Figure 2.3, which tries to explain the relation between AI and machine learning (ML). However, before getting to the details of ML, it might be good to consider what is AI outside the ML subset.

The AI techniques outside the ML subset are called *Symbolic AI* or sometimes referred to as Good Old-Fashioned AI. This is mostly based on the human expert knowledge in that specific domain, and the knowledge base here is being manually

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

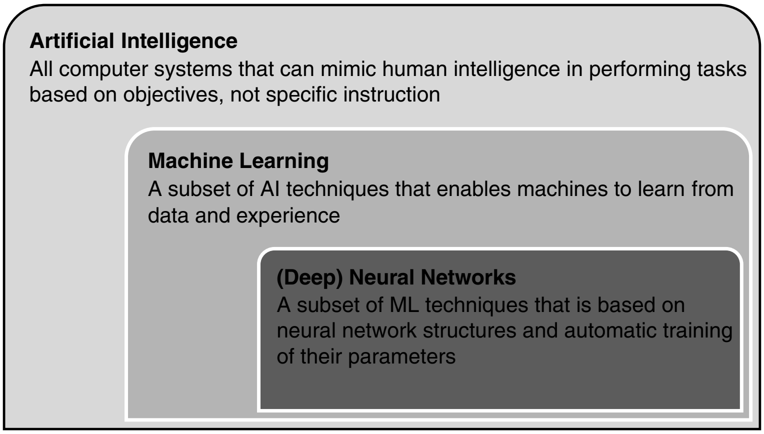

Artificial Intelligence
All computer systems that can mimic human intelligence in performing tasks based on objectives, not specific instruction
Machine Learning
A subset of AI techniques that enables machines to learn from data and experience
(Deep) Neural Networks
A subset of ML techniques that is based on neural network structures and automatic training of their parameters

FIGURE 2.3 AI versus ML

curated and encoded by the AI developers. As a result of that, it is mostly human readable (hence symbolic) and usually separable from the inference part of the system as described in the building blocks of AI earlier. Expert Systems are one of the well-known and more successful examples of symbolic AI, where their knowledge is mainly stored as 'if-then' rules.[5]

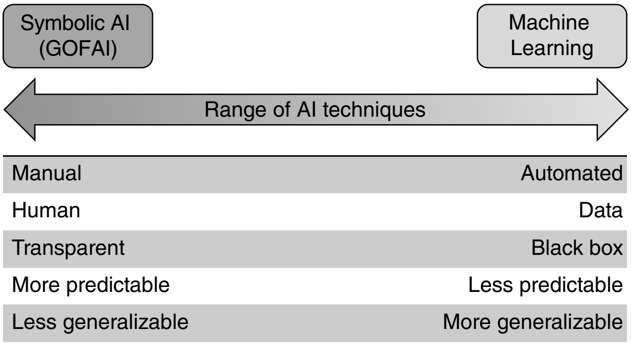
Symbolic AI systems are relatively reliable, predictable, and more explainable owing to their transparency and the readability of their knowledge base. However, the manual curation of the knowledge base makes it less generalisable and more importantly converts the knowledge acquisition or updating step into a bottleneck owing to the limited availability of the domain experts to collaborate with the developers. Symbolic AI solutions have therefore had limited success, and we have not heard much about them recently.

To obtain knowledge without experts dictating it, another approach is to observe and automatically learn from the relevant examples, which is the basis of computational learning theory and ML techniques. There is a wide range of ML techniques starting from statistical models and mathematical regression analysis to more algorithmic methods such as decision trees, support vector machines, and ANNs, which are one of the most well-known subsets of ML in the past couple of years, thanks to the huge success stories of deep neural networks.[6] When enough sample data is provided, these algorithms are capable of training models with automatically encoded knowledge that is required to achieve their objectives when put into operation. The table in Figure 2.4 summarises some of the key differences between these two groups of AI techniques.

[5] P. Jackson, *Introduction to Expert Systems* (3rd ed., Addison, Wesley, 1998), p. 2.

[6] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th ed., Pearson, 2021); T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, 2009).

