A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution

A Survey on Heterogeneous Face Recognition: Sketch, Infra-red, 3D and Low-resolution

Shuxin Ouyang[1,2,*], Timothy Hospedales[1,*], Yi-Zhe Song[1,*], Xueming Li[2,*], Chen Change Loy[3,*], Xiaogang Wang[3,*]

1. Beijing University of Posts and Telecommunications, 2.Queen Mary University of London, 3. Chinese University of Hong Kong

Abstract

Heterogeneous face recognition (HFR) refers to matching face imagery across different domains. It has received much interest from the research community as a result of its profound implications in law enforcement. A wide variety of new invariant features, cross-modality matching models and heterogeneous datasets being established in recent years. This survey provides a comprehensive review of established techniques and recent developments in HFR. Moreover, we offer a detailed account of datasets and benchmarks commonly used for evaluation. We finish by assessing the state of the field and discussing promising directions for future research.

Keywords: Cross-modality face recognition, heterogeneous face recognition, sketch-based face recognition, visual-infrared matching, 2D-3D matching, high-low resolution matching.

1. Introduction

Face recognition is one of the most studied research topics in computer vision. After over four decades of research, conventional face recognition using visual light under controlled and homogeneous conditions now approaches a mature technology [1], being deployed at industrial scale for biometric border control [2] and producing better-than-human performance [3]. Much research effort now focuses on uncontrolled, non-visual and heterogeneous face recognition, which remain open questions. Heterogeneous face recognition (HFR) refers to the problem of matching faces across different visual domains. Instead of working with just photographs, it encompasses the problems of closing the semantic gap among faces captured (i) using different sensory devices (e.g., visual light vs. near-infrared or 3D devices), (ii) under different cameras settings and specifications (e.g., high-resolution vs. low-resolution images), and (iii) manually by an artist and automatically by a digital sensor (e.g., forensic sketches vs. digital photographs).

HFR has grown in importance and interest because heterogeneous sets of facial images must be matched in many practical applications for security and law enforcement as well as multi-media indexing. For example, visual-infrared matching is important for biometric security control, because enrollment images can be taken in controlled a setting with visual light, while probe images may be taken in infra-red if visual lighting in the access control area is not controllable. Meanwhile, sketch-based recognition is important for law-enforcement, where eyewitness sketches should be matched against mugshot databases to identify suspects.

Nevertheless, HFR poses a variety of serious challenges beyond conventional homogeneous face recognition. These include: (i) comparing single versus multi-channel imagery (e.g., infra-red versus RGB visible light images), (ii) linear and non-linear variations in intensity value due to different specular reflection properties (e.g., infra-red versus RGB), (iii) different coordinate systems (e.g., 2D versus 3D depth images), (iv) reduction of appearance detail (e.g., photo versus sketch, or high versus low-resolution), (v) non-rigid distortion preventing alignment (e.g., photo versus forensic sketch). For all these reasons, it is not possible or effective to compare heterogeneous imagery directly as in conventional face recognition.

To address these challenges, the field of HFR has in recent years proposed a wide variety of approaches to bridge the cross-modal gap, thus allowing heterogeneous imagery to be compared for recognition. Research progress in bridging this gap has been assisted by a growing variety of HFR benchmark datasets allowing direct comparison of different methodologies. This paper provides a comprehensive and up-to-date review of the diverse and growing array of HFR techniques. We categorize them in terms of different modalities they operate across, as well as their strategy used to bridge the cross modal gap – bringing out some cross-cutting themes that re-occur in different pairs

*Corresponding author at: School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, No.10 Xituchen Road, Haidian District, Beijing, China, 100876, ouyangshuxin@gmail.com

[1]Queen Mary University of London,London E1 4NS, United Kingdom

[2]Beijing University of Posts and Telecommunications, No.10, Xituchen Street, Haidian District, Beijing, China

[3]Chinese University of Hong Kong, Shatin, NT, Hong Kong

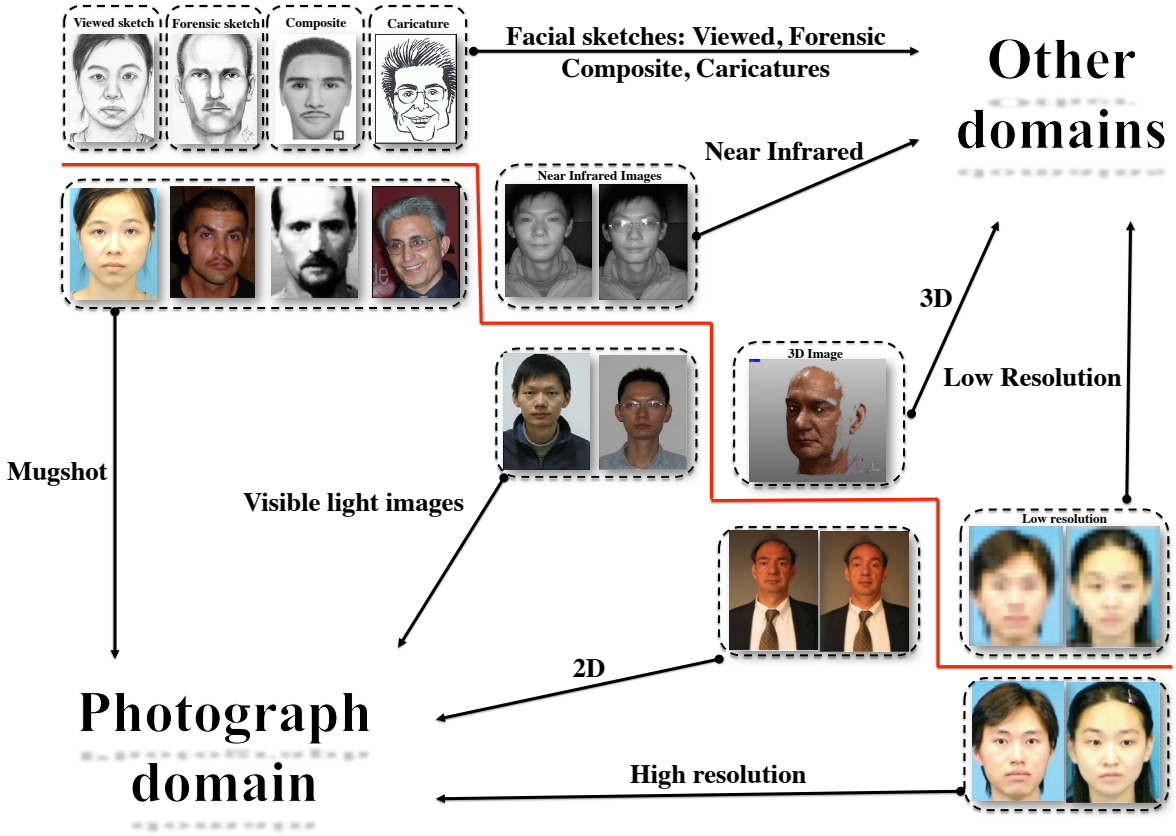

Figure 1: Scope of heterogeneous face recognition studied in this survey.

of modalities. Additionally, we summarize the available benchmark datasets in each case, and close by drawing some overall conclusions and making some recommendations for future research.

In most cases HFR involves querying a gallery consisting of high-resolution visible light face photographs using a probe image from an alternative imaging modality. We first break down HFR research in the most obvious way by the pairs of imagery considered. We consider four cross-modality applications: sketch-based, infra-red based, 3D-based and high-low resolution matching.More specifically they are:

- **Sketch**: Sketch-based queries are drawn or created by humans rather than captured by an automatic imaging device. The major example application is facial sketches made by law enforcement personal based on eye-witness description. The task can be further categorized into four variants based on level of sketch abstraction, as shown in the left of Fig. 1.

- **Near Infrared**: Near Infrared (NIR) images are captured by infrared rather than visual-light devices. NIR capture may be used to establish controlled lighting conditions in environment where visual light is not controllable. The HFR challenge comes in matching NIR probe images against visual light images. A major HFR application is access control, where enrollment images may use visual light, but access gates may use infra-red.

- **3D**: Another common access control scenario relies on an enrollment gallery of 3D images and 2D probe images. As the gallery images contain more information than the probe images, this can potentially outperform vanilla 2D-2D matching, if the heterogeneity problem can be solved.

- **Low-Resolution:** Matching low-resolution against high-resolution images is a topical challenge under contemporary security considerations. A typical scenario is that a high-resolution 'watch list' gallery is provided, and low-resolution facial images taken at standoff distance by surveillance cameras are used as probes.

Fig. 1 offers an illustrative summary of the five categories of HFR literature covered in this survey. Tab. 1 further summarizes the studies reviewed broken down by the modalities and methodological focus.

Related areas not covered by this review include (homogeneous) 3D [58] and infra-red [59] matching. View [60] and illumination [61] invariant recognition are also related, in that there exists a strong covariate shift between probe and gallery images, however we do not include these as good surveys already exist [61, 62]. Fusing modalities in

Table 1: Overview of heterogeneous face recognition steps and typical strategies for each.

Component	Approach	Sketch-Photo	VIS-NIR	2D-3D	Low-High
Representation	Analytic	Active Shape & PDMs[4, 5]			
		Relative Geometry [6]			
	Global Holistic	Whole image [7, 8, 9]	Whole image [10, 11, 12]	Whole image [13, 14, 15]	Whole image [16, 17, 18]
			Whole image [19, 20, 21]		Whole image [17, 22, 23]
		Whole image with Deep Encoder [24]			
	Global Patch	Regular grid of patches [25, 26]	Regular grid of patches [27, 12]	Regular grid of patches [28]	
		Regular grid of patches [29, 30]			
		Regular grid of patches [31, 32, 33]			
	Facial Component	Active Shape Model Detection [34]	Rectangular patches [35]		
Cross domain	Feature-based	LBP [32, 34] SSIM [33]	LBP [11, 27], LGH [21]	OGM [36]	Eigenfaces and Fisherfaces[18]
		Gabor [30], SIFT [31]	DSIFT [37] Log-DoG [38]		
		CITE [39], HOAG [29]			
	Projection	CDFE [40], Common Basis [31]	CDFE [40] LDA [41]	CCA [28, 42]	Sparse Coding [43] SDA [16]
		Kernel LDA [26]	CSR [19, 20, 35]		KCCR [17] MDS [22] CKE [23]
		RS-LDA [25], PLS [9]	CCA [10] Adaboost [35, 11]	Sparse Coding [44]	Max-margin [45]
		Sparse Coding [46]	RBMs [47]		
	Synthesis	MRF [25] Eigentransform [7, 8]	LLE [27]	3D AFM [14] 3D-FE-GEM [15]	Relationship learning [48]
		LLE [26]			DSR [48] GPA [49] RBF [50]
					Eigenface [51] S^2R^2 [18]
Matching	Multi-class	NN [7, 31, 8, 33, 4, 9]	NN [27, 52, 19, 20, 21]		NN [16]
		NN with χ^2 [29, 30]	NN with χ^2 [35, 41]	NN with χ^2 [42]	
		NN with HI [34]	NN with Cosine [10]	NN with Cosine [28]	
	Multi-class (Tr)	Bayesian [8], Metric learning [32]			Metric learning [53] SVM [48]
	Verification (Tr)	SVM [54, 24]	Similarity thresh. (Cosine) [11]		SVM[55]
		Log. Reg. [54], ANN [24]			
		Gentleboost [56, 57]			

multi-modal face recognition [63][58, 64, 59, 65, 66] is also relevant in that multiple modalities are involved. However the key difference to HFR is that multi-modal assumes both enrolment and testing images are available in all modalities, and focuses on how to fuse the cues from each, while HFR addresses matching *across* modalities with probe and enrolment image in heterogeneous modalities. Finally, a good survey about face-synthesis [67] is complementary to this work, however we consider the broader problem of cross-domain matching.

Most HFR studies focus their contribution on improved methodology to bridge the cross-modal gap, thus allowing conventional face recognition strategies to be used for matching. Even across the wide variety of application domains considered above, these methods can be broadly categorized into three groups of approaches: (i) those that synthesize one modality from another, thus allowing them to be directly compared; (ii) those that engineer or learn feature representations that are variant to person identity while being more invariant to imaging modality than raw pixels; and (iii) those that project both views into a common space where they are more directly comparable. We will discuss these in more detail in later sections.

The main contributions of this paper are summarized as follows:

1. We perform an up-to-date survey of HFR literature
2. We summarize all common public HFR datasets introduced thus far
3. We extract some cross-cutting themes face recognition with a cross-modal gap
4. We draw some conclusions about the field, and offer some recommendations about future work on HFR

The rest of this paper is organized as follow: In Section 2, we provide an overview of a HFR system pipeline, and highlight some cross-cutting design considerations. In Section 3, we provide a detailed review of methods for matching facial sketches to photos and a systematic introduction of the most widely used facial sketches datasets. In Section 4, we describe approaches for matching near-infrared to visible light face images in detail. In Section 5, we focus on matching 2D probe images against a 3D enrollment gallery. Section 6 discusses methods for matching low-resolution face images to high-resolution face images.We conclude with a discussion of current issues and recommendations about future work on HFR.

2. Outline of a HFR system

In this section, we present an abstract overview of a HFR pipeline, outlining the key steps and the main types of strategies available at each stage. A HFR system can be broken into three major components, each corresponding to an important design decision: representation, cross-modal strategy and matching strategy (Fig. 2). Of these components, the first and third have analogues in homogeneous face recognition, while the cross-modal bridge strategy is unique to HFR. Accompanying Fig. 2, Tab. 1 breaks down the papers reviewed in this survey by their choices about these design decisions.

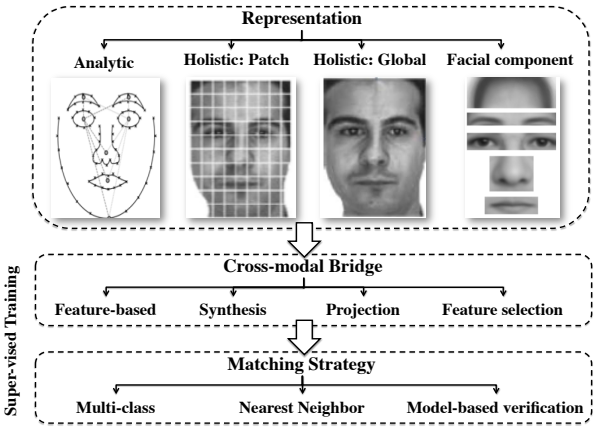

Figure 2: Overview of an abstract HFR pipeline.

2.1. Representation

The first component of a HFR system determines how the face image in each modality is represented. Common options for representations (Fig. 2, top) include analytic, component-based, patch-based, and holistic.

