

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

SHUXIN OUYANG, Beijing University of Posts and Telecommunications

TIMOTHY HOSPEDALES, Queen Mary University of London

YI-ZHE SONG, Queen Mary University of London

XUEMING LI, Beijing University of Posts and Telecommunications

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.

Categories and Subject Descriptors: A.1 [**General literature**]: Introductory and survey; I.4.9 [**Image processing and Computer Vision**]: Applications; I.5.4 [**Pattern Recognition**]: Applications

General Terms: Algorithms, Performance, Security

Additional Key Words and Phrases: 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 [Zhao et al. 2003], being deployed at industrial scale for biometric border control [Frontex 2010] and producing better-than-human performances [Sun et al. 2014]. 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

---

Author's addresses: S. Ouyang, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications ; T. Hospedales and Y. Song, Computer Vision Lab, Queen Mary University of London; X. Li, School of Digital Media and Design Art.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© YYYY ACM 0360-0300/YYYY/01-ARTA \$15.00  
DOI:<http://dx.doi.org/10.1145/0000000.0000000>

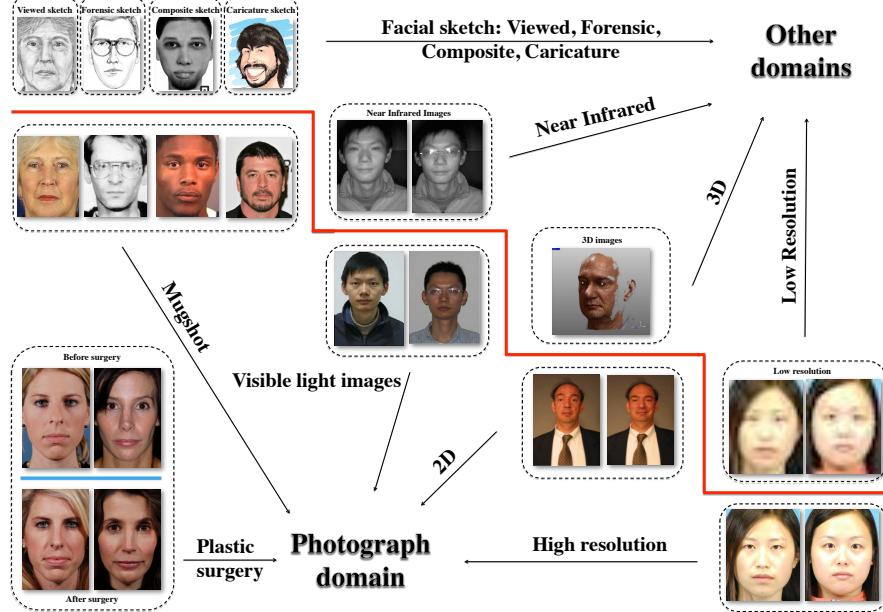


Fig. 1. Scope of heterogeneous face recognition studied in this survey.

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., infrared 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, or comparing imagery before and after plastic surgery). 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 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 con-

Table I. Overview of studies by modality. Superscript \* indicates studies that have been applied to multiple modality pairs.

Domains	Studies
Sketch-Photo	
•Viewed	[Tang and Wang 2002; Wang and Tang 2009; Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012] [Galoogahi and Sim 2012a; Bhatt et al. 2010; Klare and Jain 2010b; Pramanik and Bhattacharjee 2012] [Gao et al. 2008b; Gao et al. 2008a; Khan et al. 2012; Nejati and Sim 2011; Yuen and Man 2007] [Tang and Wang 2003; Zhang et al. 2011; Huang and Wang 2013; Liu et al. 2005] [Zhong et al. 2007; Liu et al. 2007; Choi et al. 2012; Xiao et al. 2009]* [Lin and Tang 2006; Sharma and Jacobs 2011; Huang et al. 2013]*
•Composite	[Yuen and Man 2007; Han et al. 2013; Klare and Jain 2013]
•Forensic	[Bhatt et al. 2012; Klare and Jain 2010b; Zhang et al. 2010; Uhl and da Vitoria Lobo 1996; Klare et al. 2011]
•Caricature	[Klare et al. 2012]
VIS-NIR	[Zhu et al. 2014; Liao et al. 2009; Yi et al. 2007; Chen et al. 2009; Goswami et al. 2011] [Wang et al. 2009; Pengfei et al. 2012; Klare and Jain 2010a; Lei et al. 2012; Zhu et al. 2013b] [Huang et al. 2012; Gong and Zheng 2013; Lei and Li 2009; Liu et al. 2012] [Lin and Tang 2006; Lei et al. 2012; Huang et al. 2013]*
2D-3D	[Yang et al. 2008; Huang et al. 2009; Huang et al. 2010; Toderici et al. 2010; Rama et al. 2006; Huang et al. 2012]
Low-High	[Lei and Li 2009; Zou and Yuen 2012; Hennings-Yeomans et al. 2008; Zou and Yuen 2010] [Jia and Gong 2005; Zhou et al. 2011; Li et al. 2010; Wang et al. 2013] [Biswas et al. 2012; Jiang et al. 2012; Huang and He 2011; Gunturk et al. 2003] [Zhang and He 2010; Hennings-Yeomans et al. 2009; Shekhar et al. 2011; Ren et al. 2012]* [Siena et al. 2013; Ren et al. 2011; Sharma and Jacobs 2011; Lei et al. 2012]*
Plastic surgery	[Singh et al. 2010; Aggarwal et al. 2012; LakshmiPrabha and Majumder 2012; Bhatt et al. 2013; Liu et al. 2013]

sidered. We consider four cross-modality applications: sketch-based, infra-red based, 3D-based and high-low resolution matching; as well as one within-modality application of post-surgery face 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.
- **Within-modality Heterogeneity:** Various within-modality effects can also transform facial images significantly enough to seriously challenge conventional recognition systems. In this survey we also discuss the recently topical area of recognition

across plastic surgery, that is also in demand due to implications for security and forensics.

Fig 1 offers an illustrative summary of the five categories of HFR literature covered in this survey. Tab I further summarizes all the studies reviewed broken down by the modality or modalities considered.

Related areas that are not covered by this review include (homogeneous) 3D [Bowyer et al. 2006] and infra-red [Kong et al. 2005] matching, fusing multiple modalities [Bowyer et al. 2006; Kong et al. 2005]. View [Zhang and Gao 2009] and illumination [Zou et al. 2007] invariant recognition are also related in that there exists a strong covariate between probe and gallery images, however we do not include these as good surveys already exist. A good survey about face-synthesis [Wang et al. 2014] is more relevant 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. Finally, Section 7 discusses matching faces across plastic surgery variations. 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 II breaks down all the papers reviewed in this survey by their choices about these design decisions.

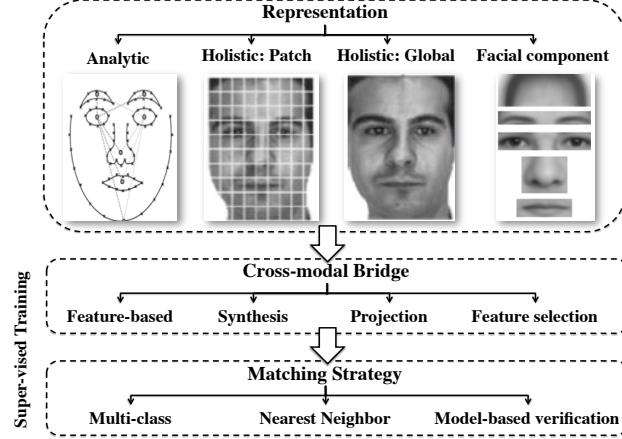


Fig. 2. Overview of an abstract HFR pipeline.

## 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** [Nejati and Sim 2011; Yuen and Man 2007; Pramanik and Bhattacharjee 2012] detect facial components and fiducial points, allowing the face to be modeled geometrically, e.g., using point distribution models [Nejati and Sim 2011; Yuen and Man 2007]. 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 [Yuen and Man 2007]. Moreover geometry is not robust to facial expression [Pramanik and Bhattacharjee 2012], 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 [Liu et al. 2012; Han et al. 2013]. This allows the informativeness of each component in matching to be measured separately [Liu et al. 2012]; and if components can be correctly detected and matched it also provides some robustness to both linear and non-linear misalignment across modalities [Han et al. 2013]. However, a component-fusion scheme is then required to produce an overall match score.

**Global holistic representations** represent the whole face image in each modality with a single vector [Yi et al. 2007; Tang and Wang 2002; Lin and Tang 2006]. 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 [Tan et al. 2006].