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press



Manual	Automated
Human	Data
Transparent	Black box
More predictable	Less predictable
Less generalizable	More generalizable

FIGURE 2.4 Symbolic AI versus ML

2.4.2 *Different Methods of Learning*

Depending on the type and specifications of the data available to learn from, there are several different methods of learning in ML algorithms. Each one of these options has strengths and weaknesses. In an application such as FRT where we might not easily access any type of dataset that we want, it is important to be aware of the potentials and limitations of different methods. Below are a few examples among many of these methods; it is an increasing list.

Supervised learning is one of the most common and broadly applied methods. It can be utilised when at the time of creating and training ML models there are enough samples of input data along with their expected output (labels). In an identity verification example under FRT domain, the trained model would normally be expected to receive two face images and give a similarity score. In such a case, the training dataset includes many pairs of facial images along with a manually allocated label, which is 1 when those are photos of the same person and 0 otherwise. In FRT applications, preparing large enough labelled datasets for supervised learning purposes is time consuming, expensive, and subject to human errors such as bias.

Unsupervised learning applies when only samples of the input data are available for the training period and the answers are unknown or unavailable. As you can imagine, this method is only useful for some specific use cases. Clustering and association models are common examples of this learning method. For example, in a facial expression categorisation application, during the training phase a model can be given lots of facial images and learns how to group them together based on similarity of the facial expression, without necessarily having a specific name for those groups. For such FRT, it might be easier to source unlabelled sample data in larger volumes, for example through web scraping. However, this is subject to privacy implications and hidden quality issues, and thus works for limited applications only.

Reinforcement learning is used when neither the samples nor the answers are available as a batch in the beginning. Rather, a reward function is maximised through trial and error while the model gradually learns in operation. For example, you can imagine an AI system that wants to display the most attractive faces from a database to its user. There is no prior dataset to train the model for each new user, however, assuming the amount of time the viewer spends before swiping to the next photo is a sign of attractiveness, the model gradually learns which facial features can maximise this target. In such situations, the learning mechanism should also balance between exploring new territories and exploiting current knowledge to avoid possibilities of local maxima traps. It is easy to imagine that only very few FRT applications can rely on such trial and error methods to learn.

Semi-supervised learning can be considered as the combination of supervised and unsupervised learning. This can be applied when there is a larger amount of training samples, but only a smaller subset of them is labelled. In such scenarios, in order to make the unlabelled subset useful in a supervised manner, some assumptions such as continuity or clustering are made to relate them to the labelled subset of the samples. Let us imagine a large set of personal photos with only a few of them labelled with names for training a facial identification model. If we know which subsets are taken from the same family albums, we may be able to associate a lot more of those unnamed photos and label them with the correct names to be used for better training of the models. Although this can help with the data labelling challenge for FRT applications, the assumptions necessarily made during this process can introduce the risk of unwanted error in the training process.

Self-supervised learning helps in another way with the challenge of labelled data availability, especially when a very large volume of training data is required, such as deep learning. Instead of a manual preparation of the training signals, this approach uses some automated processes to convert input data to meaningful relations that can be used to train the models. For example, to build and train some of the largest language models, training data is scraped from any possible source on the internet. Then, an AI developer could use, for example, a process to remove parts of the sentences, and the main model is trained to predict and fill in the blanks. In this way the answer (training signal) is automatically created, and the language model learns all meaningful structures and word relationships in human language. In the FRT domain you can think of other processes, including distortions to a face image such as shadows or rotation, or taking different frames of the same face from a video. This produces a set of different facial images that are already known to be of the same person and can be used directly for training of the models without additional manual labelling.

2.5 FACIAL RECOGNITION APPROACHES

Similar to the AI techniques, facial recognition approaches were initially more similar to Symbolic AI. They were naturally more inclined towards the way humans

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

might approach the problem and were inspired by anthropometry.[7] Owing to the difficulty of extracting all important facial features and accurate measurements that could be easily impacted by small variation in the images, there was limited success in such works until more data driven approaches were introduced; these were based on mathematical and statistical methods and had a holistic approach to face recognition, an example being Eigenfaces,[8] which is basically the eigenvectors of the training grayscale face images (An eigenvector of a matrix is a non-zero vector that, when multiplied by the matrix, results in a scaled version of itself.). This shift towards ML techniques got more mature and successful by combining the two approaches through other ideas such as neural networks in DeepFace,[9] and many other similar works. More in-depth review of the history of FRT is discussed in Chapter 3, so here we just look at technical characteristics and differences of these approaches.