Analytic representations [4, 5, 6] detect facial components and fiducial points, allowing the face to be modeled geometrically, e.g., using point distribution models [4, 5]. This representation has the advantage that if a model can be fit to a face in each modality, then the analytic/geometric representation is relatively invariant to modality, and to precise alignment of the facial images. However, it is not robust to errors in face model fitting and may require manual intervention to avoid this [5]. Moreover geometry is not robust to facial expression [6], and does not exploit texture information by default.

Component-based representations detect face parts (e.g., eyes and mouth), and represents the appearance of each individually [35, 34]. This allows the informativeness of each component in matching to be measured separately [35]; and if components can be correctly detected and matched it also provides some robustness to both linear and non-linear misalignment across modalities [34]. However, a component-fusion scheme is then required to produce an overall match score between two face images.

Global holistic representations represent the whole face image in each modality with a single vector [10, 7, 40]. Compared to analytic and component-based approaches, this has the advantage of encoding all available appearance information. However, it is sensitive to alignment and expression/pose variation, and may provide a high-dimensional feature vector that risks over-fitting [68].

Patch-based holistic representations encode the appearance of each image in patches with a feature vector per patch [25, 26, 29, 30, 32]. Subsequent strategies for using the patches vary, including for example concatenation into a very large feature vector [31] (making it in effect a holistic representation), or learning a mapping/classifier per patch [39]. The latter strategy can provide some robustness if the true mapping is not constant over the whole face, but does require a patch fusion scheme.

2.2. Cross-modal bridge strategies

The key HFR challenge of cross-modality heterogeneity typically necessitates an explicit strategy to deal with the cross-modal gap. This component uniquely distinguishes HFR systems from conventional within-modality face recognition. Most HFR studies focus their effort on developing improved strategies for this step. Common strategies broadly fall into the categories: feature design, cross-modal synthesis and subspace projection. These strategies are not exclusive, and many studies employ or contribute to more than one [31, 25].

Feature design strategies [29, 30, 31, 32] focus on engineering or learning features that are invariant to the modalities in question, while simultaneously being discriminative for person identity. Typical strategies include variants on SIFT [31] and LBP [32].

Synthesis approaches focus on synthesizing one modality based on the other [7, 25]. Typical methods include eigentransforms [7, 8], MRFs [25], and LLE [26]. The synthesized image can then be used directly for homogeneous matching. Of course, matching performance is critically dependent on the fidelity and robustness of the synthesis method.

Projection approaches aim to project both modalities of face images to a common subspace in which they are more comparable than in the original representations [40, 31, 10]. Typical methods include linear discriminant analysis (LDA) [25], canonical components analysis (CCA) [10, 28], partial least squares (PLS) and common basis [31] encoding.

A noteworthy special case of projection-based strategies is those approaches that perform *feature selection*. Rather than mapping all input dimensions to a subspace, these approaches simply discover which subset of input dimensions are the most useful (modality invariant) to compare across domains, and ignore the others [35, 11], for example using Adaboost.

2.3. Matching strategy

Once an effective representation has been chosen, and the best effort made to bridge the cross-modal heterogeneity, the final component of a HFR system is the matching strategy. Matching-strategies may be broadly categorized as multi-class classifiers (one class corresponding to each identity in the gallery), or model-based verifiers.

Multi-class classifiers pose the HFR task as a multi-class-classification problem. The probe image (after the cross-modal transform in the previous section) is classified into one of the gallery classes/identities. Typically simple classifiers are preferred because there are often only one or a few gallery image(s) per identity, which is too sparse to learn complex model-based classifiers. Thus *Nearest-Neighbor (NN)* [7, 40, 31, 10] is most commonly used to match against the gallery [7]. NN classifiers can be defined

with various distance metrics, and many studies found χ^2 [29, 30] or cosine [28] to be most effective than vanilla euclidean distance. An advantage of NN-based approaches is that they do not require an explicit training step or annotated cross-modal pairs provided as training data. However, they can be enhanced with metric-learning [32] if annotated cross-domain image pairs are available for training.

Model-based verification strategies pose HFR as a binary, rather than multi-class, classification problem [11, 54]. These take a pair of heterogeneous images as input, and output one or zero according to if they are estimated to be the same person or not. An advantage of verification over classification strategies is robustness and data sufficiency. In many HFR applications there is only one cross-modal face pair per person. Thus classification strategies have one instance per class (person), and risk over fitting when training a model-based recogniser. In contrast, by transforming the problem into a binary one, all true pairs of faces form the positive class and all false pairs form the negative class, resulting in a much larger training set, and hence a stronger and more robust classifier.

In conventional face recognition, matching strategies are often adopted according to how the proposed system is to be used at test time. If the task is to recognise a face as one of a pre-defined set of people, multi-class classifiers are a natural matching strategy. If the task is to check whether a face image matches someone on a given watchlist or not, then model-based binary-verifiers are a natural choice. However, it is worth noting that multi-class classification can be performed by exhaustive verification, so many HFR systems are realized by verification, whether the final aim is verification or recognition. A second reason for the use of verification in HFR studies is that the classic forensic sketch application scenario for HFR is an open-world verification scenario (the sketch may or may not correspond to a person in the mug-shot database). For simplicity, in this paper we use the term 'recognition' loosely to cover both scenarios, and disambiguate where necessary.

We note that some methodologies can be interpreted as either cross-domain mappings or matching strategies. For example, some papers [25] present LDA as a recognition mechanism. However, as it finds a projection that maps images of one class (person identity) closer together, it also has a role in bridging the cross-modal gap when those images are heterogeneous. Therefore for consistency, we categorize LDA and the like as cross-domain methods.

2.4. Formalizations

Many HFR methods can be seen as special cases of a general formalization given in Eq. 1. Images in two modalities $\mathbf{x}^a$ and $\mathbf{x}^b$ are input; non-linear feature extraction F may be performed; and some matching function M then compares the extracted features; possibly after taking linear transforms W^a and W^b of each feature.

$$M\left(W^a F(\mathbf{x}_i^a), W^b F(\mathbf{x}_j^b)\right). \tag{1}$$

many studies reviewed in this paper can be seen as providing different strategies for determining the mappings W^a and W^b or parameterizing functions M and F.

Matching Strategies Many matching strategies can be seen as design decisions about $M(\cdot,\cdot)$. For example, in the case of NN matching, the closest match j^* to a probe i is returned. Thus M defines the distance metric $\|\cdot\|$, as in Eq. (2). In the case of model based verification strategies, a match between i and j may be declared depending on the outcome of a model's (e.g., Logistic Regression [54], SVM [54]) evaluation of the two projections (e.g., their difference), e.g., Eq. (3). In this case, matching methods propose different strategies to determine the parameters $\mathbf{w}$ of the decision function.

$$j^* = \arg\min_j \left\|W^a F(\mathbf{x}_i^a) - W^b F(\mathbf{x}_j^b)\right\| \tag{2}$$

$$\text{match} \quad \text{iff} \quad \mathbf{w}^T \left|W^a F(\mathbf{x}_i^a) - W^b F(\mathbf{x}_j^b)\right| > 0 \tag{3}$$

Cross-domain Strategies Feature-centric cross-domain strategies [29, 30, 31, 32, 11, 27, 34, 33, 21, 39] can be seen as designing improved feature extractors F. While projection/synthesis strategies can be seen as different approaches to finding the projections W^a and W^b to help make the domains more comparable. For example synthesis strategies [12, 48] may set $W^a = I$, and search for the projection W^b so that $\left|F(\mathbf{x}_i^a) - W^b F(\mathbf{x}_i^a)\right|$ is minimized. CCA [10, 28] strategies search for W^a and W^b such that $\left|W^a F(\mathbf{x}_i^a) - W^b F(\mathbf{x}_i^a)\right|$ is minimized for cross-modal pairs of the same person i. While LDA [25] strategies search for a single projection W such that $\left|WF(\mathbf{x}_i^a) - WF(\mathbf{x}_j^a)\right|$ is minimized when $i = j$ and maximized when $i \neq j$.

2.5. Summary and Conclusions

HFR methods explicitly or implicitly make design decisions about three stages of representation, cross-domain mapping and matching (Fig. 1). An important factor in the strengths and weaknesses of each approach arises from the use of supervised training in either or both of the latter two stages (Fig. 2).

Use of training data An important property of HFR systems is whether annotated cross-modal training data is required/exploited. This has practical consequences about whether an approach can be applied in a particular application, and its expected performance. Since a large dataset of annotated cross-modal pairs may not be available, methods that require no training data (most feature-engineering and NN matching approaches [29, 30, 33, 34]) are advantageous.

On the other hand, exploiting available annotation provides a critical advantage to learn better cross-domain mappings, and many discriminative matching approaches. Methods differ in how strongly they exploit available supervision. For example CCA tries to find the subspace

where cross-modal pairs are most similar [10, 28]. In contrast, LDA simultaneously finds a space where cross-modal pairs are similar and also where different identities are well separated [25], which exploits the labeled training data more thoroughly. It is worth noting that since HFR is concerned with addressing the cross-modal gap, most approaches using training data make use of cross-domain matching pairs as annotated training data, rather than person identity annotations that are more common in conventional (within-domain) face recognition.

Heterogeneous Feature Spaces A second important model-dependent property is whether the model can deal with heterogeneous data dimensions. In some cross-modal contexts (photo-sketch, VIS-NIR), while the data distribution is heterogeneous, the data dimensions can be the same; while in 2D-3D or low-high, the data dimensionality may be fundamentally different. In the latter case approaches that require homogeneous dimensions such as LDA may not be applicable, while others such as CCA and PLS can still apply.

3. Matching facial sketches to images

The problem of matching facial sketches to photos is commonly known as sketch-based face recognition (SBFR). It typically involves a gallery dataset of visible light images and a probe dataset of facial sketches. An important application of SBFR is assisting law enforcement to identify suspects by retrieving their photos automatically from existing police databases. Over the past decades, it has been accepted as an effective tool in law reinforcement. In most cases, actual photos of suspects are not available, only sketch drawings based on the recollection of eyewitnesses. The ability to match forensic sketches to mug shots not only has the obvious benefit of identifying suspects, but moreover allows the witness and artist to interactively refine the sketches based on similar photos retrieved [25].

SBFR can be categorized based on how the sketches are generated, as shown in Fig. 3: (i) viewed sketches, where artists are given mugshots as reference, (ii) forensic sketches, where sketches are hand-drawn by professional artists based on recollections of witnesses, (iii) composite sketches, where rather than hand-drawn they were produced using specific software, and (iv) caricature sketches, where facial features are exaggerated.

The majority of existing SBFR studies focused on recognizing viewed hand drawn sketches. This is not a realistic use case – a sketch would not be required if a photo of a suspect is readily available. Yet studying them is a middle ground toward understanding forensic sketches – viewed sketch performance should reflect forensic sketch performance in the ideal case when all details are remembered and communicated correctly. Research can then focus on making good viewed sketch methods robust to lower-quality forensic sketches.

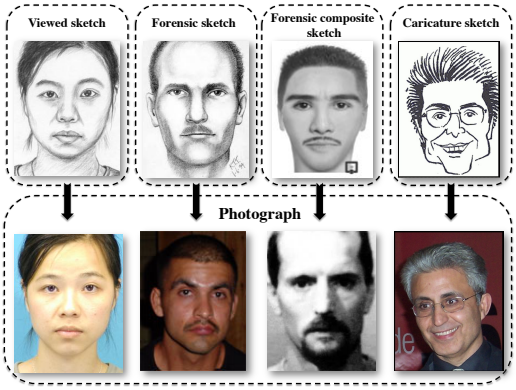

Figure 3: Facial sketches and corresponding mugshots: viewed sketch, forensic hand drawn sketch, forensic composite sketch, caricature sketch and their corresponding facial images

3.1. Categorization of facial sketches

Facial sketches can be created either by an artist or by software, and are referred to as *hand-drawn* and *composite* respectively. Meanwhile depending on whether the artist observes the actual face before sketching, they can also be categorized as *viewed* and *forensic* (unviewed). Based on these factors, we identify four typically studied categories of facial sketches:

- **Forensic hand drawn sketches**: These are produced by a forensic artist based on the description of a witness ([71]), as illustrated in the second column of Fig. 3. They have been used by police since the 19th century, however they have been less well studied by the recognition community.

- **Forensic composite sketches**: They are created by computer software (Fig. 4) with which a trained operator selects various facial components based on the description provided by a witness. An example of a resulting composite sketch is shown in the third column of Fig. 3. It is reported that 80% of law enforcement agencies use some form of software to create facial sketches of suspects [72]. The most widely used software for generating facial composite sketches are IdentiKit [70], Photo-Fit [73], FACES [69], Mac-a-Mug [73], and EvoFIT [74]. It is worth nothing that due to the limitations of such software packages, less facial detail can be presented in composite sketches compared with hand-drawn sketches.

- **Viewed hand drawn sketches**: In contrast to forensic sketches that are unviewed, these are sketches drawn by artists by while looking at a corresponding photo, as illustrated in the first column of Fig. 3. As such, they are the most similar to the actual photo.

- **Caricature**: In contrast to the previous three categories, where the goal is to render the face as accurately as possible, caricature sketches are purposefully dramatically exaggerated. This adds a layer of

(a)FACES 4.0 [69] (b)FACES [69] (c)IdentiKit [70]

Figure 4: Examples of different kind of composite sketch softwares

abstractness that makes their recognition by conventional systems much more difficult. See fourth column of Fig. 3 for an example. However, they are interesting to study because they allow the robustness of SBFR systems to be rigorously tested, and because there is evidence that humans remember faces in a caricatured form, and can recognize them even better than accurate sketches [4, 75, 76].

3.2. Facial sketch datasets

There are five commonly used datasets for benchmarking SBFR systems. Each contains pairs of sketches and photos. They differ by size, whether sketches are viewed and if drawn by artist or composited by software. Tab. 2 summaries each dataset in terms of these attributes.

CUHK Face sketch dataset (CUFS) [25] is widely used in SBFR. It includes 188 subjects from the Chinese University of Hong Kong (CUHK) student dataset, 123 faces from the AR dataset [79], and 295 faces from the XM2VTS dataset [80]. There are 606 faces in total. For each subject, a sketch and a photo are provided. The photo is taken of each subject with frontal pose and neutral expression under normal lighting conditions. The sketch is then drawn by an artist based on the photo.

CUHK Face Sketch FERET Dataset (CUFSF) [39, 25] is also commonly used to benchmark SBFR algorithms. There are 1,194 subjects from the FERET dataset [81]. For each subject, a sketch and a photo is also provided. However, compared to CUFS, instead of normal light condition, the photos in CUFSF are taken with lighting variation. Meanwhile, the sketches are drawn with shape exaggeration based on the corresponding photos. Hence, CUFSF is more challenging and closer to practical scenarios [39].