**Patch-based holistic representations** encode the appearance of each image in patches with a feature vector per patch [Wang and Tang 2009; Liu et al. 2005; Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012; Bhatt et al. 2012]. Subsequent strategies for using the patches vary, including for example concatenation into a very large feature vector [Klare and Jain 2010b] (making it in effect a holistic representation), or learning a mapping/classifier per patch [Zhang et al. 2011]. 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.

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

Component	Approach	Representative Examples
Representation	Analytic	Active Shape & Point Distribution Models [Nejati and Sim 2011; Yuen and Man 2007] Relative Geometry [Pramanik and Bhattacharjee 2012]
	Global Holistic	Whole image [Yi et al. 2007; Tang and Wang 2002; Tang and Wang 2003] Whole image[Liao et al. 2009; Wang et al. 2009; Lei and Li 2009; Lei et al. 2012]
	Global Patch	Whole image[Sharma and Jacobs 2011; Zhu et al. 2013b; Li et al. 2009; Siena et al. 2013] Regular grid of patches [Wang and Tang 2009; Liu et al. 2005; Galoogahi and Sim 2012b] Regular grid of patches [Kiani Galoogahi and Sim 2012; Klare and Jain 2010b; Bhatt et al. 2012]
	Facial Component	Regular grid of patches [Khan et al. 2012; Yang et al. 2008; Chen et al. 2009] Active Shape Model Detection [Han et al. 2013] Rectangular patches [Liu et al. 2012]
	Cross domain	LBP [Bhatt et al. 2012; Liao et al. 2009; Chen et al. 2009; Han et al. 2013] Gabor [Kiani Galoogahi and Sim 2012], SIFT [Klare and Jain 2010b], SSIM [Khan et al. 2012] CITE [Zhang et al. 2011], LGH [Zhu et al. 2013b], HOAG [Galoogahi and Sim 2012b]
	Projection	CDFE [Lin and Tang 2006], Common Basis [Klare and Jain 2010b] CSR [Lei and Li 2009; Lei et al. 2012; Liu et al. 2012] RS-LDA [Wang and Tang 2009] CCA [Yi et al. 2007; Yang et al. 2008; Li et al. 2009], PLS [Sharma and Jacobs 2011] Adaboost [Liu et al. 2012; Liao et al. 2009], Max-margin [Siena et al. 2013] Kernel LDA [Liu et al. 2005], Sparse Coding [Huang and Wang 2013; Shekhar et al. 2011]
	Synthesis	NN [Wang et al. 2009; Chen et al. 2009], MRF [Wang and Tang 2009] Eigentransform [Tang and Wang 2002; Tang and Wang 2003] LLE [Liu et al. 2005], Relationship learning [Zou and Yuen 2012]
	Matching	NN [Tang and Wang 2002; Klare and Jain 2010b; Tang and Wang 2003; Chen et al. 2009] NN [Li et al. 2009; Khan et al. 2012; Nejati and Sim 2011; Lei and Li 2009] NN [Lei et al. 2012; Sharma and Jacobs 2011; Zhu et al. 2013b] NN with $\chi^2$ [Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012; Liu et al. 2012] NN with HI [Han et al. 2013],NN with Cosine [Yi et al. 2007; Yang et al. 2008]
	Multi-class (Tr)	Bayesian [Tang and Wang 2003], MA metric learning for $\chi^2$ NN [Bhatt et al. 2012] SVM [Zou and Yuen 2012]
	Verification (Tr)	Similarity threshold (Cosine) [Liao et al. 2009], SVM [Klare et al. 2012] Logistic Regression [Klare et al. 2012]

## 2.2. Cross-modal bridge strategy

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 [Klare and Jain 2010b; Wang and Tang 2009].

**Feature design** strategies [Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012; Klare and Jain 2010b; Bhatt et al. 2012] 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 [Klare and Jain 2010b] and LBP [Bhatt et al. 2012].

**Synthesis** approaches focus on synthesizing one modality based on the other [Tang and Wang 2002; Wang and Tang 2009]. Typical methods include eigentransforms [Tang and Wang 2002; Tang and Wang 2003], MRFs [Wang and Tang 2009], and LLE [Liu et al. 2005]. 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 [Lin and Tang 2006; Klare and Jain 2010b; Yi et al. 2007]. Typical methods include linear discriminant analysis (LDA) [Wang and Tang 2009], canonical components analysis

(CCA) [Yi et al. 2007; Yang et al. 2008], partial least squares (PLS) and common basis [Klare and Jain 2010b] 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 [Liu et al. 2012; Liao et al. 2009], 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 sophisticated classifiers. Thus *Nearest-Neighbor (NN)* [Tang and Wang 2002; Lin and Tang 2006; Klare and Jain 2010b; Yi et al. 2007] is most commonly used to match against the gallery [Tang and Wang 2002]. NN classifiers can be defined with various distance metrics, and many studies found  $\chi^2$  [Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012] or cosine [Yang et al. 2008] 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 training data. However, they can be enhanced with metric-learning [Bhatt et al. 2012] if annotated cross-domain training data is available.

**Model-based verification strategies** pose HFR as a binary, rather than multi-class, classification problem [Liao et al. 2009; Klare et al. 2012]. 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. An advantage of verification over classification strategies is robustness and data sufficiency. In many scenarios there is only one cross-modal face pair per person. Thus classification strategies have one instance per class (person), and risk over fitting. 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.

We note that some methodologies can be interpreted as either cross-domain mappings or matching strategies. For example, some papers [Wang and Tang 2009] 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.

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

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 [Klare et al. 2012], SVM [Klare et al. 2012]) 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  $w$  of the decision function.

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

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

**Cross-domain Strategies** Feature-centric cross-domain strategies [Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012; Klare and Jain 2010b; Bhatt et al. 2012; Liao et al. 2009; Chen et al. 2009; Han et al. 2013; Khan et al. 2012; Zhu et al. 2013b; Zhang et al. 2011] 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 [Wang et al. 2009; Zou and Yuen 2012] may set  $W^a = I$ , and search for the projection  $W^b$  so that  $|F(\mathbf{x}_i^a) - W^b F(\mathbf{x}_i^a)|$  is minimized. CCA [Yi et al. 2007; Yang et al. 2008] strategies search for  $W^a$  and  $W^b$  such that  $|W^a F(\mathbf{x}_i^a) - W^b F(\mathbf{x}_i^a)|$  is minimized for cross-modal pairs of the same person  $i$ . While LDA [Wang and Tang 2009] strategies search for a single projection  $W$  such that  $|WF(\mathbf{x}_i^a) - WF(\mathbf{x}_j^a)|$  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 II). 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 [Galoogahi and Sim 2012b; Kiani Galoogahi and Sim 2012; Khan et al. 2012; Han et al. 2013]) 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 [Yi et al. 2007; Yang et al. 2008]. In contrast, LDA simultaneously finds a space where cross-modal pairs are similar and also where different identities are well separated [Wang and Tang 2009], which more directly optimizes the desired outcome of high cross-modal matching accuracy.

**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 homoge-

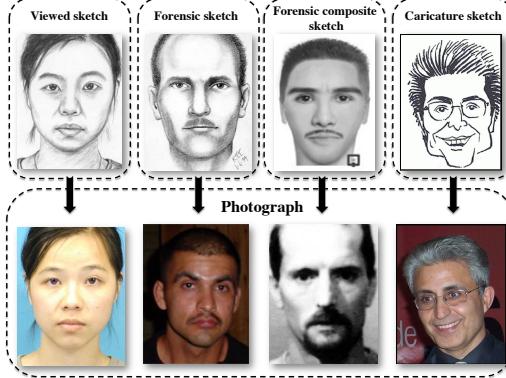


Fig. 3. Facial sketches and corresponding mugshots: viewed sketch, forensic hand drawn sketch, forensic composite sketch, caricature sketch and their corresponding facial images

neous 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 [Wang and Tang 2009].

SBFR datasets 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 the ideal forensic sketch performance when all details are remembered and communicated correctly. Research can then focus on making good viewed sketch methods robust to lower-quality forensic sketches.