Feature analysis approaches rely on the detection of facial features and their measurements. Here, each face image is converted to a numeric vector in a multi-dimensional space and the face recognition challenge is simplified to more common classification or regression problems. Similar to symbolic AI, the majority of the knowledge, if not all, is manually encoded in the form of rules that instruct how to detect the face within an image and identify each of its components to be measured accurately. These rules may rely on basic image and signal processing techniques such as edge detection and segmentation. This makes the implementation easier and, as mentioned earlier when discussing symbolic AI, the process and its decision-making is more transparent and explainable. However, intrinsic to these approaches is the limited generalisability challenge of symbolic AI. In ideal and controlled conditions these methods can be quite accurate, but changes in the imaging condition can dramatically impact the performance. This is because in the new conditions, including different angles, resolution, or shadows and partial coverage, the prescribed rules might not apply any more, and it would not be practical to manually find all these variations and customise new rules for them.

Holistic approaches became popular after the introduction of Eigenfaces in the early 1990s.[10] Rather than trying to detect facial features based on human definition of a face, these approaches consider the image in its pixel form as a vector in a high dimensional space and apply dimensionality reduction techniques combined with other mathematical and statistical approaches that do not rely on what is inside the image. This largely simplifies the problem by avoiding the facial feature extraction

[7] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, 'Identification of human faces' (1971) 59(5) *Proceedings of the IEEE* 748–760, https://doi.org/10.1109/PROC.1971.8254.

[8] M. Turk and A. Pentland, 'Eigenfaces for recognition' (1991) 3(1) *Journal of Cognitive Neuroscience* 71–86.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, 'Deepface: Closing the gap to human-level performance in face verification' (2014), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708.

[10] Turk and Pentland, 'Eigenfaces for recognition'.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

and measurement step, together with its sensitivities. This shifts the face recognition approach towards the classic ML techniques and changes the training to a data-driven problem rather than manual rule development. Unfortunately, purely holistic approaches still suffer from a few challenges, including statistical distribution assumptions behind the method that do not always apply, and any deviation from the controlled imaging condition makes it worse.

Deep neural networks made a leap in the advancement and success of face recognition approaches. After Eigenfaces and its variations, there were many other small improvements made to the holistic approaches by adding some generic feature extraction steps such as Gabor prior to the main classifier,[11] followed by some neural network-based ML approaches. However, it was not as successful until the introduction of deep learning for image processing,[12] and applying it for face recognition.[13] Convolutional neural networks convert the feature extraction and selection from the images to an unsupervised process, so it is not as challenging as manually defined facial features and not too generic like the Gabor filters used prior to some of the holistic approaches. The increasingly complex and important features that are automatically selected are used in a supervised learning layer to deliver the classification or recognition function.[14] This is the key in the success of object and face recognition of deep neural networks.

2.6 THE GIFT AND THE CURSE OF COMPLEXITY

Many variations of ANNs have been used in ML applications including face recognition. However, the so called shallow neural networks were not as successful owing to their limited learning capacity. Advancements in hardware, use of graphical processing units, and cloud computing to increase processing power along with access to more training data (big data) made the introduction of deep learning possible. In addition to novel network structures and the use of more sophisticated nodes such as convolutional functions, another important factor in the increased capacity of learning of DNNs is the overall complexity and scale of the network parameters to train. For example, the first experimental DNN used in FaceNet includes a total of 140 million parameters to train.

While the complexity of DNNs increases their success in learning to solve challenging problems such as face recognition, these new algorithms become

[11] C. Liu and H. Wechsler, 'Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition' (2002) 11(4) *IEEE Transactions on Image Processing* 467–476.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'Imagenet classification with deep convolutional neural network' in *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1 (Curran Associates Inc., 2012), pp. 1097–1105.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf (2014). Deepface: Closing the gap to human-level performance in face verification. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701–1708.