The IIIT-D Sketch Dataset [32] is another well known facial sketch dataset. Unlike CUFS and CUFSF, it contains not only viewed sketches but also semi-forensic sketches and forensic sketches, therefore can be regarded as three separate datasets each containing a particular type of sketches, namely IIIT-D viewed, IIIT-D semi-forensic and IIIT-D forensic sketch dataset. IIIT-D viewed sketch dataset comprises a total of 238 sketch-image pairs. The sketches are drawn by a professional sketch artist based on photos collected from various sources. It comprises of 67 sketch-image pairs from the FG-NET aging dataset[4], 99 sketch-digital image from Labeled Faces in Wild (LFW) dataset [82], and 72 sketch-digital image pairs from the IIIT-D student & staff dataset [82]. In the IIIT-D semi-forensic dataset, sketches are drawn based on an artist's memory instead of directly based on the photos or the description of an eye-witness. These sketches are termed semi-forensic sketches. The semi-forensic dataset is based on 140 digital images from the Viewed Sketch dataset. In the IIIT-D forensic dataset there are 190 forensic sketches and face photos. It contains 92 and 37 forensic sketch-photo pairs from [83] and [84] respectively, as well as 61 pairs from various sources on the internet.

The Pattern Recognition and Image Processing (PRIP) Viewed Software-Generated Composite (PRIP-VSGC) database [34] contains 123 subjects from AR database. For each photograph, three composites were created. Two of composites are created using FACES [69] and the third was created using Identi-Kit [70].

The Pattern Recognition and Image Processing (PRIP) Hand-Drawn Composite (PRIP-HDC) database [77] includes 265 hand-drawn and composite facial sketches, together with corresponding mugshots. Those facial sketches are drawn based on the verbal description by the eyewitness or victim. Among all those facial sketches, 73 were drawn by Lois Gibson, 43 were provided by Karen Taylor, 56 were provided by the Pinellas County Sheriff's Office (PCSO), 46 were provided by Michigan State Police, and 47 were downloaded from the Internet. So far, only those 47 facial sketches collected from Internet are publicly available.

All sketches collected by previous attempts are coarsely grouped as either viewed or unviewed, without tracking the time-delay between viewing and forensic sketching – a factor that has critical impact on the fidelity of human facial memory [85]. To address this [78] introduce the first Memory Gap Database which not only includes viewed and unviewed sketch, but uniquely sketches rendered at different time-delays between viewing and sketching. Memory Gap Database (MGDB) [78] includes 100 real subjects

[4]Downloadable at http://www-prima.inrialpes.fr/FGnet/html/home.html

Table 2: Existing facial sketch benchmark datasets.

Datasets	Pairs of Sketch/Photo	Viewed or Forensic	Composite or Hand drawn	Availability
CUFS [25]	606	Viewed	Hand drawn	CUHK: Free to download AR: Request permission XM2VTS: Pay a fee
CUFSF [39, 25]	1,194	Viewed	Hand drawn	Sketch: Free to download Photo: Request permission
IIIT-D viewed sketch [32]	238	Viewed	Hand drawn	Request permission
IIIT-D semi-forensic sketch [32]	140	Semi-Forensic	Hand drawn	Request permission
IIIT-D forensic sketch [32]	190	Forensic	Hand drawn and Composite	Request permission
PRIP-VSGC database[34]	123	Viewed	Composite	Free to download
PRIP-HDC database [77]	265	Forensic	Hand drawn	Part (47) of free to download
Memory Gap Database [78]	100	Forensic/Viewed	Hand drawn	Request permission

(mugshots sampled from mugshot.com). Each subject has frontal face photo and four facial sketches drawn at various time-delays: viewed sketch, 1 hour sketch, 24 hour sketch and unviewed sketches. In total, 400 hand-drawn sketches are provided by the MGDB. This database is aimed to help modellers disentangle modality, memory, and communication factors in forensic sketch HFR.

It is worth noting that the accessibility of these datasets varies, with some not being publicly available. [31] created a forensic dataset from sketches cropped from two books (also contained in IIIT-D forensic), which is thus limited by copyright. Klare et al. also conducted experiments querying against a real police database of 10,000 mugshots, but this is not publicly available.

3.3. Viewed sketch face recognition

Viewed sketch recognition is the most studied subproblem of SBFR. Although a hypothetical problem (in practice a photo would be used directly if available, rather than a viewed sketch), it provides an important step toward ultimately improving forensic sketch accuracy. It is hypothesized that based on an ideal eyewitness description, unviewed sketches would be equivalent to viewed ones. Thus performance on viewed sketches should be an upper bound on expected performance on forensic sketches.

Viewed sketch-based face recognition studies can be classified into synthesis, projection and feature-based methods according to their main contribution to bridging the cross-modal gap.

3.3.1. Synthesis-based approaches

The key strategy in synthesis-based approaches is to synthesize a photo from corresponding sketch (or vice-versa), after which traditional homogeneous recognition methods can be applied (see Fig. 5). To convert a photo into a sketch, [7] propose an eigensketch transformation approach, wherein a new sketch is constructed using a linear combination of training sketch samples, with linear coefficients obtained from corresponding photos via eigen decomposition. Classification is then accomplished by the obtained eigensketch features. To exploit the strong correlation exists among face images, the Karhunen-Loeve Transform (KLT) is applied to represent and recognise faces. The eigensketch transformation algorithm reduced the discrepancies between photo and sketch. The resulting rank-10 accuracy is reasonable. However, the work lacks in the small size of the dataset (188 pairs) used and weak rank-1 accuracy.

It was soon discovered that synthesizing facial sketches holistically via linear processes might not be sufficient, in that synthesized sketches lack details which will in turn negatively impact final matching accuracy. Liu et al. [26] proposed a Local Linear Embedding (LLE) inspired method to convert photos into sketches based on image patches, rather than holistic photos. For each image patch to be converted, it finds the nearest neighbors in the training set. Reconstruction weights of neighbouring patches are then computed, and used to generate the final synthesized patch. Wang and Tang [25] further improved [26] by synthesizing local face structures at different scales using Markov Random Fields (MRF), as shown in Fig. 5(a). By modelling the relationship between local patches through a compatibility function, the multi-scale MRF jointly reasons the selection of the sketch patch corresponding to each photo patch during photo-sketch conversion. In each case photos/sketche conversion reduces the modality gap, allowing the two domains to be matched effectively. In both [7] and [25], after photos/sketches are synthesized, many standard methods like PCA [76], Bayesianface [86], Fisherface [87], null-space LDA [88], dual-space LDA [89] and Random Sampling LDA (RS-LDA) [90, 91] are straightforwardly applied for homogeneous face recognition.

The embedded hidden Markov model (E-HMM) is applied by Zhong et al. [92] to transform a photo to a sketch. The nonlinear relationship between a photo/sketch pair is modeled by E-HMM. Then, learned models are used to generate a set of pseudo-sketches. Those pseudo-sketches

Figure 5: Examples of sketch synthesis: (Left) photo to sketch by synthesized sketches (Right) sketch to photo by synthesized photos

are used to synthesize a finer face pseudo-sketch based on a selective ensemble strategy. E-HMMs are also used by Gao et al. [93, 94] to synthesis sketches from photos. On the contrary, Xiao et al. [95] proposed a E-HMM based method to synthesis photos from sketches. Liu et al. [96] proposed a synthesis method based on Bayesian Tensor Inference. This method can be used to synthesize both sketches from photos and photos from sketches.

A common problem shared by most sketch synthesis methods is that they can not handle non-facial factors such as hair style, hairpins and glasses well. To tackle this problem, Zhang et al. [97] combined sparse representation and bayesian inference in synthesizing facial sketches. Sparse representation is used to model photo patches, where nearest neighbor search with learned prior knowledge is applied to compute similarity scores across patches. After selecting candidate sketch patches using these similarity scores, MRF is employed to reconstruct the final sketch by calculating the probability between photo patches and candidate sketch patches.

Most sketch synthesis methods rely on many training pairs to work, which naturally makes them deficient in modelling subtle non-facial features. Zhang et al. [98] recognised this and proposed a method that is capable of handling non-facial factors only using a single photo-sketch pair. Sparse representation based greedy search is used to select candidate patches and bayesian inference is then used for finalise sketch synthesis. A cascaded image synthesis strategy is further applied to improve the quality of the synthesized sketch.

All aforementioned methods synthesize facial sketches using pixel intensities alone. Peng et al. [99] explored a multi-representation approach to face sketch modelling. Filters such as DoG, and features like SURF and LBP are employed to generate different representations and a Markov network is deployed to exploit the mutual relationship among neighbouring patches. They conduct forensic sketch recognition experiments using sketches from CUHK and AR datasets as probe, and 10,000 face photo images from LFW-a dataset as gallery.

3.3.2. Projection based approaches

Rather than trying to completely reconstruct one modality from the other as in synthesis-based approaches; projection-based approaches attempt to find a lower-dimensional sub-space in which the two modalities are directly comparable (and ideally, in which identities are highly differentiated).

Lin and Tang [40] proposed a linear transformation which can be used between different modalities (sketch/photo, NIR/VIS), called common discriminant feature extraction (CDFE). In this method, images from two modalities are projected into a common feature space in which matching can be effectively performed.

Sharma et al. [9] use Partial Least Squares (PLS) to linearly map images of different modalities (e.g., sketch, photo and different poses, resolutions) to a common subspace where mutual covariance is maximized. This is shown to generalize better than CCA. Within this subspace, final matching is performed with simple NN.

In [46], a unified sparse coding-based model for coupled dictionary and feature space learning is proposed to simultaneously achieve synthesis and recognition in a common subspace. The learned common feature space is used to perform cross-modal face recognition with NN.

In [26] a kernel-based nonlinear discriminant analysis (KNDA) classifier is adopted by Liu et al. for sketch-photo recognition. The central contribution is to use the nonlinear kernel trick to map input data into an implicit feature space. Subsequently, LDA is used to extract features in that space, which are non-linear discriminative features of the input data.

3.3.3. Feature based approaches

Rather mapping photos into sketches, or both into a common subspace; feature-based approaches focus on designing a feature descriptor for each image that is intrinsically invariant to the modality, while being variant to the identity of the person. The most widely used image feature descriptors are Scale-invariant feature transform (SIFT), Gabor transform, Histogram of Averaged Oriented Gradients (HAOG) and Local Binary Pattern (LBP). Once sketch and photo images are encoded using these descrip-

tors, they may be matched directly, or after a subsequent projection-based step as in the previous section.

Klare et al. [31] proposed the first direct sketch/photo matching method based on invariant SIFT-features [100]. SIFT features provide a compact vector representation of an image patch based on the magnitude, orientation, and spatial distribution of the image gradients [31]. SIFT feature vectors are first sampled uniformly from the face images and concatenated together separately for sketch and photo images. Then, Euclidean distances are computed between concatenated SIFT feature vectors of sketch and photo images for NN matching.

Later on, Bhatt et al. [101] proposed an method which used extended uniform circular local binary pattern descriptors to tackle sketch/photo matching. Those descriptors are based on discriminating facial patterns formed by high frequency information in facial images. To obtained the high frequency cues, sketches and photos are decomposed into multi-resolution pyramids. After extended uniform circular local binary pattern based descriptors are computed, a Genetic Algorithm (GA) [102] based weight optimization technique is used to find optimum weights for each facial patch. Finally, NN matching is performed by using weighted Chi square distance measure.

Khan et al. [33] proposed a self-similarity descriptor. Features are extracted independently from local regions of sketches and photos. Self-similarity features are then obtained by correlating a small image patch within its larger neighborhood. Self-similarity remains relatively invariant to the photo/sketch-modality variation therefore reduces the modality gap before NN matching.

A new face descriptor, Local Radon Binary Pattern (LRBP) was proposed by Galoogahi et al. [103] to directly match face photos and sketches. In the LRBP framework, face images are first transformed into Radon space, then transformed face images are encoded by Local Binary Pattern (LBP). Finally, LRBP is computed by concatenating histograms of local LBPs. Matching is performed by a distance measurement based on Pyramid Match Kernel (PMK) [104]. LRBP benefits from low computational complexity and the fact that there is no critical parameter to be tuned [103].

Zhang et al. [105] introduced another face descriptor based on coupled information-theoretic encoding which uniquely captures discriminative local facial structures. Through maximising mutual information between photos and sketches in the quantised feature spaces, they obtained a coupled encoding using an information-theoretic projection tree. The method was evaluated with 1,194 faces sampled from the FERET database.

Galoogahi et al. consequently proposed another two face descriptors: Gabor Shape [30] which is variant of Gabor features and Histogram of Averaged Oriented Gradient (HAOG) features [29] which is variant of HOG for sketch/photo directly matching, the latter achieves perfect 100% accuracy on the CUFS dataset.

Klare et al. [31] further exploited their SIFT descriptor, by combining it with a 'common representation space' projection-based strategy. The assumption is that even if sketches and photos are not directly comparable, the distribution of *inter-face similarities* will be similar within the sketch and photo domain. That is, the (dis)similarity between a pair of sketches will be roughly the same as the (dis)similarity between the corresponding pair of photos. Thus each sketch and photo is re-encoded as a vector of their euclidean distances to the training set of sketches and photos respectively. This common representation should now be invariant to modality and sketches/photos can be compared directly. To further improve the results, direct matching and common representation matching scores are fused to generate the final match [31]. The advantage of this approach over mappings like CCA and PLS is that it does not require the sketch-photo domain mapping to be linear. The common representation strategy has also been used to achieve cross-view person recognition [107], where it was shown to be dependent on sufficient training data.

In contrast to the previous methods which are appearance centric in their representation, Pramanik et al. [6] evaluate an analytic geometry feature based recognition system. Here, a set of facial components such as eyes, nose, eyebrows, lips, are extracted their aspect ratio are encoded as feature vectors, followed by K-NN as classifier.

Overall, because viewed sketches and photos in the CUFS database are very well-aligned and exaggeration between photo and sketch is minimal, appropriate feature engineering, projection or synthesis approaches can all deliver near-perfect results, as shown in Tab. 3.

3.4. Forensic sketch face recognition

Forensic sketches pose greater challenge than viewed sketch recognition because, beyond modality shift, they contain incomplete or inaccurate information due to the subjectivity of the description, and imperfection of the witness' memory [85].

Due to its greater challenge, and the lesser availability of forensic sketch datasets, research in this area has been less than for viewed sketches. Uhl et al. [108] proposed the first system for automatically matching police artist sketches to photographs. In their method, facial features are first extracted from sketches and photos. Then, the sketch and photo are geometrically standardized to facilitate comparison. Finally, eigen-analysis is employed for matching. Only 7 probe sketches were used in experimental validation, their method is antiquated with respect to modern methods. Nonetheless, Uhl and Lobo's study highlighted the complexity and difficulty in forensic sketch based face recognition and drew other researchers towards forensic sketch-based face recognition.