#### 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:

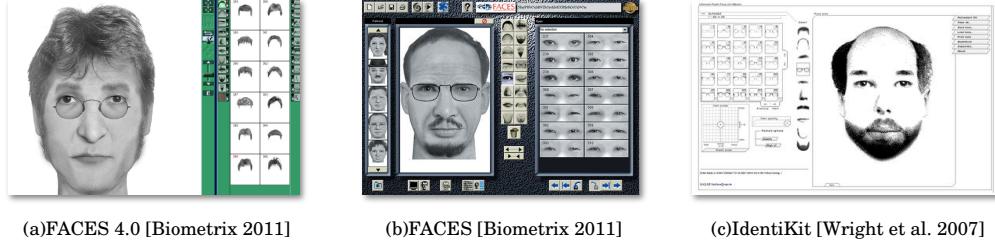


Fig. 4. Examples of different kind of composite sketch softwares

Table III. Existing facial sketch benchmark datasets.

Datasets	Pairs of Sketch/Photo	Viewed or Forensic	Composite or Hand drawn	Availability
CUFS [Wang and Tang 2009]	606	Viewed	Hand drawn	CUHK: Free to download AR: Request permission XM2VTS: Pay a fee
CUFSF [Zhang et al. 2011; Wang and Tang 2009]	1,194	Viewed	Hand drawn	Sketch: Free to download Photo: Request permission
IIIT-D viewed sketch [Bhatt et al. 2012]	238	Viewed	Hand drawn	Request permission
IIIT-D semi-forensic sketch [Bhatt et al. 2012]	140	Semi-Forensic	Hand drawn	Request permission
IIIT-D forensic sketch [Bhatt et al. 2012]	190	Forensic	Hand drawn and Composite	Request permission

- **Forensic hand drawn sketches:** These are produced by a forensic artist based on the description of a witness ([Klum et al. 2013]), 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 [McQuiston-Surrett et al. 2006]. The most widely used software for generating facial composite sketches are Identikit [Wright et al. 2007], Photo-Fit [G.Wells and Hasel 2007], FACES [Biometrix 2011], Mac-a-Mug [G.Wells and Hasel 2007], and EvoFIT [Frowd et al. 2004]. It is worth noting 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 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 [Nejati and Sim 2011; Zhang et al. 2010; Turk and Pentland 1991].

### 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 III summaries each dataset in terms of these attributes.

CUHK Face sketch dataset (CUFS) [Wang and Tang 2009] 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 [Martinez and Benavente 1998], and 295 faces from the XM2VTS dataset [Messer et al. 1999]. 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) [Zhang et al. 2011; Wang and Tang 2009] is also commonly used to benchmark SBFR algorithms. There are 1,194 subjects from the FERET dataset [Phillips et al. 2000]. 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 [Zhang et al. 2011].

The IIIT-D Sketch Dataset [Bhatt et al. 2012] 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<sup>1</sup>, 99 sketch-digital image from Labeled Faces in Wild (LFW) dataset [Huang et al. 2007], and 72 sketch-digital image pairs from the IIIT-D student & staff dataset [Huang et al. 2007]. 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 [Gibson 2008] and [Taylor 2001] respectively, as well as 61 pairs from various sources on the internet.

It is worth noting that the accessibility of these datasets varies, with some not being publicly available. [Klare and Jain 2010b] 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 sub-problem 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.

---

<sup>1</sup>Downloadable at <http://www-prima.inrialpes.fr/FGnet/html/home.html>

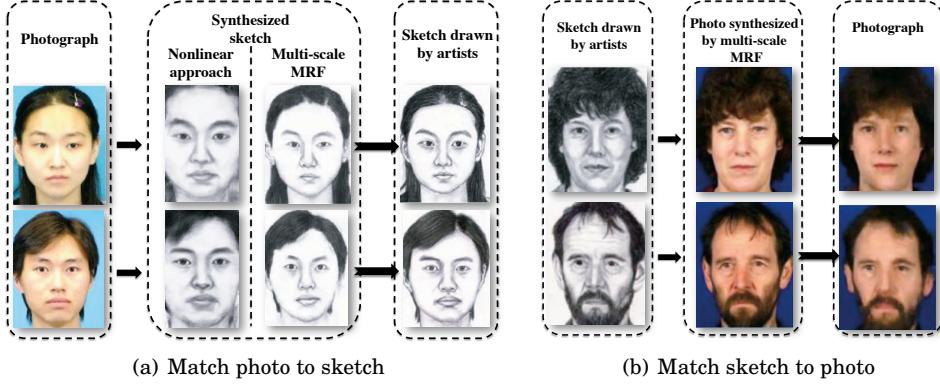


Fig. 5. Examples of sketch synthesis: (a) photo to sketch by synthesized sketches (b) sketch to photo by synthesized photos

**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, an eigensketch transformation algorithm is proposed by [Tang and Wang 2002]. 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.

Liu et al. [Liu et al. 2005] proposed a Local Linear Embedding (LLE) inspired method to convert photos into sketches based on image patches. Those sketches are geometry preserving synthetic sketches. For each image patch to be converted, it finds the nearest neighbors in the training set, and uses their corresponding sketch patches to synthesize the sketch patch. Tang and Wang [Wang and Tang 2009] further improved [Liu et al. 2005] by developing an approach to synthesize local face structures at different scales using a Markov Random Fields (MRF), as shown in Fig 5(a). In the latter work, a multi-scale MRF learns local patches and scales jointly instead of independently as in [Liu et al. 2005]. The scale of learned local face structures is based on the size of overlapped patches. With a multi-scale MRF, the joint photo-sketch model is learned at multiple scales. By converting face photos or sketches into same modality, the modality-gap is reduced, thus allowing the two domains to be matched effectively.

In both [Tang and Wang 2002] and [Wang and Tang 2009], after photos/sketches are synthesized, many standard methods like PCA [Turk and Pentland 1991], Bayesian-face [Moghaddam and Pentland 1997], Fisherface [Belhumeur et al. 1997], null-space LDA [Chen et al. 2000], dual-space LDA [Wang and Tang 2004a] and Random Sampling LDA (RS-LDA) [Wang and Tang 2004b; Wang and Tang 2006a] are straightforwardly applied for homogeneous face recognition.

The embedded hidden Markov model (E-HMM) is applied by Zhong et al. [Zhong et al. 2007] 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 are used to synthesize a finer face pseudo-sketch based on a selective ensemble strategy. E-HMMs are also used by Gao et al. [Gao et al. 2008b; Gao et al. 2008a] to synthesis sketches from photos. On the

contrary, Xiao et al. [Xiao et al. 2009] proposed a E-HMM based method to synthesis photos from sketches. Liu et al. [Liu et al. 2007] proposed a synthesis method based on Bayesian Tensor Inference. This method can be used to synthesize both sketches from photos and photos from sketches.

*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 [Lin and Tang 2006] 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. [Sharma and Jacobs 2011] 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 [Huang and Wang 2013], 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 [Liu et al. 2005] 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 descriptors, they may be matched directly, or after a subsequent projection-based step as in the previous section.

Klare et al. [Klare and Jain 2010b] proposed the first direct sketch/photo matching method based on invariant SIFT-features [Lowe 2004]. SIFT features provide a compact vector representation of an image patch based on the magnitude, orientation, and spatial distribution of the image gradients [Klare and Jain 2010b]. 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. [Bhatt et al. 2010] 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) [Goldberg 1989] 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.

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

Method	Publications	Recognition Approach	Dataset	Feature	Train:Test	Accuracy
Synthesis based	[Tang and Wang 2002]	KLT	CUHK	Eigen-sketch features	88:100	about 60%
	[Liu et al. 2005]	KNDA	CUFS		306:300	87.67%
	[Wang and Tang 2009]	RS-LDA	CUFS	Multiscale MRF	306:300	96.3%
	[Zhong et al. 2007]		CUFS	E-HMM	—	95.24%
	[Liu et al. 2007]		CUFS	E-HMM+Selective ensemble	—	100%
Projection based	[Klare and Jain 2010b]	Common representation	CUFS	SIFT	100:300	96.47%
	[Sharma and Jacobs 2011]	PLS	CUHK		88:100	93.60%
	[Choi et al. 2012]	PLS regression	CUFS,CUFSF	Gabor and CCS-POP	0:1800	99.94%
Feature based	[Klare and Jain 2010b]	NN	CUFS	SIFT	100:300	97.87%
	[Khan et al. 2012]	NN	CUFS	Self Similarity	161:150	99.53%
	[Galoogahi and Sim 2012a]	NN,PMK,Chi-square	CUFS	LRBP	—	99.51%
	[Kiani Galoogahi and Sim 2012]	NN,Chi-square	CUFS	Gabor Shape	306:300	99.14%
	[Bhatt et al. 2010]	Weighted Chi-square	CUFS	EUCLBP	78:233	94.12%
	[Galoogahi and Sim 2012b]	NN,Chi-square	CUFS	HAOG	306:300	100.00%
	[Kiani Galoogahi and Sim 2012]	NN,Chi-square	CUFSF	Gabor Shape	500:694	96.32%
	[Galoogahi and Sim 2012a]	NN,PMK,Chi-square	CUFSF	LRBP	—	91.12%
	[Pramanik and Bhattacharjee 2012]	K-NN	CUHK	Geometric features	108:80	80.00%
	[Bhatt et al. 2010]	Weighted Chi-square	IIT-D	EUCLBP	58:173	78.58%

Khan et al. [Khan et al. 2012] 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. [Galoogahi and Sim 2012a] 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) [Lazebnik et al. 2006]. LRBP benefits from low computational complexity and the fact that there is no critical parameter to be tuned [Galoogahi and Sim 2012a].