[14] Hannes Schulz and Sven Behnke, 'Deep learning' (2012) 26 *KI – Künstliche Intelligenz* 357–363, https://doi.org/10.1007/s13218-012-0198-z.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

increasingly data hungry. Without going into too much detail, if the number of training samples are too small compared with the number of parameters of the model, rather than learning a generalised solution for solving the problem it overfits the model and memorises the answer only for that specific subset. This causes the model to perform very well for the training samples, as it has memorised the correct answers for the training set, but fail when it comes to test and unseen samples, owing to a lack of generalisation and fitting the model only to the previously seen examples. Therefore, such successful face recognition systems based on DNNs or a variation of them are actually trained on large training facial datasets, which can be the source of new risks and concerns.

Privacy and security concerns are one of the first to pay attention to. It is difficult and expensive to create new and large face datasets with all appropriate consents in place. Many of these large datasets are collected from the web and from a few different sources where copyright and privacy statements raise problems from both legal and ethics perspectives. Additionally, after collection, such datasets could be potentially a good target for cyber-attacks, especially if the images can be correlated to other information that may be publicly available about the same person.

Data labelling is the next challenge after collection of the suitable dataset. It is labour intensive to manually label such large datasets to be used as a supervised learning signal for the models. As discussed earlier, self-supervised learning is one of the next best choices for data-heavy algorithms such as DNNs. However, this introduces the risk of incorrect assumptions in the self-supervised logic and the missing of some problems in the training process even when performance measures seem to be adequate.

Hidden data quality issues might be the key to most of the well-known face recognition failures. Usually, a lot of automation or crowdsourcing is involved in the preparation of such large face datasets. This can prevent thorough quality checks across the samples and labels, which can lead to flawed models and cause unexpected behaviour in special cases despite high performance results during the test and evaluation. Bias and discrimination are among the most common misbehaviours of FRT models, which can be either due to such hidden data quality issues or simply the difficulty of obtaining a well-balanced large sample across all cohorts.

2.7 CONCLUSION

Face recognition is one of the complex applications of AI and inherits many of its limitations and challenges. We have made a quick review of some of the important considerations, choices, and potential pitfalls of AI techniques and more specifically FRT systems. Given this is a relatively new technology being used in our daily lives, it is crucial to increase the awareness and literacy of such technologies and their potential implications from a multi-disciplinary angle for all its stakeholders, from its developers and providers to the operators, regulators, and the end users.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press

Now that with DNNs the reported performance of FR models is reaching or surpassing human performance,[15] a critical question is why we still hear so many examples of failure and find FR models insufficiently reliable in practice. Among many reasons, such as data quality discussed earlier, the difference between development and operation conditions can be one of the common factors. The dataset that the model is trained and tested on may not be a good representation of what the model will receive when put into operation. Such differences can be due to imaging conditions, demographic distribution, or other factors. Additionally, we should not forget that the performance tests are usually done directly on the FRT model. However, an FRT-based solution has a lot of other software components and configurable decision-making logic that will be applied to the facial image similarity scores. For example, such surrounding configurable logic can easily introduce human bias to a FRT solution with a good performing model at core. Finally, it is worth reminding that like many other software and digital solutions, FRT systems can be subject to adversarial attacks. It might be a lot easier to fool a DNN-based FR model using adversarial samples or patches compared with the human potential for identifying such attempts.[16]

Hence, considering all such intentional and unintentional risks, are the benefits of FRT worth it? Rather than giving a blanket yes/no answer, it should be concluded that this depends on the application and impact levels. However, making a conscious decision based on a realistic understanding of potentials and limitations of technology, along with having humans in the loop, can significantly help to minimise these risks.

[15] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J-C. Chen, C. D. Castillo, R. Chellappa, D. White, and A. J. O'Toole, 'Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms' (2018) 115(24) *Proceedings of the National Academy of Sciences* 6171–6176, https://doi.org/10.1073/pnas.1721355115

[16] Yaoyao Zhong and Weihong Deng, 'Towards transferable adversarial attack against deep face recognition' (2020) 16 *IEEE Transactions on Information Forensics and Security* 1452–1466, doi: 10.1109/TIFS.2020.3036801.

https://doi.org/10.1017/9781009321211.004 Published online by Cambridge University Press