Klare et al. [109] performed the first large scale study in 2011, with an approach combining feature-based and projection-based contributions. SIFT and MLBP features were extracted, followed by training a LFDA projection to minimize the distance between corresponding sketches

Table 3: Sketch-Photo matching methods: Performance on benchmark datasets.

Method	Publications	Recognition Approach	Dataset	Feature	Train:Test	Accuracy
Synthesis based	[7]	KLT	CUHK	Eigen-sketch features	88:100	about 60%
	[26]	KNDA	CUFS		306:300	88%
	[25]	RS_LDA	CUFS	Multiscale MRF	306:300	96%
	[92]		CUFS	E-HMM	—	95%
	[96]		CUFS	E-HMM+Selective ensemble	—	100%
	[97]		CUHK/XM2VTS	Sparse representations	—	—
	[98]		CUHK/XM2VTS	Single sketch-photo pair	—	—
	[99]	Fisherface	CUFS/IIIT-D	Multiple representation+Markov Network	88:100	98.3%
Projection based	[31]	Common representation	CUFS	SIFT	100:300	96%
	[9]	PLS	CUHK		88:100	93%
	[106]	PLS regression	CUFS,CUFSF	Gabor and CCS-POP	0:1800	99%
Feature based	[31]	NN	CUFS	SIFT	100:300	98%
	[33]	NN	CUFS	Self Similarity	161:150	99%
	[105]	PCA+LDA	CUFSF	CITE	500:694	99%
	[30]	NN,Chi-square	CUFS	Gabor Shape	306:300	99%
	[101]	Weighted Chi-square	CUFS	EUCLBP	78:233	94%
	[29]	NN,Chi-square	CUFS	HAOG	306:300	100%
	[30]	NN,Chi-square	CUFSF	Gabor Shape	500:694	96%
	[103]	NN,PMK,Chi-square	CUFSF	LRBP	—	91%
	[103]	NN,PMK,Chi-square	CUFSF	LRBP	—	91%
	[6]	K-NN	CUHK	Geometric features	108:80	80%
	[101]	Weighted Chi-square	IIIT-D	EUCLBP	58:173	79%

and photos while maximizing the distance between distinct identities. They analyse a dataset of 159 pairs of forensic hand drawn sketches and mugshot photos. The subjects in this dataset were identified by the law enforcement agencies. They also included 10,159 mugshot images provided by Michigan State Police to better simulate a realistic police search against a large gallery. With this realistic scenario, they achieved about 15 percent success rate.

To improve recognition performance, Bhatt et al. [32] proposed an algorithm that also combines feature and projection-based contributions. They use multi-scale circular Webber's Local descriptor to encode structural information in local facial regions. Memetic optimization was then applied to every local facial region as a metric learner to find the optimal weights for Chi squared NN matching [32]. The result outperforms [109] using only the forensic set as gallery.

Different to previous studies that tackle forensic sketch matching using a single model, Ouyang et al. [78] developed a database and methodology to decouple the multiple distinct challenges underlying forensic matching: the modality change, the eyewitness-artist communication, and the memory loss of the eyewitness. Their MGDB has 400 sketches created under different conditions such as memory time-delays. Using this MGDB, they applied multi-task Gaussian process regression to synthesise facial sketches accounting for each of these factors. They evaluated this model on IIIT-D forensic sketch and a large (10,030) mugshot database similar to that used in [109] and achieved state-of-the-art results.

3.5. Composite sketch based face recognition

Several studies have now considered face recognition using composite sketches. The earliest used both local and global features to represent sketches and is proposed by Yuen et al. [5]. This method also investigated user input in the form of relevance feedback in the recognition phase. Studies have focused on holistic [110, 24] component based [34, 56, 111, 57] and hybrid [5, 77] representations respectively.

The holistic method [110] uses similarities between local features computed on uniform patches across the entire face image. Following tessellating a facial sketch/mugshot into 154 uniform patches, SIFT [100] and multi-scale local binary pattern (MLBP) [112] invariant features are extracted from each patch. With this feature encoding, as improved version of the common representation intuition from [31] is applied, followed by RS-LDA [91] to generate a discriminative subspace for NN matching with cosine distance. The scores generated by each feature and patch are fused for final recognition.

In contrast, the component based method [34] uses similarities between individual facial components to compute

an overall sketch to mughsot match score. Facial landmarks in composite sketches and photos are automatically detected by an active shape model (ASM) [113]. Mutiscale local binary patterns (MLBPs) are then applied to extract features of each facial component, and similarity is calculated for each component: using histogram intersection distance for the component's appearance and cosine distance for its shape. The similarity scores of each facial component are normalized and fused to obtain the overall sketch-photo similarity. [56] also used features extracted according to facial landmarks. Daisy descriptors were extracted from patches centred on facial landmarks. The cross-modal Chi square distances of these descriptors at each landmark are then used as the input feature to train a binary verifier based on GentleBoost. This was improved by a subsequent study [57] which improved the representation by using Self Similarity Descriptors (SSD) as features, followed by encoding them in terms of distance to a dictionary of faces (analogously to the common representation in [31]) – before applying GentleBoost verification again.

In contrast to engineered descriptors, deep learning can provide an effective way to learn discriminative and robust representations. However these methods tend to require large data volumes relative to the size of available HFR datasets. To address this [24] use deep auto encoders and deep belief networks to learn an effective face representation based on a large photo database. This is then fine-tuned on a smaller heterogeneous database to adapt it to the HFR task. Binary verification is then performed using SVM and NN classifiers.

Finally, [77] focuses on building a practically accurate, efficient and deployable sketch-based interaction system by improving and fusing the holistic and component-based algorithms in [110] and [34] respectively. The implication of different sources of training data is also investigated.

3.6. Caricature based face recognition

The human visual system's ability to recognise a person from a caricature is remarkable, as conventional face recognition approaches fail in this setting of extreme intra-class variability (Fig. 6). The caricature generation process can be conceptualised as follows: If we assume a face space in which each face lies. Then by drawing a line to connect the mean face to each face, the corresponding caricature will lie beyond that face along the line. That is to say, a caricature is an exaggeration of a face away from the mean [114].

Studies have suggested that people may encode faces in a caricatured manner [115]. Moreover they may be more capable of recognizing a familiar person through a caricature than an accurate rendition [116, 117]. The effectiveness of a caricature is due to its emphasis of deviations from average faces [54]. Developing efficient approaches in caricature based face recognition could help drive more robust and reliable face and heterogeneous face recognition systems.

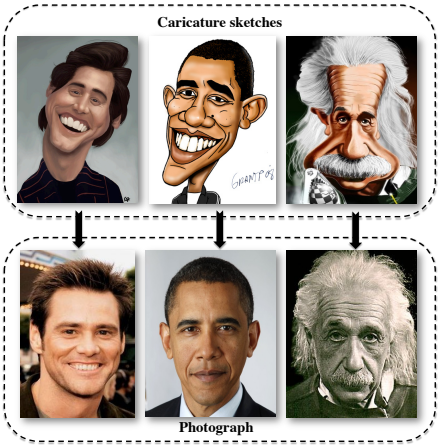

Figure 6: Caricatures and corresponding mugshots

Klare et al. [54] proposed a semi-automatic system to match caricatures to photographs. In this system, they defined a set of qualitative facial attributes that describe the appearance of a face independently of whether it is a caricature or photograph. These mid-level facial features were manually annotated for each image, and used together with automatically extracted LBP [112] features. These two feature types were combined with an ensemble of matching methods including NN and discriminatively trained logistic regression SVM, MKL and LDA. The results showed that caricatures can be recognized slightly better with high-level qualitative features than low-level LBP features, and that they are synergistic in that combining the two can almost double the performance up to 22.7% rank 1 accuracy. A key insight here is that – in strong contrast to viewed sketches that are perfectly aligned – the performance of holistic feature based approaches is limited because the exaggerated nature of caricature sketches means that detailed alignment is impossible.

A limitation of the above work is that the facial attributes must be provided, requiring manual intervention at run-time. Ouyang et al. [118] provided a fully automated procedure that uses a classifier ensemble to robustly estimate facial attributes separately in the photo and caricature domain. These estimated facial attributes are then combined with low-level features using CCA to generate a robust domain invariant representation that can be matched directly. This study also contributed facial attribute annotation datasets that can be used to support this line of research going forward.

3.7. Summary and Conclusions

Tab. 3 summarizes the results of major studies in terms of distance metric, dataset, feature representation, train to test ratio, and rank-1 accuracy, of feature-based and projection-based approaches respectively[5]. As viewed

[5]Note that some results on the same dataset are not directly comparable because of differing test set sizes.

sketch datasets exhibit near perfect alignment and detail correspondence between sketches and photos, well designed approaches of any type achieve near perfect accuracies. Forensic sketch in contrast is an open problem, but the fewer and less comparable studies here also makes it hard to identify the most promising techniques. What seems clear is that representations assuming simple perfect correspondence such as dense-HOG and simple linear projections are unlikely to be the answer, and that purely image-processing approaches may be significantly improved by understanding the involved human factors[78].

Methodologies. All three categories of approaches – synthesis, projection and discriminative features – have been well studied for SBFR. Interestingly, while synthesis approaches have been one of the more popular categories of methods, they have only been demonstrated to work in viewed-sketch situations where the sketch-photo transformation is very simple and alignment is perfect. It seems unlikely that they can generalize effectively to forensic sketches, where the uncertainty introduced by forensic process (eyewitness subjective memory) significantly completes the matching process.

An interesting related issue that has not been systematically explored by the field is the dependence on the sketching artists. Al Nizami et al. [119] demonstrated significant intra-personal variation in sketches drawn by different artists. This may challenge systems that rely on learning a single cross-modal mapping. This issue will become more significant in the forensic sketch case where there is more artist discretion, than in viewed-sketches which are more like copying exercises.

Challenges and Datasets. The majority of SBFR research has focused on viewed sketch-based recognition, with multiple studies now achieving near-perfect results on the CUFS dataset. This is due to the fact that viewed sketches are professionally rendered copies of photographed faces, and thus close in likeness to real faces, so non-linear misalignment and all the attendant noise introduced by verbal descriptions communicated from memory are eliminated. This point is strongly made by Choi et al. [106], who criticize the existing viewed-sketch datasets and the field's focus on them. They demonstrate that with minor tweaks, an off the shelf PLS-based *homogeneous* face recognition system can outperform existing cross-modality approaches and achieve perfect results on the CUFS dataset. They conclude that existing viewed-sketch datasets are unrealistically easy, and not representative of realistic forensic sketch scenarios.

It is thus important that the field should move to more challenging forensic, composite and caricature sketches with more realistic non-linear misalignment and heteroskedastic noise due to the forensic process. This will reveal whether current state of the art methods from viewed-sketches are indeed best, or are brittle to more realistic data; and will drive the generation of new insights, methods and practically relevant capabilities. Research here, although less mature, has begun to show promising results. However, it has been hampered by lack of readily obtainable forensic datasets. Constructing realistic and freely available datasets should be a priority [106], and is beginning to happen [78].

Training Data Source. Many effective SBFR studies have leveraged annotated training data to learn projections and/or classifiers [31]. As interest has shifted to forensic sketches, standard practice has been to train such models on viewed-sketch datasets and test on forensic datasets [109]. An interesting question going forward is whether this is the best strategy. The first study explicitly addressing this issue concluded that it may not be [77]. Since viewed-sketches under-represent sketch-photo heterogeneity, this means that learning methods are learning a model that is not matched to the data (forensic sketches) that they will be tested on. This poses an additional challenge of *domain shift* [120] (photo/viewed→photo/unviewed), to be solved. This issue also further motivates the creation of larger forensic-sketch datasets for training, which will be necessary to thoroughly investigate the best training strategy.

Automated Matching versus Human Recognition. Finally we notice that the vision and biometrics communities have largely focused on automated cross-domain matching, while an important outstanding question in forensic sketch for law enforcement has been left largely un-studied [85]. Rather than cross-domain mapping for HFR matching of a sketch against a photo database, police are often interested in generating a sketch/photo which *can be best recognised by a person* who might be familiar with the suspect; rather than generating photo that can be matched to a mugshot database by a machine. From a cross-domain synthesis perspective, rather than simply generate the most accurate photo, the task here is to generate a more *human recognisable image*, which has a different set of requirements [121] than conventional metrics.

4. Matching NIR to Visible Light Images

NIR face recognition has attracted increasing attention recently because of its much desired attribute of (visible-light) illumination invariance, and the decreasing cost of NIR acquisition devices. It encompasses matching near infrared (NIR) to visible light (VIS) face images. In this case, the VIS enrollment samples are images taken under visible light spectrum (wavelength range $0.4\mu m - 0.7\mu m$), while query images are captured under near infrared (NIR) condition (just beyond the visible light range, wavelengths between $0.7\mu m$ - $1.4\mu m$) [41]. NIR images are close enough to the visible light spectrum to capture the structure of the face, while simultaneously being far enough to be invariant

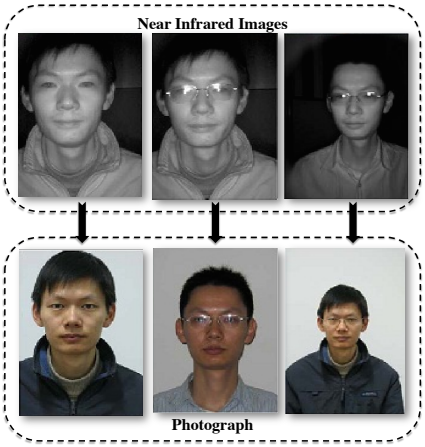

Figure 7: VIS and NIR face images.

to visible light illumination changes. Fig. 7 illustrates differences between NIR and VIS images. Matching NIR to VIS face images is of interest, because it offers the potential for face recognition where controlling the visible environment light is difficult or impossible, such as in night-time surveillance or automated gate control.

In NIR based face recognition, similar to sketch based recognition, most studies can be categorized into synthesis, projection and discriminant feature based approaches, according to their contribution to bridging the cross-modal gap.

4.1. Datasets

There are five main heterogeneous datasets covering the NIR-VIS condition. The CASIA HFB dataset [122], composed of visual (VIS), near infrared (NIR) and 3D faces, is widely used. In total, it includes 100 subjects: 57 males and 43 females. For each subject, there are 4 VIS and 4 NIR face images. Meanwhile, there are also 3D images for each subject (92 subjects: 2 for each, 8 subjects: 1 for each). In total, there are 800 images for NIR-VIS setting and 200 images for 3D studies.

CASIA NIR-VIS 2.0 [123] is another widely used NIR dataset. 725 subjects are included, with 50 images (22 VIS and 28 NIR) per subject, for a total of 36,250 images.