Galoogahi et al. consequently proposed another two face descriptors: Gabor Shape [Kiani Galoogahi and Sim 2012] which is variant of Gabor features and Histogram of Averaged Oriented Gradient (HAOG) features [Galoogahi and Sim 2012b] which is variant of HOG for sketch/photo directly matching, the latter achieves perfect 100% accuracy on the CUFS dataset.

Klare et al. [Klare and Jain 2010b] 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 [Klare and Jain 2010b]. 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 [Anand et al. 2013], 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. [Pramanik and Bhattacharjee 2012] 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.

### 3.4. Forensic sketch face recognition

Forensic sketches pose greater challenges than viewed sketch recognition because forensic sketches contain less, incomplete or inaccurate information. This issue due to the subjectivity of the description, and imperfection of the witness' memory.

There are therefore two sets of challenges in forensic sketch-based recognition: (1) recognizing across modalities and (2) performing recognition despite inaccurate, incomplete and harder to align depictions of the face. 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. [Uhl and da Vitoria Lobo 1996] 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. [Klare et al. 2011] 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 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. [Bhatt et al. 2012] 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 [Bhatt et al. 2012]. The result outperforms [Klare et al. 2011] using only the forensic set as gallery.

### 3.5. Composite sketches based face recognition

Three studies have thus far focused on face recognition using composite sketches. The first one uses both local and global features to represent sketches and is proposed by Yuen et al. [Yuen and Man 2007]. This method required user input in the form of relevance feedback in the recognition phase. The second two focus on holistic [Klare and Jain 2013] and component based [Han et al. 2013] representations respectively.

The holistic method [Klare and Jain 2013] 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 [Lowe 2004] and multi-scale local binary pattern (MLBP) [Ojala et al. 2002] invariant features are extracted from each patch. With this feature encoding, as improved version of the common representation intuition from [Klare and Jain 2010b] is applied, followed by RS-LDA [Wang and Tang

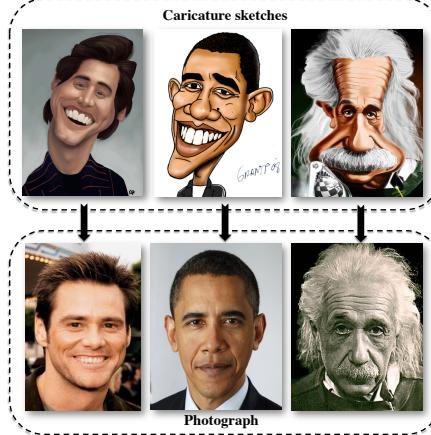


Fig. 6. Caricatures and corresponding mugshots

2006b] 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 [Han et al. 2013] uses similarities between individual facial components to compute an overall sketch to mugshot match score. Facial landmarks in composite sketches and photos are automatically detected by an active shape model (ASM) [Milborrow and Nicolls 2008]. Multiscale 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.

### 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 [Lanckriet et al. 2004].

Studies have suggested that people may encode faces in a caricatured manner [Sinha et al. 2006]. Moreover they may be more capable of recognizing a familiar person through a caricature than an accurate rendition [Mauro and Kubovy 1992; Rhodes et al. 1987]. The effectiveness of a caricature is due to its emphasis of deviations from average faces [Klare et al. 2012]. Developing efficient approaches in caricature based face recognition could help drive more robust and reliable face and heterogeneous face recognition systems.

Klare et al. [Klare et al. 2012] 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 [Ojala et al. 2002] 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. [Ouyang et al. 2014] 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 IV 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. As viewed sketch datasets exhibit near perfect alignment and detail correspondence between sketches and photos, well designed approaches achieve near perfect accuracies. Note that some results on the same dataset are not directly comparable because of differing test set sizes.

*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. [Nizami et al. 2009] demonstrated significant intra-personal variation in sketches drawn by different artists. This may challenge systems that rely on learning a simple uni-modal 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. [Choi et al. 2012], 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

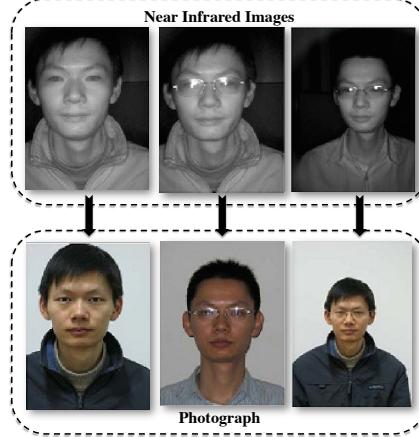


Fig. 7. VIS and NIR face images.

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 is being hampered by lack of readily obtainable forensic datasets. Constructing realistic and freely available datasets should be a priority [Choi et al. 2012].

*Training Data Source.* Many effective SBFR studies have leveraged annotated training data to learn projections and/or classifiers [Klare and Jain 2010b]. As interest has shifted onto forensic sketches, standard practice has been to train such models on viewed-sketch datasets and test on forensic datasets [Klare et al. 2011]. An interesting question going forward is whether this is the right strategy. 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* [Pan and Yang 2010], to be solved; as well as further motivating the creation of larger forensic-sketch datasets with which it will be possible to discover whether training on forensic pairs is more effective than training on viewed-pairs.

#### 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  $0.7\mu m - 1.4\mu m$ ) [Klare and Jain 2010a]. 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 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.

Table V. Summary of existing NIR-VIS benchmark datasets

Dataset	Wavelength	No.of Subjects	No.of Images	3D	Pose variations	Expression variations
CASIA HFB [Li et al. 2009]	850nm	100	992	✓	✗	✗
CASIA NIR-VIS 2.0 [Li et al. 2013]	850nm	725	17580	✓	✓	✓
Cross Spectral Dataset [Zhang et al. 2010]	800-1000nm	430	4189	✓	✓	✗
PolyU [Zhang et al. 2010]	780-1100nm	335	33500	✓	✓	✓

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 four main heterogeneous datasets covering the NIR-VIS condition. The CASIA HFB dataset [Li et al. 2009], 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 [Li et al. 2013] 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 [Goswami et al. 2011] 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 [Zhang et al. 2010] 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. The active light source used to create this dataset is in the NIR spectrum between 780mm to 1,100mm.

A summary of the main NIR-VIS datasets can be found in Tab V. 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. [Wang et al. 2009] 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 [Xie and Lam 2006], 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. [Chen et al. 2009] 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 [Wang et al. 2009], 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. [Pengfei et al. 2012] 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.

#### 4.3. Projection based approaches

Lin et al. [Lin and Tang 2006] 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. [Yi et al. 2007] proposed a learning-based approach for cross-modality matching. In this approach, linear discriminant analysis (LDA) is used to extract features and reduce the dimension of the feature vectors. Then, a canonical correlation analysis (CCA) [Hotelling 1992] 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 limitation, Liao et al. [Liao et al. 2009] 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) [Shengcui et al. 2007] 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. [Klare and Jain 2010a] build on [Liao et al. 2009], 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 [Klare and Jain 2010a] learn a shared projection with reduced over fitting. Finally, NN and Sparse Representation based matching are performed for matching.

Lei et al. [Lei and Li 2009] 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 feature 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 [Yan et al. 2007] and spectral regression [Cai et al. 2007] combined with regularization techniques. They later improve the same framework [Lei et al. 2012], to better exploit the cross-modality supervision and sample locality.

Huang et al. [Huang et al. 2013] 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.

#### 4.4. Feature based approaches

Zhu et al. [Zhu et al. 2013b] interpret the VIS-NIR problem as a highly illumination-variant task. They address it by designing an effective illumination invariant descrip-

tor, the logarithm gradient histogram (LGH). This outperforms the LBP and SIFT descriptors used by [Liao et al. 2009] and [Klare and Jain 2010a] respectively. As a purely feature-based approach, no training data is required.

Huang et al. [Huang et al. 2012], 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. [Goswami et al. 2011] 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 [Gong and Zheng 2013] 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.

Finally, Zhu et al. [Zhu et al. 2014] 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.

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) often play a greater role in bridging the gap. This is because it is to some extent possible to model the cross-modal lighting gap more directly and explicitly than other HFR scenarios that 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. Owing 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 [Bowyer et al. 2006]. 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 ex-

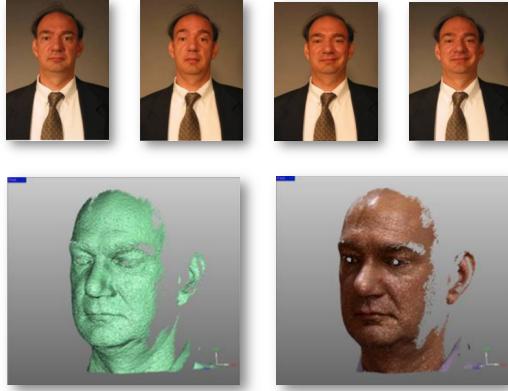


Fig. 8. 2D images and 3D images from FRGC dataset.

ample, 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. 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.

### 5.1. Datasets

The face Recognition Grand Challenge (FRGC) V2.0 dataset<sup>2</sup> 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, and both range and texture cues are included. An example from the FRGC V2.0 dataset is shown in Fig 8.

UHDB11 [Toderici et al. 2014] 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.

### 5.2. Projection based approaches

Yang et al. [Yang et al. 2008] 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. [Huang et al. 2009] 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.

<sup>2</sup>Downloadable at <http://www.nist.gov/itl/iad/ig/frgc.cfm>

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. [Huang et al. 2010] 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.

Petrou et al. [Toderici et al. 2010] introduced a 2D-3D face recognition method based on a novel bidirectional relighting algorithm. A subject-specific 3D annotated model is built by using its raw 3D data and 2D texture. With this model, 2D images are projected onto a normalized image space. The lighting from the probe image is then transferred to the gallery image for more robust matching.

Rama et al. [Rama et al. 2006] generated 180-degree cylindrical face images in the 3D enrollment phase. At test time, after dimensionally reduction by  $P^2CA$ , 2D face images are compared against all subwindows of the 180-degree enrollment image. The score of the best-matching subwindow is taken as the score of the probe image.

### 5.3. Feature based approaches

A biologically inspired feature, Oriented Gradient Maps (OGMs), is introduced by Huang et al. in [Huang et al. 2012]. 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 [Huang et al. 2009; Huang et al. 2010].

### 5.4. Summary and Conclusions

2D image based face recognition systems often fail in situations where facial depictions exhibit strong pose and 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 have done some kind of explicit 3D reasoning about matching the 2D probe image to the 3D model [Toderici et al. 2010; Rama et al. 2006], and others that have relied on discriminative features and learning a single cross-domain mapping such as CCA [Huang et al. 2009; Huang et al. 2010; Huang et al. 2012; Yang et al. 2008]. 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 be necessary.

## 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 and surveillance. Interestingly we know this should be possible, because humans can recognize low-resolution faces down to  $16 \times 16$  pixels [Sinha et al. 2006]. 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 [Yang et al. 2010; van Ouwerkerk 2006] 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. Synthesis based approaches

Hennings-Yeomans et al. [Hennings-Yeomans et al. 2008] 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 [Hennings-Yeomans et al. 2009]. However these methods have high computational cost.

Zou et al. [Zou and Yuen 2012] 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. [Zou and Yuen 2010] 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. [Jia and Gong 2005] 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. [Jiang et al. 2012] 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.

Huang et al. [Huang and He 2011] proposed a nonlinear mapping based 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. [Gunturk et al. 2003] 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.

### 6.2. Projection-based approaches

Li et al. [Li et al. 2010] 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. [Zhou et al. 2011] 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. [Wang et al. 2013] 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 [Sharma and Jacobs 2011] 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. [Biswas et al. 2012] to simultaneously embed LR and HR images in a common space. In this common space, the distance between LR and HR approximates the distance between corresponding HR images.

Shekhar et al. [Shekhar et al. 2011] 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.

Ren et al. [Ren et al. 2012] 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. [Siena et al. 2013] 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.

Ren et al. [Ren et al. 2011] addressed discriminative subspace learning in LR-HR recognition from the angle of evaluating and combining multiple encodings of each image type. Different image descriptors including RsL2, LBP, Gradientface and IMED are considered and a multiple kernel learning strategy used to learn a good projection with a weighted combination of them.

In [Li et al. 2009], 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 than super-resolution first strategies, while being applicable to other types of ‘degraded’ images besides LR, such as blur and occlusion.

Deng et al. [Deng et al. 2010] utilized color information to tackle LR face recognition as color cues are less variant to resolution change. They improved on [Li et al. 2010] to introduce a regularized coupled mapping to project both LR and HR face images into a common discriminative space.

### 6.3. 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. Interest-

Table VI. The details of plastic surgery database

Plastic Surgery Procedure	Number of Individuals
Dermabrasion	32
Brow lift(Forehead surgery)	60
Otoplasty(Ear surgery)	74
Blepharoplasty(Eyelid surgery)	105
Rhinoplasty(Nose surgery)	192
Others(Mentoplasty, Malar augmentation, Craniofacial,Lip,augmentation,Fat,injection)	56
Skin peeling(Skin resurfacing)	73
Rhytidectomy(Face lift)	308

ingly, while this discriminative cue has been used relatively less frequently in SBFR, NIR and 3D matching, it has been used almost throughout in HR-LR matching.

*LR Dataset realism.* With few exceptions [Shekhar et al. 2011], the vast majority of LR-HR studies simulate LR data by downsampling 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 [Hospedales et al. 2012; Shekhar et al. 2011]. Thus existing systems are likely to under perform in practice. A benchmark dataset of realistic LR surveillance captures and associated HR mugshots would be advantageous to drive research. This may lead into integrating super-resolution and recognition with simultaneous de-blurring [Cho et al. 2007; Levin et al. 2011], re-lighting [Shekhar et al. 2011] and pose alignment [Zhang and Gao 2009].

## 7. WITHIN-MODALITY HETEROGENEITY

Independently of the sensing modality used, other covariates such as disguises and plastic surgery are also key factors that affect the performance of face recognition systems. These do not change the intrinsic quality or type of the images, but can still provide a strong image-space transformation between probe and gallery faces. Disguise is an interesting and challenging covariate of face recognition. It includes intentional or unintentional changes through which one can either hide his/her identity or appear to be someone else. Dhamecha et al. [Dhamecha et al. 2013] have summarized the existing disguise detection and face recognition algorithms. In this survey, we rather focus on matching across plastic surgery variations.

With plastic surgery requiring reduced cost and time, its popularity has increased dramatically. It provides a significant covariate that seriously degrades the performance of conventional fare recognition systems, for example loosing 25-30% rank 1 accuracy [Singh et al. 2010]. Aside from its generally rising popularity, this result provides an incentive for individuals to conceal their identity and evade recognition via plastic surgery. Hence it is of interest to develop HFR systems capable of recognition across pre and post-surgical images.

### 7.1. Database

Singh et al. [Singh et al. 2010] provided a face database which encompasses 900 individuals who have plastic surgery. There are 900 subjects in the database corresponding to 1800 full frontal face images. A wide range of cases are included, such as nose surgery, as shown in Fig (9), eyelid surgery, skin peeling, brow lift, and face lift. The details of images in the plastic surgery database are given in Tab VI.

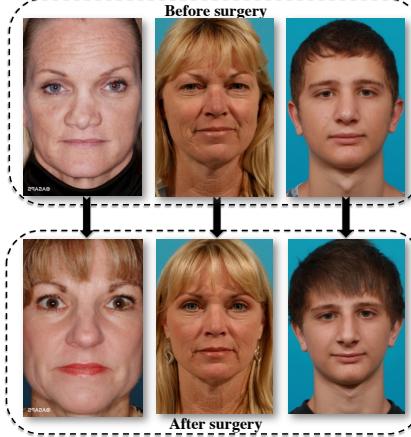


Fig. 9. Pre and post-operative samples from the plastic surgery database.

## 7.2. Feature-based approaches

Aggarwal et al. [Aggarwal et al. 2012] locate facial components using active shape models, and then learn a sparse coding representation for each component. Components are matched across domains according to their sparse coding reconstruction errors. The overall face match is performed by sum fusion of the per-component scores.