The Cross Spectral Dataset [124] is proposed by Goswami et al. It consists of 430 subjects from various ethnic backgrounds (more than 20% of non-European origin). At least one set of 3 poses (-10 degree / 0 degree / 10 degree) are captured for each subject. In total, there are 2,103 NIR images and 2,086 VIS images.

The PolyU NIR face dataset [125] is proposed by the biometric research center at Hong Kong Polytechnic University. This dataset includes 33,500 images from 335 subjects. Besides frontal face images and faces with expression, pose variations are also included. It is created with an active light source in the NIR spectrum between 780nm to 1,100nm.

The main NIR-VIS datasets are summarised in Tab. 4. Each column categorizes the datasets by wavelength of NIR light, no. of subject, no. of images, and whether they include 3D images, pose and expression variations, respectively.

4.2. Synthesis based approaches

Wang et al. [12] proposed an analysis-by-synthesis framework, that transforms face images from NIR to VIS. To achieve the conversion, facial textures are extracted from both modalities. NIR-VIS texture patterns extracted at corresponding regions of different face pairs collectively compose a training set of matched pairs. After illumination normalization [126], VIS images can be synthesized patch-by-patch by finding the best matching patch for each patch of the input NIR image.

Chen et al. [27] also synthesize VIS from NIR images using a similar inspiration of learning a cross-domain dictionary of corresponding VIS and NIR patch pairs. To more reliably match patches, illumination invariant LBP features are used to represent them. Synthesis of the VIS image is further improved compared to [12], by using locally-linear embedding (LLE) inspired patch synthesis rather than simple nearest-neighbor. Finally homogeneous VIS matching is performed with NN classifier on the LBP representations of the synthesized images.

Xiong et al. [127] developed a probabilistic statistical model of the mapping between two modalities of facial appearance, introducing a hidden variable to represent the transform to be inferred. To eliminate the influences of facial structure variations, a 3D model is used to perform pose rectification and pixel-wise alignment. Difference of Gaussian (DOG) filter is further used to normalize image intensities.

Recently, Xu et al. [128] introduced a dictionary learning approach for VIS-NIR face recognition. It first learns a cross-modal mapping function between the two domains following a cross-spectral joint l_0 minimization approach. Facial images can then be reliably reconstructed by applying the mapping in either direction. Experiments conducted on the CASIA NIR-VIS v2.0 database show state-of-the-art performance.

4.3. Projection based approaches

Lin et al. [40] proposed a matching method based on Common Discriminant Feature Extraction (CDFE), where two linear mappings are learned to project the samples from NIR and VIS modalities to a common feature space. The optimization criterion aims to both minimize the intra-class scatter while maximizing the inter-class scatter. They further extended the algorithm to deal with more challenging situations where the sample distribution is non-gaussian by kernelization, and where the transform is multi-modal.

After analysing the properties of NIR and VIS images, Yi et al. [10] proposed a learning-based approach for cross-modality matching. In this approach, linear discriminant

Table 4: Summary of existing NIR-VIS benchmark datasets

Dataset	Wavelength	No.of Subjects	No.of Images	3D	Pose variations	Expression variations
CASIA HFB [122]	850nm	100	992	√	×	×
CASIA NIR-VIS 2.0 [123]	850nm	725	17580	√	√	√
Cross Spectral Dataset [125]	800-1000nm	430	4189	√	√	×
PolyU [125]	780-1100nm	335	33500	√	√	√

analysis (LDA) is used to extract features and reduce the dimension of the feature vectors. Then, a canonical correlation analysis (CCA) [129] based mechanism is learned to project feature vectors from both modalities into CCA subspaces. Finally, nearest-neighbor with cosine distance is used matching score.

Both of methods proposed by Lin and Yi tend to overfit to training data. To overcome this, Liao et al. [11] present a algorithm based on learned intrinsic local image structures. In training phase, Difference-of-Gaussian filtering is used to normalize the appearance of heterogeneous face images in the training set. Then, Multi-scale Block LBP (MB-LBP) [130] is applied to represent features called Local Structure of Normalized Appearance (LSNA). The resting representation is high-dimensional, so Adaboost is used for feature selection to discover a subset of informative features. R-LDA is then applied on the whole training set to construct a discriminative subspace. Finally, matching is performed with a verification-based strategy, where cosine distance between the projected vectors is compared with a threshold to decide a match.

Klare et al. [41] build on [11], but improve it in a few ways. They add HOG to the previous LBP descriptors to better represent patches, and use an ensemble of random LDA subspaces [41] learn a shared projection with reduced over fitting. Finally, NN and Sparse Representation based matching are performed for matching.

Lei et al. [19] presented a method to match NIR and VIS face images called Coupled Spectral Regression (CSR). Similar to other projection-based methods, they use two mappings to project the heterogeneous data into a common subspace. In order to further improve the performance of the algorithm (efficiency and generalisation), they use the solutions derived from the view of graph embedding [131] and spectral regression [132] combined with regularization techniques. They later improve the same framework [20], to better exploit the cross-modality supervision and sample locality.

Huang et al. [133] proposed a discriminative spectral regression (DSR) method that maps NIR/VIS face images into a common discriminative subspace in which robust classification can be achieved. They transform the subspace learning problem into a least squares problem. It is asked that images from the same subject should be mapped close to each other, while these from different subjects should be as separated as possible. To reflect category relationships in the data, they also developed two novel regularization terms.

Yi et al. [47] applied Restricted Boltzmann Machines (RBMs) to address the non-linearity of the NIR-VIS projection. After extracting Gabor features at localised facial points, RBMs are used to learn a shared representation at each facial point. These locally learned representations are stacked and processed by PCA to yield a final holistic representation.

4.4. Feature based approaches

Zhu et al. [21] interpret the VIS-NIR problem as a highly illumination-variant task. They address it by designing an effective illumination invariant descriptor, the logarithm gradient histogram (LGH). This outperforms the LBP and SIFT descriptors used by [11] and [41] respectively. As a purely feature-based approach, no training data is required.

Huang et al. [134], in contrast to most approaches, perform feature extraction after CCA projection. CCA is used to maximize the correlations between NIR and VIS image pairs. Based on low-dimensional representations obtained by CCA, they extract three different modality-invariant features, namely, quantized distance vector (QDV), sparse coefficients (SC), and least square coefficients (LSC). These features are then represented with a sparse coding framework, and sparse coding coefficients are used as the encoding for matching.

Goswami et al. [124] introduced a new dataset for NIR/VIS (VIS/NIR) face recognition. To establish baselines for the new dataset they compared a series of photometric normalization techniques, followed by LBP-based encoding and LDA to find an invariant subspace. They compared classification with Chi-squared and Cosine as well as establishing a logistic-regression based verification model that obtained the best performance by fusing the weights from each of the model variants.

Gong and Zheng [135] proposed a learned feature descriptor, that adapts parameters to maximize the correlation of the encoded face images between two modalities. With this descriptor, the within-class variations can be reduced at the feature extraction stage, therefore offering better recognition performance. This descriptor outperforms classic HOG, LBP and MLBP, however unlike the others it requires training.

To tackle cross spectral face recognition, Dhamecha et al. [37] evaluated the effectiveness of a variety of HoG variants. They concluded that DSIFT with subspace LDA outperforms other features and algorithms.

Finally, Zhu et al. [38] presented a new logarithmic Difference of Gaussians (Log-DoG) feature, derived based on mathematical rather than merely empirical analysis of various features properties for recognition. Beyond this, they also present a framework for projecting to a non-linear discriminative subspace for recognition. In addition to aligning the modalities, and regularization with a manifold, their projection strategy uniquely exploits the unlabelled test data transductively.

4.5. Summary and Conclusions

Given their decreasing cost, NIR acquisition devices are gradually becoming an integrated component of everyday surveillance cameras. Combined with the potential to match people in a (visible-light) illumination independent way, this has generated increasing interest in NIR-VIS face recognition.

Tab. 5 summarizes the results of major cross-spectral studies in terms of recognition approach, dataset, feature representation, train to test ratio, and rank-1 accuracy. Results are promising, but lack of standardization in benchmarking prevents direct quantitative comparison across methods.

As with all the HFR scenarios reviewed here, NIR-VIS studies have addressed bridging the cross-modal gap with a variety of synthesis, projection and feature-based techniques. One notable unique aspect of NIR-VIS is that it is the change in illumination type that is the root of the cross-modal challenge. For this reason image-processing or physics based photometric normalization methods (e.g., gamma correction, contrast equalization, DoG filtering) are often able to play a greater role. This is because it is to some extent possible to model the cross-modal lighting change more analytically and explicitly than other HFR scenarios that must rely entirely on machine learning or invariant feature extraction methods.

5. Matching 2D to 3D

The majority of prior HFR systems work with 2D images, whether the face is photographed, sketched or composited. Owning to the 2D projection nature of these faces, such systems often exhibit high sensitivity to illumination and pose. Thus 3D-3D face matching has been of interest for some time [58]. However, 3D-3D matching is hampered in practice by the complication and cost of 3D compared to 2D equipment. An interesting variant of interest is thus the cross-modal middle ground, of using 3D images for enrollment, and 2D images for probes. This is useful, for example, in access control where enrollment is centralized (and 3D images are easy to obtain), but the access gate can be deployed with simpler and cheaper 2D equipment.

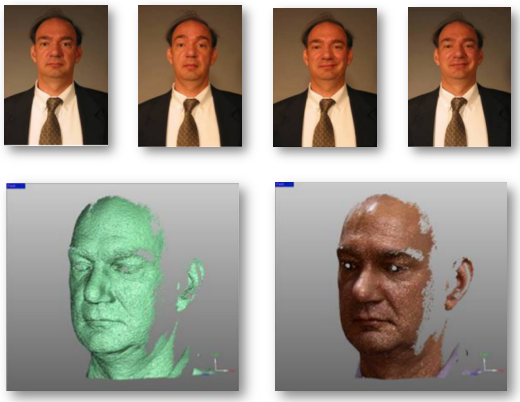

Figure 8: 2D images and 3D images from FRGC dataset.

In this case, 2D probe images can potentially be matched more reliably against the 3D enrollment model than a 2D enrollment image – if the cross-domain matching problem can be solved effectively.

A second motivation for 2D-3D HFR indirectly arises in the situation where pose-invariant 2D-2D matching is desired [136, 15, 14]. In this case the faces can be dramatically out of correspondence, so it may be beneficial to project one face to 3D in order to better reason about alignment, or synthesize a better aligned or lit image for better matching.

5.1. Datasets

The face Recognition Grand Challenge (FRGC) V2.0 dataset[6] is widely used for 2D-3D face recognition. It consists of a total of 50,000 recordings spread evenly across 6,250 subjects. For each subject, there are 4 images taken in controlled light, 2 images taken under uncontrolled light and 1 3D image. The controlled images were taken in a studio setting while uncontrolled images were taken in changing illumination conditions. The 3D images were taken by a Minolta Vivid 900/910 series sensor, including both range and texture cues. An example from the FRGC V2.0 dataset is shown in Fig. 8.

UHDB11 [137] is another popular dataset in 2D-3D face recognition. It consists of samples from 23 individuals, for each of which it has 2D high-resolution images spanning across six illumination conditions and 12 head-pose variations (72 variations in total), and a textured 3D facial mesh models. Each capture consists of both 2D images captured using a Canon DSLR camera and a 3D mesh captured by 3dMD 2-pod optical 3D system. UHDB12 [138] is an incremental update to UHDB11 [137]. 3D data were captured using a 3dMD 2-pod optical scanner, while 2D images were collected using a commercial Cannon DSLR camera. The 2D acquisition setup has six diffuse lights that vary lighting conditions. For each subject, a single

[6]Downloadable at http://www.nist.gov/itl/iad/ig/frgc.cfm

Table 5: NIR-VIS matching methods: Performance on benchmark datasets.

Method	Publications	Recognition Approach	Dataset	Feature	Train:Test	Accuracy
Synthesis based	[12]	Analysis-by-synthesis framework	Self-collected	Texture patterns	200:200	about 90%
	[27]	LLE	Self-collected	LBP	250:250	94%
	[127]	Probabilistic statistical model	CASIA HFB		200:200	40%
	[128]	Cross-Spectral joint dictionary learning	CASIA NIR-VIS 2.0		8600:6358	79%
Projection based	[40]	CDFE	Self-collected		800:64	68%
	[11]	LSNA	CASIA NIR-VIS 2.0		3464:1633	68%
	[41]	random LDA subspace	CASIA NIR-VIS 2.0	HoG + LBP	2548:2548	93%
	[19]	Coupled Spectral Regression	CASIA NIR-VIS 2.0	LBP	2549:2548	97%
	[133]	Discriminative Spectral Regression	CASIA NIR-VIS 2.0	LBP	2549:2548	95%
	[47]	Restricted Boltzmann Machines	CASIA HFB	Gabor	about 2500:2500	99%
			CASIA NIR-VIS 2.0	Gabor	–	87%
Feature based	[21]	NN	CASIA HFB	LGH	400:400	46%
	[38]	THFM	CASIA HFB	Log-DoG	400:400	99%

3D scan and 6 2D images under different lighting conditions were captured. Overall, there are 26 subjects with a total of 26 3D scans and 800 2D images. The most recent UHDB31 [14] dataset includes 3D models and facial images from 21 view points for each of the 77 subjects used. All data were captured using 21 3dMD^{TM} high resolution cameras properly set-up in a semi-sphere configuration. The average facial ROI in 2D images is around 800×600.

5.2. Synthesis based approaches

Toderici et al. [13] projected the probe 2D face image onto a normalized image space with the help of a subject-specific 3D model in the gallery. This allows them to introduce a relighting algorithm that transfers the probe image lighting to the gallery images in order to generate more accurate matching.

Wu et al. [14] compared two 3D-aided face recognition where the 3D model is used for either image normalisation or rendering. Once normalised/rendered, images are encoded with various descriptors such as HOG, LBP, etc. The conclusion is that rendering-based strategies and subsequent HOG encoding perform best.

To deal with unconstrained matching against background clutter with varied expressions, Moeini et al. [15] proposed a 3D facial expression generic elastic model (3D-FE-GEM) that reconstructs a 3D model of each human face using only a single 2D frontal image. 3D-FE-GEM improves accuracy compared to alternatives by better handling expression variation through the elastic model, and introducing a pose-robust sparse encoding of LBP for a descriptor.

Dou et al. [136] reconstruct a 3D Annoted Face Models (AFM) directly from 2D facial images using only a sparse set of 2D facial landmarks. With the help of self-occlusion masks, they are able to extract novel face signatures that were shown to be effective in matching faces in a pose-robust fashion.

5.3. Projection based approaches

Yang et al. [28] used CCA to correspond the 2D and 3D face modalities and deal with their heterogeneous dimensionality. Once projected into a common space, NN matching with Cosine distance is applied. To deal with the 2D-3D mapping being more complicated than a single linear transform, the CCA mapping is learned per-patch, and the matching scores fused at decision level.