LakshmiPrabha et al. [LakshmiPrabha and Majumder 2012] proposed a face recognition system invariant to plastic surgery using shape local binary texture (SLBT) feature in a two step cascade. In the first step, ASM is used to warp two images to be matched into alignment for global appearance based comparison – thus partially addressing changes in face structure due to surgery. In the second step of the cascade per-component comparisons are made. However, this method requires manually annotated facial landmarks.

Bhatt et al. [Bhatt et al. 2013] developed an evolutionary granular algorithm to address the issues of automatic matching of face across plastic surgery variations. In this method, facial patches (granules) at multiple locations and resolutions are extracted, and two features (SIFT and EUCLBP) used to describe them. Evolutionary algorithms are used in order to select among granules, select the feature type for each, and determine their weighting in comparison using weighted Chi-square distance.

Liu et al. [Liu et al. 2013] proposed an ensemble of Gabor Patch classifiers via Rank-Order list Fusion (GPROF). Dividing face images into regular patches, Gabor features together with Fisher Linear Discriminant Analysis is exploited to generate a descriptor for each patch. The descriptors are then further transformed into a new invariant representation similar in inspiration to common representation [Klare and Jain 2010b]: the rank ordering of their most similar gallery patches. Finally, the overall score fuses the result of each patch.

## 7.3. Conclusions and discussion

The key challenge of heterogeneity due to plastic surgery, is of course the intra-class variability introduced by the surgical process. Some previous cases of heterogeneity discussed in this survey such as sketch and NIR have more or less uniform and non-geometry distorting transformations (assuming good frontal poses). In contrast, surgical modifications can take a variety of forms including: similarly global but non structural/geometric modifications (e.g., skin resurfacing), global and structure/geometry

distorting transforms (e.g., face lift), and highly localized transformations (e.g., nose surgery) [Singh et al. 2010; Bhatt et al. 2013]. The multi-modality and non-uniformity of surgical transformations may explain why all of the studies so far have primarily been feature based approaches, rather than the synthesis and projection based approaches commonly seen in other HFR contexts. The prevalence of localized transformations also explains the heavier reliance in this area on component-based representations compared to other HFR settings. If a single facial component is modified, then the matching noise introduced is limited to the score of a single component.

*Databases.* An interesting issue is whether HFR systems for plastic surgery should be trained on non-surgery, or surgery databases. In the latter case, significantly more training data is likely to be available, but discriminatively trained models [Bhatt et al. 2013] have then not been exposed to the variations which they will be tested on, and will thus under-perform. In the latter case the reverse is true, models will have been exposed to appropriate cross-modal variations at training time, but the amount of training data in HFR databases is less. These approaches were compared in [Singh et al. 2010], where training with 360 surgical pairs was reported to give better results than on 900 non-surgical pairs. This issue is somewhat analogous to the previously mentioned question of whether to train on viewed or forensic sketches for forensic sketch recognition.

## 8. CONCLUSION AND 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, 3D and plastic surgery. In case there is a strong driving application factor in security/law-enforcement/forensics. We draw the following observations and conclusions:

### 8.1. 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 besides plastic surgery. 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 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.

*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 representations of course exploit the specific face structure most heavily. Component-based methods are commonly used in recognition across plastic surgery. However, interestingly, the majority of methods reviewed do not ex-

ploit 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 do leverage 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 realistic LR recognition where accurate alignment is impossible.

*Dataset over-fitting.* Recognition tasks in broader computer vision have recently been shown to suffer from over-fitting to entire datasets, as researchers engineer methods to maximize benchmark scores on insufficiently diverse datasets [Torralba and Efros 2011]. Current HFR datasets, notably in Sketch, NIR and plastic surgery 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.

## 8.2. 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 [Wang and Tang 2009; Bhatt et al. 2012; Wang and Tang 2009; Zhang et al. 2011], are much smaller than those used in homogeneous face recognition [Huang et al. 2007] and broader computer vision [Deng et al. 2009] problems. As larger heterogeneous datasets are collected in future, more sophisticated non-linear models may gain the edge.

*Openness & Components.* Many studies make a contribution both to feature representation, and to some projection/synthesis/matching method. It is often hard to dis-entangle which part provides the benefit. It would be beneficial for the field if researchers: (i) always present their experiments breaking out feature and learning model contributions to the overall results, and (ii) released their features and learning methods, so that those who want to focus on one part can take best practice from the other without re-inventing the wheel.

*Alignment.* Unlike homogeneous face recognition which has moved onto recognition ‘in the wild’ [Huang et al. 2007], 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 [Klare et al. 2011] 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, 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 [Lampert et al. 2009]. They have successfully been applied to cross-domain modeling for person (rather than face) recognition [Layne et al. 2012]. 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 [Klare et al. 2012]. Recent studies [Ouyang et al. 2014] 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 rapidly improving facial attribute extraction techniques [Luo et al. 2013], 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 attempt 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.

*Technical Methodologies.* CCA, PLS, Sparse Coding and various generalizations thereof have been used extensively in the studies reviewed here. Some promising methodologies that have been under-exploited in HFR include metric learning and deep learning. Metric learning approaches have had great success in the related area of cross-view person recognition [Hirzer et al. 2012]. Early studies have shown the potential for HFR to improve the cross-domain and matching steps [Bhatt et al. 2012; Bhatt et al. 2013].

Deep learning in contrast has transformed broader computer vision problems by learning significantly more effective feature representations [Krizhevsky et al. 2012]. Deep features may provide scope for bridging the cross-modal gap. For example, they have recently been applied for pose-invariant face-recognition [Zhu et al. 2013a]. However, this requires significantly larger scale heterogeneous datasets than is available for most of the HFR settings reviewed here.

## REFERENCES

- G. Aggarwal, S. Biswas, P.J. Flynn, and K.W. Bowyer. 2012. A sparse representation approach to face matching across plastic surgery. In *IEEE Workshop on Applications of Computer Vision (WACV)*. 113–119.
- Le Anand, Mehran Kafaian, Songfan Yang, and Bir Bhanu. 2013. Reference Based Person ReIdentification. In *Advanced Video and Signal Based Surveillance (AVSS)*. 244–249.
- P.N. Belhumeur, J.P. Hespanha, and D. Kriegman. 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *TPAMI* (1997), 711–720.
- H.S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. 2010. On matching sketches with digital face images. In *Biometrics: Theory Applications and Systems (BTAS)*. 1–7.
- H.S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. 2012. Memetically Optimized MCWLD for Matching Sketches With Digital Face Images. *IEEE Transactions on Information Forensics and Security* (2012), 1522–1535.
- H.S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. 2013. Recognizing Surgically Altered Face Images Using Multiobjective Evolutionary Algorithm. *IEEE Transactions on Information Forensics and Security* (2013), 89–100.
- IQ Biometrix. 2011. FACE 4.0. *IQ Biometrix* (2011).
- Soma Biswas, Kevin W. Bowyer, and Patrick J. Flynn. 2012. Multidimensional Scaling for Matching Low-Resolution Face Images. *TPAMI* (2012), 2019–2030.
- Kevin W. Bowyer, Kyong Chang, and Patrick Flynn. 2006. A Survey of Approaches and Challenges in 3D and Multi-modal 3D + 2D Face Recognition. *CVIU* (2006), 1–15.
- Deng Cai, Xiaofei He, and Jiawei Han. 2007. Spectral Regression for Efficient Regularized Subspace Learning. In *ICCV*. 1–8.
- Jie Chen, Dong Yi, Jimei Yang, Guoying Zhao, S.Z. Li, and M. Pietikainen. 2009. Learning mappings for face synthesis from near infrared to visual light images. In *CVPR*. 156–163.
- Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu. 2000. A new LDA-based face recognition system which can solve the small sample size problem. *PR* (2000), 1713–1726.