Huang et al. [42] presented a scheme to improve results by fusing 2D and 3D matching. 2D LBP features are extracted from both the 2D image and the 2D projection of the 3D image; and then compared with Chi-squared distance. Meanwhile LBP features are also extracted from both the 2D face and 3D range image. These are mapped into a common space using CCA and compared with cosine distance. The two scores are fused at decision level, and the desired result of 2D-3D matching outperforming 2D-2D matching is demonstrated.

To further improve recognition performance, Huang et al. [44] proposed a 2D-3D face recognition approach with two separate stages: First, for 2D-2D matching, Sparse Representation Classifier (SRC) is used; Second, CCA is exploited to learn the projections between 3D and 2D face images. The two scores are again fused synergistically.

5.4. Feature based approaches

A biologically inspired feature, Oriented Gradient Maps (OGMs), is introduced by Huang et al. in [36]. OGMs simulate the complex neurons response to gradients within a pre-defined neighborhood. They have the benefit of being able to describe local texture of 2D faces and local geometry of 3D faces simultaneously. Using this feature, they are able to improve on both the 2D-2D and 2D-3D components of their previous work [42, 44].

5.5. Summary and Conclusions

2D image based face recognition systems often fail in situations where facial depictions exhibit strong pose and

Table 6: 3D-2D matching methods: Performance on benchmark datasets.

Method	Publications	Recognition Approach	Dataset	Feature	Train:Test	Accuracy
Synthesis based	[13]	Bidirectional relighting algorithm	UHDB11/UHDB12		—	91%
	[14]	Rendering/Normalisation	UHDB31	HOG, LBP, SIFT	—	83%
	[136]	3D Annotated Face Models (AFM)	UHDB11	DFD, LBP	—	94%
	[15]	3D-FE-GEM	CMU-PIE/FERET	LBP, Sparse Coding	—	—
Projection based	[28]	CCA	FRGC V2.0		172:28	87%
	[42]	CCA	FRGC V2.0	LBP	170:30	82%
	[44]	SRC+CCA	FRGC V2.0		410:410	93%
Feature based	[36]	CCA	FRGC V2.0	OGMs	3541:466	95%

illumination variations. Introducing 3D models instead naturally solves these problems since poses are fully encoded and illumination can be modeled. However, matching 3D models generally is more computational resource demanding and incurs relatively higher cost (labor and hardware) in data acquisition. 2D-3D matching is thus gaining increasing interest as a middle ground to obtain improved pose invariance, with cheaper and easier data acquisition at test time. In this area studies can be broken down into those that do some kind of explicit 3D reasoning about matching 2D probe images to 3D models [13, 139, 136, 15, 13], and others that have relied on discriminative features and learning a single cross-domain mapping such as CCA [42, 44, 36, 28]. The latter approaches are somewhat more straightforward, but to fully realize the potential pose-invariance benefits of 2D-3D matching, methods that explicitly reason about pose mapping of each test image are likely to have greater potential. Tab. 6 summarises the performance of major 2D-3D matching studies.

6. Matching low and high-resolution face images

The ability to match low-resolution (LR) to high-resolution (HR) face images has clear importance in security, forensics an surveillance. Interestingly we know this should be possible, because humans can recognize low-resolution faces down to 16×16 pixels [115]. In practice, face images with high-resolution such as mug-shots or passport photos need to be compared against low-resolution surveillance images captured at a distance by CCTV, PTZ and wearable cameras. In this case there is a dimension mismatch between the LR probe images and HR gallery images. Simple image processing upscaling the probe images, or down-scaling the HR images is a direct solution to this, but it is possible to do better.

In matching across resolution, existing approaches can be categorized into synthesis based and projection-based. Synthesis based approaches, attempt to transform LR into HR images for matching. Super-resolution [140] is used to reconstruct a HR representation of LR probe image. Then matching can be performed with any state of the art homogeneous face recognition systems. In projection-based approaches, HR gallery images and LR probes are projected into a common space in which classification is performed.

6.1. Datasets

Most LR-HR matching studies simulate LR data by downsampling HR data. SCface provides a 'natural' multi-resolution dataset [141]. It includes 4160 images of 130 subjects taken by five surveillance cameras and a high resolution SLR. The different surveillance cameras result in LR images from 144×108 to 224×168 pixels in size. Some simple PCA baselines for cross-resolution recognition are also provided.

6.2. Synthesis based approaches

Hennings-Yeomans et al. [18] presented a simultaneous super-resolution and recognition (S^2R^2) algorithm to match the low-resolution probe image to high-resolution gallery. Training this algorithm learns a super-resolution model with the simultaneous objective that the resulting images should be discriminative for identity. In followup work, they further improved the super-resolution prior and goodness of fit feature used for classification [142]. However these methods have high computational cost.

Zou et al. [48] propose a similarly inspired discriminative super resolution (DSR) approach. The relationship between the two modalities is learned in the training procedure. Then, test time procedure, the learned relationship is used to reconstruct the HR images. In order to boost the effectiveness of the reconstructed HR images, a new discriminative constraint that exploits identity information in the training data is introduced. With these, the reconstructed HR images will be more discriminative for recognition.

Zou et al. [143] proposed a nonlinear super resolution algorithm to tackle LR-HR face matching. The kernel trick is used to tractably learn a nonlinear mapping from low to high-resolution images. A discriminative regularization term is then included that requires the high-resolution reconstructions to be recognizable.

Jia et al. [144] presented a bayesian latent variable approach to LR-HR matching. Tensor analysis is exploited to perform simultaneous super-resolution and recognition. This framework also has the advantage of simultaneously addressing other covariates such as view and lighting.

Jiang et al. [49] super-resolved LR probe images by Graph Discriminant Analysis on Multi-Manifold (GDAMM), before HR matching. GDAMM exploits manifold learning, with discriminative constraints to minimize within-class scatter and maximize across-class scatter. However to learn a good manifold multiple HR samples per person are required.

Shekhar et al. [43] proposed an algorithm to address low-high resolution face recognition, while maintaining illumination invariance required for practical problems. HR training images are relighted and downsampled, and LR sparse coding dictionaries are learned for each person. At test time LR images are classified by their reconstruction error using each specific dictionary.

Huang et al. [50] proposed a nonlinear mapping approach for LR-HR matching. First, CCA is employed to align the PCA features of HR and LR face images. Then a nonlinear mapping is built with radial basis functions (RBF)s in this subspace. Matching is carried out by simple NN classifier.

Instead of super-resolving a LR image for matching with HR images, Gunturk et al. [51] proposed an algorithm which constructs the information required by the recognition system directly in the low dimensional eigenface domain. This is more robust to noise and registration than general pixel based super-resolution.

6.3. Projection-based approaches

Li et al. [145] proposed a method that projects face images with different resolutions into a common feature space for classification. Coupled mappings that minimize the difference between the correspondences (i.e., low-resolution and its corresponding high-resolution image) are learned. The online phase of this algorithm is a simple linear transformation, so it is more efficient than many alternatives that perform explicit synthesis/super-resolution.

Zhou et al. [16] proposed an approach named Simultaneous Discriminant Analysis (SDA). In this method, LR and HR images are projected into a common subspace by the mappings learned respectively by SDA. The mapping is designed to preserve the most discriminative information. Conventional classification methods can then be applied in the common space.

Wang et al. [17] present a projection-based approach called kernel coupled cross-regression (KCCR) for matching LR face images to HR ones. In this method, the relationship between LR and HR is described in a low dimensional embedding by a coupled mappings model and graph embedding analysis. The kernel trick is applied to make this embedding non-linear. They realize the framework with spectral regression to improve computational efficiency and generalization.

Sharma and Jacobs's cross-modality model [9] discussed previously can also be used for LR-HR matching. PLS is used to linearly map images of LR and HR to a common subspace. The matching results show that PLS can be used to obtain state-of-the-art face recognition performance in matching LR to HR face images.

Multidimensional Scaling (MDS) is used by Biswas et al. [22] to simultaneously embed LR and HR images in a common space. In this space, the distance between LR and HR approximates the distance between corresponding HR images.

Ren et al. [23] tackle the low-high resolution face recognition by coupled kernel embedding (CKE). With CKE, they non-linearly map face images of both resolutions into an infinite dimensional Hilbert space where neighborhoods are preserved. Recognition is carried out in the new space.

Siena et al. [45] introduced a Maximum-Margin Coupled Mappings (MMCM) approach for low-high resolution face recognition. A Maximum-margin strategy is used to learn the projections which maps LR and HR data to a common space where there is the maximum margin of separation between pairs of cross-domain data from different classes.

In [52], Li et al. generalize CCA to use discriminative information in learning a low dimensional subspace for LR-HR image recognition. This is an closed-form optimization that is more efficient that super-resolution first strategies, while being applicable to other types of 'degraded' images besides LR, such as blur and occlusion.

Deng et al. [146] utilized color information to tackle LR face recognition as color cues are less variant to resolution change. They improved on [145] to introduce a regularized coupled mapping to project both LR and HR face images into a common discriminative space.

Representation learning and metric learning were combined and optimized jointly by [53]. Matching is finally performed using NN with the learned metric.

Finally [55] addressed LR-HR matching while simultaneously addressing the sparsity of annotated data by combining the ideas of co-training and transfer learning. They pose learning HFR as a transfer learning problem of adapting an (easier to train) HR-HR matching model to a (HFR) HR-LR matching task. The base model is binary-verification SVM based on LPQ and SIFT features. To address sparsity of *annotated* cross-domain training data, they perform co-training which exploits a large but un-annotated pool of cross-domain data to improve the matching model.

6.4. Summary and Conclusions

Both high-resolution synthesis and sub-space projection methods have been successfully applied to LR-HR recognition. In both cases the key insight to improve performance has been to use discriminative information in the reconstruction/projection, so that the new representation is both accurate and discriminative for identity. Interestingly, while this discriminative cue has been used relatively

Table 7: LR-HR matching methods: Performance on benchmark datasets.

Method	Publications	Recognition Approach	Dataset	Feature	Train:Test	High Resolution	Low Resolution	Accuracy
Synthesis based	[18]	S^2R^2	FRGC V2.0	CFA	5120:300	24×24	6×6	80%
	[48]	DSR	FRGC V2.0	LBP	2488:622	56×48	7×6	56%
	[143]	Nonlinear kernel	FRGC V2.0		2220:4168	64×56	14×16	84%
	[144]	Bayesian latent variable	$AR + FERET + Yale$		2655:295	56×36	14×9	75%
	[49]	GDAMM	AR		441:441	24×24	7×8	73%
	[50]	RBFs	FERET		1196:1195	72×72	12×12	84%
Projection based	[145]	CMs	FERET		1002:1195	72×72	12×12	92%
	[16]	SDA	FERET		1002:1195	72×72	12×12	93%
	[17]	KCCR	FERET		1002:1195	72×72	12×12	91%
	[9]	PLS	FERET		90:100	76×66	5×4	60%
	[22]	MDS	FRGC		183:608	45×39	9×7	56%
	[23]	CKE	Multi-PIE		108:229	Original Images	6×6	88%
	[146]	CMs	AR	Color	700:700	33×24	7×6	85%
	[55]	SVM	SCface	LPQ+SIFT		72×72	24×24	70%
	[53]	CBD	SCface		510:130	30×24	15×12	58%

less frequently in SBFR, NIR and 3D matching, it has been used almost throughout in HR-LR matching. Tab. 7 summarizes the results of major LR-HR matching studies; although again lack of consistency in experimental settings prevents direct quantitative comparison.

LR Dataset realism. With few exceptions [43, 55], the majority of LR-HR studies *simulate* LR data by down-sampling HR face images. Similarly to SBFR's focus on viewed-sketches, it is unclear that this is a realistic simulation of a practical LR-HR task. In practice, LR surveillance images are unavoidably captured with many other artefacts such as lighting change, motion-blur, shadows, non-frontal alignment and so on [43, 55]. Thus existing systems are likely to under perform in practice. This may lead into integrating super-resolution and recognition with simultaneous de-blurring [147, 148], re-lighting [43] and pose alignment [60].

7. Discussion

As conventional within-modality face-recognition under controlled conditions approaches a solved problem, heterogeneous face recognition has grown in interest. This has occurred independently across a variety of covariates – Sketch, NIR, LR and 3D. In case there is a strong driving application factor in security/law-enforcement/forensics. We draw the following observations and conclusions:

7.1. Common Themes

Model types. Although the set of modality pairs considered has been extremely diverse (Sketch-Photo, VIS-NIR, HR-LR, 2D-3D), it is interesting that a few common themes emerge about how to tackle modality heterogeneity. Synthesis and subspace-projection have been applied in each case. Moreover, integrating the learned projection with a discriminative constraint that different identities should be separable, has been effectively exploited in a variety of ways. On the other hand, feature engineering approaches, while often highly effective, have been largely limited to situations where the input-representation itself is not intrinsically heterogeneous (Sketch-Photo, and VIS-NIR).

Learning-based or Engineered. An important property differentiating cross-domain recognition systems is whether they require training data or not (and if so how much). Most feature-engineering based approaches have the advantage of requiring no training data, and thus not requiring a (possibly hard to obtain) dataset of annotated image pairs to be obtained before training for any particular application. On the other hand, synthesis and projection approaches (and some learning-based feature approaches), along with discriminatively trained matching strategies, can potentially perform better at the cost of requiring such a dataset. A third less-explored alternative is approaches that can perform effective unsupervised representation learning, such as auto-encoders and RBMs [47].

Exploiting Face Structure. The methods reviewed in this survey varied in how much face-specific information is exploited; as opposed to generic cross-domain methods. Analytic and component-based face image representations exploit face structure, but these are less common than patch-based or holistic representations. Methods in the 2D-3D HFR setting often use explicit face representations in order to exploit 3D's ability to align and correct for lighting shift. However, the majority of methods reviewed do not exploit face-specific domain knowledge, relying on simple holistic or patch based representations with generally applicable synthesis/projection steps (e.g., CCA, PLS, sparse coding). Many methods rely on the assumption of a fairly

accurate and rigid correspondence in order use simple representations and mappings (such as patches with CCA). Going forward, this may be an issue in some circumstances like forensic sketch and 'in the wild' LR recognition where accurate alignment is difficult.

Dataset over-fitting. Recognition tasks in broader computer vision have recently been shown to suffer from overfitting to entire datasets, as researchers engineer methods to maximize benchmark scores on insufficiently diverse datasets [149]. Current HFR datasets, notably in Sketch are also small and likely insufficiently diverse. As new larger and more diverse datasets are established, it will become clear whether existing methods do indeed generalize, and if the current top performers continue to be the most effective.

7.2. Issues and Directions for Future Research

Training data Volume. An issue for learning-based approaches is how much training data is required. Simple mappings to low-dimensional sub-spaces may require less data than more sophisticated non-linear mappings across modalities, although the latter are in principle more powerful. Current heterogeneous face datasets, for example in sketch [25, 32, 25, 39], are much smaller than those used in homogeneous face recognition [82] and broader computer vision [150] problems. As larger heterogeneous datasets are collected in future, more sophisticated non-linear models may gain the edge. This is even more critical for future research into HFR with deep-learning based methodologies which have proven especially powerful in conventional face recognition, but require thousands to millions of annotated images [3].

Alignment. Unlike homogeneous face recognition which has moved onto recognition 'in the wild' [82], heterogeneous recognition generally relies on accurately and manually aligned facial images. As a result, it is unclear how existing approaches will generalize to practical applications with inaccurate automatic alignment. Future work should address HFR methods that are robust enough to deal with residual alignment errors, or integrate alignment into the recognition process.

Side Information and Soft Biometrics. Side information and soft-biometrics have been used in a few studies [109] to prune the search space to improve matching performance. The most obvious examples of this are filtering by gender or ethnicity. Where this information is provided as metadata, filtering to reduce the matching-space is trivial. Alternatively, such soft-biometric properties can be estimated directly from data, and then the estimates used to refine the search space. However, better biometric estimation and appropriate fusion methods then need to be developed to balance the contribution of the biometric cue versus the face-matching cue.

Facial Attributes. Related to soft-biometrics is the concept of facial attributes. Attribute-centric modelling has made huge impact on broader computer vision problems [151]. They have successfully been applied to cross-domain modeling for person (rather than face) recognition [152]. Early analysis using manually annotated attributes highlighted their potential to help bridge the cross-modal gap by representing faces at a higher-level of abstraction [54]. Recent studies [118] have begun to address fully automating the attribute extraction task for cross-domain recognition, as well as releasing facial attribute annotation datasets (both caricature and forensic sketch) to support research in this area. In combination with improving facial attribute recognition techniques [153], this is a promising avenue to bridge the cross-modal gap.

Computation Time. For automated surveillance, or search against realistically large mugshot datasets, we may need to recognise faces in milliseconds. Test-time computation is thus important, which may be an implication for models with sophisticated non-linear mappings across modalities; or in the LR-HR case, synthesis (super-resolution) methods that are often expensive. Deep learning techniques may help here, as while they are costly to train, they can provide strong non-linear mappings with modest run-time cost.

Technical Methodologies. CCA, PLS, Sparse Coding, MRFs, metric learning and various generalizations thereof have been used extensively in the studies reviewed here. Going forward, there are other promising methodologies that are currently under-exploited in HFR, notably transfer learning and deep learning.

Deep Learning. Deep learning has transformed many problems in computer vision by learning significantly more effective feature representations [154]. These representations can be unsupervised or discriminatively trained, and have been used to archieve good effect in conventional face recognition [3, 155]. They have also been effectively applied for many HFR-related problems including face-recognition across pose [156], facial attribute recognition [153] (which provides a more abstract domain/modality invariant representation), and super-resolution [157] (which could potentially be used to address the HR-LR variant of HFR). Preliminary studies found that conventional photo face recognition DNNs do not provide an excellent out of the box representation for HFR [78], suggesting that they need to be trained and/or designed specifically for HFR.

Only a few studies have begin to consider application of Deep Neural Networks (DNN)s to HFR [24], thus there is significant scope for deep learning to make impact in future. In terms of our abstract HFR pipeline, deep learning approaches to HFR would combine both feature-based and synthesis or projection approaches by learning a deep hierarchy of features that together bridge the cross-modal gap. Both cross-modal HFR synthesis or match-

ing would be possible with deep learning: E.g., by fully-convolutional networks such as used in super-resolution [157], image-image encoders such as used for cross-pose matching [156], or multi-branch verification/ranking networks such used in other matching problems [158, 159]. To fully exploit DNNs for the HFR problem, a key challenge is HFR datasets, which likely need to grow to support training dat requirements of DNNs, or developing methods for training DNNs with sparse data [160]. Nevertheless, if this can be solved, DNNs are expected to provide improved feature and cross-modal projection learning compared to existing approaches. Like CCA style projections, but unlike many other reviewed methods, they can match across heterogenous dimensionality, e.g., as required for 2D-3D matching. Moreover they provide the opportunity to integrate a number of other promising strategies discussed earlier including multi-task learning for integrating attribute/biometric information with matching [161], jointly reasoning about alignment and matching [159], and fast yet non-linear matching.

Transfer Learning. Transfer Learning (including Domain Adaptation (DA)) [120] is also growing in importance in other areas of computer vision [162], and has begun to influence, e.g., view and lighting invariant face recognition [163]. This research area addresses adapting models to a different-but-related task or domain to those which they were trained [120, 162]. Approaches to adapt both specific models [120, 24] and model-agnostic approaches that adapt low-level features both exist [120, 163]. Some also require annotated target domain training data [24] while others do not [163]. A straightforward application of DA to HFR would be adapting a within-domain model (e.g., HR-HR) to another within domain setting (e.g., LR-LR). The outstanding research question for HFR is how to use these ideas to support cross-modal matching, which is just beginning to be addressed [24, 55]. Finally, we note that TL is potentially synergistic with deep leaning, in potentially allowing a strong DNN trained from large conventional recognition datasets to be adapted to HFR tasks.

7.3. Conclusion

In this survey we have reviewed the state of the art methodology and datasets in heterogeneous face recognition across multiple modalities including Photo-Sketch, VIS-NIR, 2D-3D and HR-LR. We provided a common framework to breakdown and understand the individual components of a HFR pipeline and a typology of approaches, that can be used to relate methods both within and across these diverse HFR settings. Based on this analysis we extract common themes, drawing connections across the somewhat distinct communities of HFR research, as well as identifying challenges for the field and directions for future research.

[1] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, Face recognition: A literature survey, Journal ACM Computing Surveys (CSUR) (2003) 399–458.

[2] Frontex, Biopass II: Automated biometric border crossing systems based on electronic passports and facial recognition: Rapid and smartgate (2010).

[3] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, NIPS.

[4] H. Nejati, T. Sim, A study on recognizing non-artistic face sketches, in: IEEE Workshop on Applications of Computer Vision (WACV), 2011, pp. 240–247.

[5] P. Yuen, C. H. Man, Human face image searching system using sketches, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans(TSMC) (2007) 493–504.

[6] S. Pramanik, D. Bhattacharjee, Geometric feature based face-sketch recognition, in: Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012, pp. 409–415.

[7] X. Tang, X. Wang, Face photo recognition using sketch, in: ICIP, 2002, pp. 257–260.

[8] X. Wang, X. Tang, Face sketch synthesis and recognition, in: ICCV, 2003, pp. 687–694.

[9] A. Sharma, D. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch, in: CVPR, 2011, pp. 593–600.

[10] D. Yi, R. Liu, R. Chu, Z. Lei, S. Li, Face matching between near infrared and visible light images, in: Advances in Biometrics, Springer, 2007, pp. 523–530.

[11] S. Liao, D. Yi, Z. Lei, R. Qin, S. Z. Li, Heterogeneous face recognition from local structures of normalized appearance, in: International Conference on Advances in Biometrics, 2009, pp. 209–218.

[12] R. Wang, J. Yang, D. Yi, S. Li, An analysis-by-synthesis method for heterogeneous face biometrics, in: Advances in Biometrics, Springer, 2009, pp. 319–326.

[13] G. Toderici, G. Passalis, S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Theoharis, I. Kakadiaris, Bidirectional relighting for 3d-aided 2d face recognition, in: CVPR, 2010, pp. 2721–2728.

[14] Y. WU, S. K. Shah, I. A. Kakadiaris, Rendering or normalization? an analysis of the 3d-aided pose-invariant face recognition, in: ISBA, 2016, pp. 1–8.

[15] A. Moeini, H. Moeini, K. Faez, Unrestricted pose-invariant face recognition by sparse dictionary matrix, IVC.

[16] C. Zhou, Z. Zhang, D. Yi, Z. Lei, S. Li, Low-resolution face recognition via simultaneous discriminant analysis, in: The International Joint Conference on Biometrics (IJCB), 2011, pp. 1–6.

[17] Z. Wang, Z. Miao, Y. Wan, Z. Tang, Kernel coupled cross-regression for low-resolution face recognition, Mathematical Problems in Engineering (2013) 1–20.

[18] P. Hennings-Yeomans, S. Baker, B. Kumar, Simultaneous super-resolution and feature extraction for recognition of low-resolution faces, in: CVPR, 2008, pp. 1–8.

[19] Z. Lei, S. Li, Coupled spectral regression for matching heterogeneous faces, in: CVPR, 2009, pp. 1123–1128.

[20] Z. Lei, C. Zhou, D. Yi, A. K. Jain, S. Z. Li, An improved coupled spectral regression for heterogeneous face recognition., in: The IAPR International Conference on Biometrics (ICB), 2012, pp. 7–12.

[21] J. Y. Zhu, W. S. Zheng, J.-H. Lai, Logarithm gradient histogram: A general illumination invariant descriptor for face recognition, in: FG, 2013, pp. 1–8.

[22] S. Biswas, K. W. Bowyer, P. J. Flynn, Multidimensional scaling for matching low-resolution face images, TPAMI (2012) 2019–2030.

[23] C. X. Ren, D. Q. Dai, H. Yan, Coupled kernel embedding for low-resolution face image recognition, TIP (2012) 3770–3783.

[24] P. Mittal, M. Vatas, R.Singh, Composite sketch recognition via deep network-a transfer learning approach, in: IAPR International Conference on Biometrics, 2015, pp. 251–256.

[25] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, TPAMI (2009) 1955–1967.

[26] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, in: CVPR, 2005, pp.

1005–1010.
[27] J. Chen, D. Yi, J. Yang, G. Zhao, S. Li, M. Pietikainen, Learning mappings for face synthesis from near infrared to visual light images, in: CVPR, 2009, pp. 156–163.
[28] W. Yang, D. Yi, Z. Lei, J. Sang, S. Li, 2d-3d face matching using cca, in: Automatic Face Gesture Recognition, 2008, pp. 1–6.
[29] H. Galoogahi, T. Sim, Inter-modality face sketch recognition, in: ICME, 2012, pp. 224–229.
[30] H. Kiani Galoogahi, T. Sim, Face photo retrieval by sketch example, in: ACM M, 2012, pp. 1–4.
[31] B. Klare, A. K. Jain, Sketch-to-photo matching: a feature-based approach, in: Biometric Technology for Human Identification VII.SPIE, 2010, pp. 1–10.
[32] H. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, Memetically optimized mcwld for matching sketches with digital face images, TIFS (2012) 1522–1535.
[33] Z. Khan, Y. Hu, A. Mian, Facial self similarity for sketch to photo matching, in: Digital Image Computing Techniques and Applications (DICTA), 2012, pp. 1–7.
[34] H. Han, B. Klare, K. Bonnen, A. Jain, Matching composite sketches to face photos: A component-based approach, IEEE Transactions on Information Forensics and Security (2013) 191–204.
[35] S. Liu, D. Yi, Z. Lei, S. Li, Heterogeneous face image matching using multi-scale features, in: The IAPR International Conference on Biometrics (ICB), 2012, pp. 79–84.
[36] D. Huang, M. Ardabilian, Y. Wang, L. Chen, Oriented gradient maps based automatic asymmetric 3d-2d face recognition, in: The IAPR International Conference on Biometrics (ICB), 2012, pp. 125–131.
[37] R. S. Tejas Indulal Dhamecha, Praneet Sharma, M. Vatsa, On effectiveness of histogram of oriented gradient features for visible to near infrared face matching, in: International Conference on Pattern Recognition (ICPR), 2014, pp. 1788–1793.
[38] J. Y. Zhu, W. S. Zheng, J.-H. Lai, S. Li, Matching nir face to vis face using transduction, TIFS (2014) 501–514.
[39] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: CVPR, 2011, pp. 513–520.
[40] D. Lin, X. Tang, Inter-modality face recognition, in: ECCV, 2006, pp. 13–26.
[41] B. Klare, A. Jain, Heterogeneous face recognition: Matching nir to visible light images, in: International Conference on Pattern Recognition (ICPR), 2010, pp. 1513–1516.
[42] D. Huang, M. Ardabilian, Y. Wang, L. Chen, Asymmetric 3d/2d face recognition based on lbp facial representation and canonical correlation analysis, in: ICIP, 2009, pp. 3325–3328.
[43] S. Shekhar, V. Patel, R. Chellappa, Synthesis-based recognition of low resolution faces, in: The 2011 International Joint Conference on Biometrics (IJCB), 2011, pp. 1–6.
[44] D. Huang, A. M., Y. Wang, L. Chen, Automatic asymmetric 3d-2d face recognition, in: International Conference on Pattern Recognition (ICPR), 2010, pp. 1225–1228.
[45] S. Siena, V. Boddeti, B. Kumar, Maximum-margin coupled mappings for cross-domain matching, in: Biometrics: Theory, Applications and Systems (BTAS), 2013, pp. 1–8.
[46] D. A. Huang, Y. C. F. Wang, Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition, in: ICCV, 2013, pp. 2496–2503.
[47] D. Yi, Z. Lei, S. Z. Li, Shared representation learning for heterogenous face recognition, in: FG, 2015, pp. 1–15.
[48] W. Zou, P. Yuen, Very low resolution face recognition problem, TIP (2012) 327–340.
[49] J. Jiang, R. Hu, Z. Han, K. Huang, T. Lu, Graph discriminant analysis on multi-manifold: A novel super-resolution method for face recognition, in: ICIP, 2012, pp. 1465–1468.
[50] H. Huang, H. He, Super-resolution method for face recognition using nonlinear mappings on coherent features, IEEE Transactions on Neural Networks (2011) 121–130.
[51] B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, R. Mersereau, Eigenface-domain super-resolution for face recognition, TIP (2003) 597–606.
[52] B. Li, H. Chang, S. Shan, X. Chen, Coupled metric learning for face recognition with degraded images, in: Advances in Machine Learning, Asian Conference on Machine Learning (ACML), 2009, pp. 220–233.
[53] P. Moutafis, I. Kakadiaris, Semi-coupled basis and distance metric learning for cross-domain matching: Application to low-resolution face recognition, in: IEEE International Joint Conference on Biometrics (IJCB), 2014, pp. 1–8.
[54] B. Klare, S. Bucak, A. Jain, T. Akgul, Towards automated caricature recognition, in: The IAPR International Conference on Biometrics (ICB), 2012, pp. 139–146.
[55] H. Bhatt, R. Singh, M. Vatsa, N. Ratha, Improving cross-resolution face matching using ensemble-based co-transfer learning, IEEE Transactions on Image Processing(TIP) (2014) 5654–5669.
[56] P. Mittal, A. Jain, R. Singh, M. Vatsa, Boosting local descriptors for matching composite and digital face images, in: ICIP, 2013, pp. 2797–2801.
[57] P. Mittal, A. Jain, G. Goswami, R. Singh, M. Vatsa, Recognizing composite sketches with digital face images via ssd dictionary, in: IEEE International Joint Conference on Biometrics (IJCB), 2014, pp. 1–6.
[58] K. W. Bowyer, K. Chang, P. Flynn, A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition, CVIU (2006) 1–15.
[59] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, M. A. Abidi, Recent advances in visual and infrared face recognition—a review, CVIU (2005) 103–135.
[60] X. Zhang, Y. Gao, Face recognition across pose: A review, PR (2009) 2876–2896.
[61] X. Zou, J. Kittler, K. Messer, Illumination invariant face recognition: A survey, in: Biometrics: Theory, Applications, and Systems, 2007, pp. 1–8.
[62] X. Zhang, Y. Gao, Face recognition across pose: A review, PR (2009) 2876–2896.
[63] Y. Wang, C. S. Chua, Face recognition from 2d and 3d images using 3d gabor filters, IVC (2005) 1018–1028.
[64] C. S. Chua, Y. Wang, Robust face recognition from 2d and 3d images using structural hausdorff distance, IVC (2006) 176–185.
[65] G. P. Kusuma, C. S. Chua, Pca-based image recombination for multimodal 2d and 3d face recognition, IVC (2011) 306–316.
[66] É. Marchand, P. Bouthemy, F. Chaumette, A 2d–3d model-based approach to real-time visual tracking, IVC (2001) 941–955.
[67] N. Wang, D. Tao, X. Gao, X. Li, J. Li, A comprehensive survey to face hallucination, IJCV 106 (1) (2014) 9–30.
[68] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Face recognition from a single image per person: A survey, PR (2006) 1725–1745.
[69] I. Biometrix, Face 4.0, IQ Biometrix.
[70] P. Wright, J. Corder, M. Glazier, Identi-kit (2007).
[71] S. Klum, H. Han, A. K. Jain, B. Klare, Sketch based face recognition: Forensic vs. composite sketches, in: The International Conference on Biometrics (ICB), 2013, pp. 1–8.
[72] D. McQuiston-Surrett, L. D. Topp, R. S. Malpass, Use of facial composite systems in us law enforcement agencies, Psychology, Crime and Law (2006) 505–517.
[73] G.Wells, L. Hasel, Facial composite production by eyewitnesses, Current Directions in Psychological (2007) 6–10.
[74] C. D. Frowd, P. J. B. Hancock, D. Carson, Evofit: A holistic, evolutionary facial imaging technique for creating composites., Journal ACM Transactions on Applied Perception(TAP) (2004) 19–39.
[75] Y. Zhang, C. McCullough, J. Sullins, C. Ross, Hand-drawn face sketch recognition by humans and a pca-based algorithm for forensic applications, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans(TSMC) (2010) 475–485.
[76] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive

Neuroscience (1991) 71–86.
[77] S. Klum, H. Han, B. Klare, A. Jain, The facesketchid system: Matching facial composites to mugshots, IEEE Transactions on Information Forensics and Security 9 (12) (2014) 2248–2263.
[78] S. Ouyang, T. M. Hospedales, Y.-Z. Song, X. Li, Forgetmenot: Memory-aware forensic facial sketch matching, in: CVPR, 2016, pp. 1–8.
[79] A. M. Martinez, R. Benavente, The ar face database, Tech. rep., CVC Technical Report 24 (1998).
[80] K. Messer, J. Matas, J. Kittler, K. Jonsson, Xm2vtsdb: The extended m2vts database, in: In Second International Conference on Audio and Video-based Biometric Person Authentication, 1999, pp. 1–6.
[81] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The feret evaluation methodology for face-recognition algorithms, TPAMI (2000) 1090–1104.
[82] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Tech. rep., University of Massachusetts, Amherst (2007).
[83] L. Gibson, Forensic Art Essentials, Academic Press, 2008.
[84] K. Taylor, Forensic Art and Illustration, CRC Press, 2001.
[85] C. D. Frowd, W. B. Erickson, J. M. Lampinen, F. C. Skelton, A. H. McIntyre, P. J. Hancock, A decade of evolving composites: regression- and meta-analysis, The Journal of Forensic Practice 17 (4) (2015) 319–334.
[86] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, TPAMI (1997) 696–710.
[87] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, TPAMI (1997) 711–720.
[88] L. F. Chen, H. Y. M. Liao, M. T. Ko, J.-C. Lin, G. J. Yu, A new lda-based face recognition system which can solve the small sample size problem, PR (2000) 1713–1726.
[89] X. Wang, X. Tang, Dual-space linear discriminant analysis for face recognition, in: CVPR, 2004, pp. 564–569.
[90] X. Tang, Random sampling lda for face recognition, in: CVPR, 2004, pp. 1–7.
[91] X. Wang, Random sampling for subspace face recognition, in: IJCV, 2006, pp. 91–104.
[92] J. Zhong, X. Gao, C. Tian, Face sketch synthesis using e-hmm and selective ensemble, in: Acoustics, Speech and Signal Processing(ICASSP), 2007, pp. 485–488.
[93] X. Gao, J. Zhong, D. Tao, X. Li, Local face sketch synthesis learning, Neurocomputing (2008) 1921–1930.
[94] X. Gao, J. Zhong, J. Li, C. Tian, Face sketch synthesis algorithm based on e-hmm and selective ensemble, IEEE Transactions on Circuits and Systems for Video Technology (2008) 487–496.
[95] B. Xiao, X. Gao, D. Tao, X. Li, A new approach for face recognition by sketches in photos, Signal Processing (2009) 1576–1588.
[96] W. Liu, X. Tang, J. Liu, Bayesian tensor inference for sketch-based facial photo hallucination, in: The international joint conference on Artifical intelligence, 2007, pp. 2141–2146.
[97] S. Zhang, X. Gao, N. Wang, J. Li, M. Zhang, Face sketch synthesis via sparse representation-based greedy search, TIP.
[98] S. Zhang, X. Gao, N. Wang, Face sketch synthesis from a single photo-sketch pair, IEEE Transactions on Circuits and Systems for Video Technology.
[99] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, J. Li, Multiple representations-based face sketch-photo synthesis, IEEE Transactions on Neural Networks and Learning Systems.
[100] D. Lowe, Distinctive image features from scale-invariant keypoints, IJCV (2004) 91–110.
[101] H. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, On matching sketches with digital face images, in: BTAS, 2010, pp. 1–7.
[102] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Longman Publishing Co., 1989.
[103] H. Galoogahi, T. Sim, Face sketch recognition by local radon binary pattern: Lrbp, in: ICIP, 2012, pp. 1837–1840.
[104] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006, pp. 2169–2178.
[105] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: CVPR, 2011, pp. 513–520.
[106] J. Choi, A. Sharma, D. Jacobs, L. Davis, Data insufficiency in sketch versus photo face recognition, in: CVPR, 2012, pp. 1–8.
[107] L. Anand, M. Kafaiand, S. Yang, B. Bhanu, Reference based person reidentification, in: Advanced Video and Signal Based Surveillance(AVSS), 2013, pp. 244–249.
[108] J. Uhl, R.G., N. da Vitoria Lobo, A framework for recognizing a facial image from a police sketch, in: CVPR, 1996, pp. 586–593.
[109] B. Klare, Z. Li, A. Jain, Matching forensic sketches to mug shot photos, TPAMI (2011) 639–646.
[110] B. F. Klare, A. K. Jain, Heterogeneous face recognition using kernel prototype similarities, TPAMI (2013) 1410–1422.
[111] T. Chugh, H. Bhatt, R. Singh, M. Vatsa, Matching age separated composite sketches and digital face images, in: BTAS, 2013, pp. 1–6.
[112] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, TPAMI (2002) 971–987.
[113] S. Milborrow, F. Nicolls, Locating facial features with an extended active shape model, in: ECCV, 2008, pp. 504–513.
[114] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, M. I. Jordan, Learning the kernel matrix with semidefinite programming, Journal of Machine Learning Research (2004) 27–72.
[115] P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: Nineteen results all computer vision researchers should know about, Proceedings of the IEEE (2006) 1948–1962.
[116] R. Mauro, M. Kubovy, Caricature and face recognition, Memory & Cognition (1992) 433–440.
[117] G. Rhodes, S. Brennan, S. Carey, Identification and ratings of caricatures: Implications for mental representations of faces, Cognitive Psychology (1987) 473–497.
[118] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, Cross-modal face matching: beyong viewed sketches, Accepted by ACCV.
[119] H. Nizami, J. Adkins-Hill, Y. Zhang, J. Sullins, C. McCullough, S. Canavan, L. Yin, A biometric database with rotating head videos and hand-drawn face sketches, in: Biometrics: Theory, Applications, and Systems, 2009, pp. 1–6.
[120] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering (TKDE) (2010) 1345–1359.
[121] C. Frowd, V. Bruce, A. McIntyre, P. Hancock, The relative importance of external and internal features of facial composites, British Journal of Psychology 98 (1) (2007) 61–77.
[122] S. Li, Z. Lei, M. Ao, The hfb face database for heterogeneous face biometrics research, in: CVPR, 2009, pp. 1–8.
[123] S. Li, D. Yi, Z. Lei, S. Liao, The casia nir-vis 2.0 face database, in: CVPR, 2013, pp. 348–353.
[124] D. Goswami, C.-H. Chan, D. Windridge, J. Kittler, Evaluation of face recognition system in heterogeneous environments (visible vs nir), in: ICCV, 2011, pp. 2160–2167.
[125] B. Zhang, L. Zhang, D. Zhang, L. Shen, Directional binary code with application to polyu near-infrared face database, Pattern Recognition Letters (2010) 2337–2344.
[126] X. Xie, K.-M. Lam, An efficient illumination normalization method for face recognition, Pattern Recognition Letters 27 (2006) 609–617.
[127] X. Pengfei, L. Huang, C. Liu, A method for heterogeneous face image synthesis, in: The IAPR International Conference on Biometrics (ICB), 2012, pp. 1–6.
[128] F. J. Xu, D. K. Pal, M. Savvides, Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction, in: CVPR Workshops, 2015, pp. 141–150.

[129] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in Statistics, Biometrika, 1992, pp. 162–190.
[130] S. Liao, X. Zhu, Z. Lei, L. Zhang, S. Li, Learning multi-scale block local binary patterns for face recognition, in: Advances in Biometrics, Springer, 2007, pp. 828–837.
[131] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: A general framework for dimensionality reduction, TPAMI (2007) 40–51.
[132] D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, in: ICCV, 2007, pp. 1–8.
[133] X. Huang, Z. Lei, M. Fan, X. Wang, S. Li, Regularized discriminative spectral regression method for heterogeneous face matching, TIP (2013) 353–362.
[134] L. Huang, J. Lu, Y.-P. Tan, Learning modality-invariant features for heterogeneous face recognition, in: International Conference on Pattern Recognition (ICPR), 2012, pp. 1683–1686.
[135] D. Gong, J. Zheng, A maximum correlation feature descriptor for heterogeneous face recognition, in: Asian Conference on Pattern Recognition (ACPR), 2013, pp. 135–139.
[136] P. Dou, L. Zhang, Y. Wu, S. K. Shah, I. A. Kakadiaris, Pose-robust face signature for multi-view face recognition, in: BTAS, 2015, pp. 1–8.
[137] Uhdb11 face database, UH Computational Biomedicine Lab, 2009.
[138] Uhdb12 face database, UH Computational Biomedicine Lab, 2009.
[139] A. Rama, F. Tarres, D. Onofrio, S. Tubaro, Mixed 2d-3d information for pose estimation and face recognition, in: Acoustics, Speech and Signal Processing, 2006, pp. 361–364.
[140] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image super-resolution via sparse representation, TIP (2010) 2861–2873.
[141] M. Grgic, K. Delac, S. Grgic, Scface- surveillance cameras face database, Multimedia Tools Appl. (2011) 863–879.
[142] P. H. H. Yeomans, B. V. Kumar, S. Baker, Robust low-resolution face identification and verification using high-resolution features, in: ICIP, 2009, pp. 33–36.
[143] W. W. Zou, P. C. Yuen, Learning the relationship between high and low resolution images in kernel space for face super resolution, in: International Conference on Pattern Recognition(ICPR), 2010, pp. 1152–1155.
[144] K. Jia, S. Gong, Multi-modal tensor face for simultaneous super-resolution and recognition, in: ICCV, 2005, pp. 1683–1690.
[145] B. Li, H. Chang, S. Shan, X. Chen, Low-resolution face recognition via coupled locality preserving mappings, Signal Processing Letters, IEEE (2010) 20–23.
[146] Z.-X. Deng, D.-Q. Dai, X.-X. Li, Low-resolution face recognition via color information and regularized coupled mappings, in: Chinese Conference on Pattern Recognition (CCPR), 2010, pp. 1–5.
[147] S. Cho, Y. Matsushita, S. Lee, Removing non-uniform motion blur from images, in: ICCV, 2007, pp. 1–8.
[148] A. Levin, Y. Weiss, F. Durand, W. T. Freeman, Understanding blind deconvolution algorithms, TPAMI (2011) 2354–2367.
[149] A. Torralba, A. A. Efros, Unbiased look at dataset bias, in: CVPR, 2011, pp. 1521–1528.
[150] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.
[151] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: CVPR, 2009, pp. 951–958.
[152] R. Layne, T. M. Hospedales, S. Gong, Person re-identification by attributes, in: BMVC, 2012, pp. 1–8.
[153] P. Luo, X. Wang, X. Tang, A deep sum-product architecture for robust facial attributes analysis, in: ICCV, 2013, pp. 2864–2871.
[154] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012, pp. 1–9.
[155] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, T. Hospedales, When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition, arXiv preprint arXiv:1504.02351.
[156] Z. Zhu, P. Luo, X. Wang, X. Tang, Deep learning identity-preserving face space, in: ICCV, 2013, pp. 113–120.
[157] C. Dong, C. C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: ECCV, 2014, pp. 1–16.
[158] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, C. C. Loy, Sketch me that shoe, in: CVPR, 2016, pp. 1–8.
[159] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: CVPR, 2014, pp. 152–159.
[160] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, J. Verbeek:, Frankenstein: Learning deep face representations using small data., CoRR abs/1603.06470.
[161] Z. Zhang, P. Luo, C. C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: ECCV, 2014, pp. 1–15.
[162] V. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: A survey of recent advances, IEEE Signal Processing Magazine.
[163] H. T. Ho, R. Gopalan, Model-driven domain adaptation on product manifolds for unconstrained face recognition, IJCV.