- Sunghyun Cho, Y. Matsushita, and Seungyong Lee. 2007. Removing Non-Uniform Motion Blur from Images. In *ICCV*. 1–8.
- Jonghyun Choi, A. Sharma, D.W. Jacobs, and L.S. Davis. 2012. Data insufficiency in sketch versus photo face recognition. In *CVPR*. 1–8.
- Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. 248–255.
- Zhao-Xiong Deng, Dao-Qing Dai, and Xiao-Xin Li. 2010. Low-Resolution Face Recognition via Color Information and Regularized Coupled Mappings. In *Chinese Conference on Pattern Recognition (CCPR)*. 1–5.
- T.I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa. 2013. Disguise detection and face recognition in visible and thermal spectrums. In *The International Conference on Biometrics (ICB)*. 1–8.
- Frontex. 2010. BIOPASS II: Automated biometric border crossing systems based on electronic passports and facial recognition: RAPID and SmartGate.
- Charlie D. Frowd, Peter J. B. Hancock, and Derek Carson. 2004. EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *Journal ACM Transactions on Applied Perception(TAP)* (2004), 19–39.
- H.K. Galoogahi and T. Sim. 2012a. Face sketch recognition by Local Radon Binary Pattern: LRPB. In *ICIP*. 1837–1840.
- H.K. Galoogahi and T. Sim. 2012b. Inter-modality Face Sketch Recognition. In *ICME*. 224–229.
- Xinbo Gao, Juanjuan Zhong, Jie Li, and Chunna Tian. 2008a. Face Sketch Synthesis Algorithm Based on E-HMM and Selective Ensemble. *IEEE Transactions on Circuits and Systems for Video Technology* (2008), 487–496.
- Xinbo Gao, Juanjuan Zhong, Dacheng Tao, and Xuelong Li. 2008b. Local face sketch synthesis learning. *Neurocomputing* (2008), 1921–1930.
- L. Gibson. 2008. *Forensic Art Essentials*.
- David E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*.
- Dihong Gong and Jiangyu Zheng. 2013. A Maximum Correlation Feature Descriptor for Heterogeneous Face Recognition. In *Asian Conference on Pattern Recognition (ACPR)*. 135–139.
- D. Goswami, Chi-Ho Chan, D. Windridge, and J. Kittler. 2011. Evaluation of face recognition system in heterogeneous environments (visible vs NIR). In *ICCV*. 2160–2167.
- B.K. Gunturk, AU. Batur, Y. Altunbasak, M.H. Hayes, and R.M. Mersereau. 2003. Eigenface-domain super-resolution for face recognition. *TIP* (2003), 597–606.
- G. Wells and L. Hasel. 2007. Facial composite production by eyewitnesses. *Current Directions in Psychological* (2007), 6–10.
- Hu Han, B.F. Klare, K. Bonnen, and A.K. Jain. 2013. Matching Composite Sketches to Face Photos: A Component-Based Approach. *IEEE Transactions on Information Forensics and Security* (2013), 191–204.
- P.H. Hennings-Yeomans, S. Baker, and B.V.K.V. Kumar. 2008. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *CVPR*. 1–8.
- P.H. Hennings-Yeomans, B.V.K.V. Kumar, and S. Baker. 2009. Robust low-resolution face identification and verification using high-resolution features. In *ICIP*. 33–36.
- Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. 2012. Relaxed Pairwise Learned Metric for Person Re-identification. In *ECCV*. 780–793.
- T. M. Hospedales, S. Gong, and T. Xiang. 2012. A real-time dictionary based approach to super-resolution for surveillance. In *SPIE Security and Defence Conference*. 1–8.
- Harold Hotelling. 1992. Relations Between Two Sets of Variates. In *Breakthroughs in Statistics*. 162–190.
- Di Huang, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. 2009. Asymmetric 3D/2D Face Recognition Based on LBP Facial Representation and Canonical Correlation Analysis. In *ICIP*. 3325–3328.
- Di Huang, M. Ardabilian, Yunhong Wang, and Liming Chen. 2010. Automatic Asymmetric 3D-2D Face Recognition. In *International Conference on Pattern Recognition (ICPR)*. 1225–1228.
- Di Huang, M. Ardabilian, Yunhong Wang, and Liming Chen. 2012. Oriented Gradient Maps based automatic asymmetric 3D-2D face recognition. In *The IAPR International Conference on Biometrics (ICB)*. 125–131.
- De-An Huang and Yu-Chiang Frank Wang. 2013. Coupled Dictionary and Feature Space Learning with Applications to Cross-Domain Image Synthesis and Recognition. In *ICCV*. 2496–2503.
- Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report. University of Massachusetts, Amherst.

- Hua Huang and Huiting He. 2011. Super-Resolution Method for Face Recognition Using Nonlinear Mappings on Coherent Features. *IEEE Transactions on Neural Networks* (2011), 121–130.
- Likun Huang, Jiwen Lu, and Yap-Peng Tan. 2012. Learning modality-invariant features for heterogeneous face recognition. In *International Conference on Pattern Recognition (ICPR)*. 1683–1686.
- Xiangsheng Huang, Zhen Lei, Mingyu Fan, Xiao Wang, and S.Z. Li. 2013. Regularized Discriminative Spectral Regression Method for Heterogeneous Face Matching. *TIP* (2013), 353–362.
- Kui Jia and Shaogang Gong. 2005. Multi-modal tensor face for simultaneous super-resolution and recognition. In *ICCV*. 1683–1690.
- Junjun Jiang, Ruimin Hu, Zhen Han, Kebin Huang, and Tao Lu. 2012. Graph discriminant analysis on multi-manifold (GDAMM): A novel super-resolution method for face recognition. In *ICIP*. 1465–1468.
- Z. Khan, Yiqun Hu, and A. Mian. 2012. Facial Self Similarity for Sketch to Photo Matching. In *Digital Image Computing Techniques and Applications (DICTA)*. 1–7.
- Hamed Kiani Galoogahi and Terence Sim. 2012. Face photo retrieval by sketch example. In *The 20th ACM international conference on Multimedia*. 949–952.
- B.F. Klare, S.S. Bucak, A.K. Jain, and T. Akgul. 2012. Towards automated caricature recognition. In *The IAPR International Conference on Biometrics (ICB)*. 139–146.
- B. Klare and A.K. Jain. 2010a. Heterogeneous Face Recognition: Matching NIR to Visible Light Images. In *International Conference on Pattern Recognition (ICPR)*. 1513–1516.
- Brendan Klare and Anil K. Jain. 2010b. Sketch-to-photo matching: a feature-based approach. In *Biometric Technology for Human Identification VII.SPIE*. 1–10.
- B.F. Klare, Zhifeng Li, and A.K. Jain. 2011. Matching Forensic Sketches to Mug Shot Photos. *TPAMI* (2011), 639–646.
- Brendan F. Klare and Anil K. Jain. 2013. Heterogeneous Face Recognition Using Kernel Prototype Similarities. *TPAMI* (2013), 1410–1422.
- Scott Klum, Hu Han, Anil K. Jain, and Brendan Klare. 2013. Sketch based face recognition: Forensic vs. composite sketches. In *The International Conference on Biometrics (ICB)*. 1–8.
- Seong G. Kong, Jingu Heo, Besma R. Abidi, Joonki Paik, and Mongi A. Abidi. 2005. Recent advances in visual and infrared face recognition—a review. *CVIU* (2005), 103–135.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1–9.
- N.S. Lakshmiprabha and S. Majumder. 2012. Face recognition system invariant to plastic surgery. In *Intelligent Systems Design and Applications (ISDA)*. 258–263.
- C. H. Lampert, H. Nickisch, and S. Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*. 951–958.
- Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. 2004. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research* (2004), 27–72.
- Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. 2012. Person Re-identification by Attributes. In *BMVC*. 1–8.
- S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*. 2169–2178.
- Zhen Lei and S.Z. Li. 2009. Coupled Spectral Regression for matching heterogeneous faces. In *CVPR*. 1123–1128.
- Zhen Lei, Changtao Zhou, Dong Yi, Anil K. Jain, and Stan Z. Li. 2012. An improved coupled spectral regression for heterogeneous face recognition.. In *The IAPR International Conference on Biometrics (ICB)*. 7–12.
- A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. 2011. Understanding Blind Deconvolution Algorithms. *TPAMI* (2011), 2354–2367.
- Bo Li, Hong Chang, Shiguang Shan, and Xilin Chen. 2009. Coupled Metric Learning for Face Recognition with Degraded Images. In *Advances in Machine Learning*. In proceeding of Advances in Machine Learning, First Asian Conference on Machine Learning, ACML, 220–233.
- Bo Li, Hong Chang, Shiguang Shan, and Xilin Chen. 2010. Low-Resolution Face Recognition via Coupled Locality Preserving Mappings. *Signal Processing Letters, IEEE* (2010), 20–23.
- S.Z. Li, Zhen Lei, and Meng Ao. 2009. The HFB Face Database for Heterogeneous Face Biometrics research. In *CVPR*. 1–8.
- S.Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. 2013. The CASIA NIR-VIS 2.0 Face Database. In *CVPR*. 348–353.

- Shengcai Liao, Dong Yi, Zhen Lei, Rui Qin, and Stan Z. Li. 2009. Heterogeneous Face Recognition from Local Structures of Normalized Appearance. In *Proceedings of the Third International Conference on Advances in Biometrics*. 209–218.
- Dahu Lin and Xiaoou Tang. 2006. Inter-modality Face Recognition. In *ECCV*. 13–26.
- Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. 2005. A nonlinear approach for face sketch synthesis and recognition. In *CVPR*. 1005–1010.
- Sifei Liu, Dong Yi, Zhen Lei, and S.Z. Li. 2012. Heterogeneous face image matching using multi-scale features. In *The IAPR International Conference on Biometrics (ICB)*. 79–84.
- Wei Liu, Xiaoou Tang, and Jianzhuang Liu. 2007. Bayesian tensor inference for sketch-based facial photo hallucination. In *The international joint conference on Artificial intelligence*. 2141–2146.
- Xin Liu, Shiguang Shan, and Xilin Chen. 2013. Face Recognition after Plastic Surgery: A Comprehensive Study. In *ACCV*. 565–576.
- DavidG. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV* (2004), 91–110.
- Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2013. A Deep Sum-Product Architecture for Robust Facial Attributes Analysis. In *ICCV*. 2864–2871.
- A. M. Martinez and R. Benavente. 1998. *The AR Face Database*. Technical Report. CVC Technical Report 24.
- Robert Mauro and Michael Kubovy. 1992. Caricature and face recognition. *Memory & Cognition* (1992), 433–440.
- Dawn McQuiston-Surrett, Lisa D. Topp, and Roy S. Malpass. 2006. Use of facial composite systems in US law enforcement agencies. *Psychology, Crime and Law* (2006), 505–517.
- K. Messer, J. Matas, J. Kittler, and K. Jonsson. 1999. XM2VTSDB: The Extended M2VTS Database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*. 72–77.
- Stephen Milborrow and Fred Nicolls. 2008. Locating Facial Features with an Extended Active Shape Model. In *ECCV*. 504–513.
- B. Moghaddam and A. Pentland. 1997. Probabilistic visual learning for object representation. *TPAMI* (1997), 696–710.
- H. Nejati and T. Sim. 2011. A study on recognizing non-artistic face sketches. In *IEEE Workshop on Applications of Computer Vision (WACV)*. 240–247.
- H. Nizami, J.P. Adkins-Hill, Yong Zhang, J.R. Sullins, C. McCullough, S. Canavan, and Lijun Yin. 2009. A biometric database with rotating head videos and hand-drawn face sketches. In *Biometrics: Theory, Applications, and Systems*. 1–6.
- T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI* (2002), 971–987.
- Shuxin Ouyang, Tim Hospedales, Yi zhe Song, and Xueming Li. 2014. Cross-modal face matching: beyond viewed sketches. *Accepted by ACCV* (2014).
- Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2010), 1345–1359.
- Xiong Pengfei, Lei Huang, and Changping Liu. 2012. A method for heterogeneous face image synthesis. In *The IAPR International Conference on Biometrics (ICB)*. 1–6.
- P.J. Phillips, Hyeyonjoon Moon, S.A. Rizvi, and P.J. Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *TPAMI* (2000), 1090–1104.
- S. Pramanik and D. Bhattacharjee. 2012. Geometric feature based face-sketch recognition. In *Pattern Recognition, Informatics and Medical Engineering (PRIME)*. 409–415.
- A. Rama, F. Tarres, Davide Onofrio, and S. Tubaro. 2006. Mixed 2D-3D Information for Pose Estimation and Face Recognition. In *Acoustics, Speech and Signal Processing*. 361–364.
- Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. 2011. Low resolution facial image recognition via multiple kernel criterion. In *First Asian Conference on Pattern Recognition (ACPR)*. 204–208.
- Chuan-Xian Ren, Dao-Qing Dai, and Hong Yan. 2012. Coupled Kernel Embedding for Low-Resolution Face Image Recognition. *TIP* (2012), 3770–3783.
- Gillian Rhodes, Susan Brennan, and Susan Carey. 1987. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology* (1987), 473–497.
- A. Sharma and D.W. Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *CVPR*. 593–600.
- S. Shekhar, V.M. Patel, and R. Chellappa. 2011. Synthesis-based recognition of low resolution faces. In *The 2011 International Joint Conference on Biometrics (IJCB)*. 1–6.

- Liao Shengcai, Zhu Xiangxin, Lei Zhen, Zhang Lun, and Li StanZ. 2007. Learning Multi-scale Block Local Binary Patterns for Face Recognition. In *Advances in Biometrics*. 828–837.
- S. Siena, V.N. Boddeti, and B.V.K.V. Kumar. 2013. Maximum-Margin Coupled Mappings for cross-domain matching. In *Biometrics: Theory, Applications and Systems (BTAS)*. 1–8.
- R. Singh, M. Vatsa, H.S. Bhatt, S. Bharadwaj, A. Noore, and S.S. Nooreyezdan. 2010. Plastic Surgery: A New Dimension to Face Recognition. *IEEE Transactions on Information Forensics and Security* (2010), 441–448.
- P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. 2006. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proc. IEEE* (2006), 1948–1962.
- Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep Learning Face Representation by Joint Identification-Verification. *Technical report, arXiv* (2014), 1–9.
- Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. 2006. Face recognition from a single image per person: A survey. *PR* (2006), 1725–1745.
- Xiaoou Tang and Xiaogang Wang. 2002. Face photo recognition using sketch. In *ICIP*. 257–260.
- Xiaoou Tang and Xiaogang Wang. 2003. Face sketch synthesis and recognition. In *ICCV*. 687–694.
- K. Taylor. 2001. *Forensic Art and Illustration*. CRC Press.
- George Toderici, Georgios Evangelopoulos, Tianhong Fang, Theoharis Theoharis, and IoannisA. Kakadiaris. 2014. UHDB11 Database for 3D-2D Face Recognition. In *Image and Video Technology*. 73–86.
- G. Toderici, G. Passalis, S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Theoharis, and I. A. Kakadiaris. 2010. Bidirectional relighting for 3D-aided 2D face recognition. In *CVPR*. 2721–2728.
- Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR*. 1521–1528.
- Matthew Turk and Alex Pentland. 1991. Eigenfaces for Recognition. *J. Cognitive Neuroscience* (1991), 71–86.
- Jr. Uhl, R.G. and N. da Vitoria Lobo. 1996. A framework for recognizing a facial image from a police sketch. In *CVPR*. 586–593.
- J.D. van Ouwerkerk. 2006. Image super-resolution survey. *Image and Vision Computing* (2006), 1039 – 1052.
- Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. 2014. A Comprehensive Survey to Face Hallucination. *IJCV* 1 (2014), 9–30.
- Rui Wang, Jimei Yang, Dong Yi, and StanZ. Li. 2009. An Analysis-by-Synthesis Method for Heterogeneous Face Biometrics. In *Advances in Biometrics*. 319–326.
- Xiaogang Wang and Xiaoou Tang. 2004a. Dual-space linear discriminant analysis for face recognition. In *CVPR*. 564–569.
- Xiaogang Wang and Xiaoou Tang. 2004b. Random sampling LDA for face recognition. In *CVPR*. 259–265.
- Xiaogang Wang and Xiaoou Tang. 2006a. Random sampling for subspace face recognition. *IJCV* (2006), 91–104.
- Xiaogang Wang and Xiaoou Tang. 2006b. Random Sampling for Subspace Face Recognition. *IJCV* (2006), 91–104.
- Xiaogang Wang and Xiaoou Tang. 2009. Face Photo-Sketch Synthesis and Recognition. *TPAMI* (2009), 1955–1967.
- Zhifei Wang, Zhenjiang Miao, Yanli Wan, and Zhen Tang. 2013. Kernel Coupled Cross-Regression for Low-Resolution Face Recognition. *Mathematical Problems in Engineering* (2013), 1–20.
- Paul Wright, John Corder, and Matt Glazier. 2007. Identikit. (2007).
- Bing Xiao, Xinbo Gao, Dacheng Tao, and Xuelong Li. 2009. A new approach for face recognition by sketches in photos. *Signal Processing* (2009), 1576–1588.
- Xudong Xie and Kin-Man Lam. 2006. An efficient illumination normalization method for face recognition. *Pattern Recognition Letters* 27 (2006), 609–617.
- Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and S. Lin. 2007. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *TPAMI* (2007), 40–51.
- J. Yang, J. Wright, T. S. Huang, and Y. Ma. 2010. Image Super-Resolution via Sparse Representation. *TIP* (2010), 2861–2873.
- Weilong Yang, Dong Yi, Zhen Lei, Jitao Sang, and S.Z. Li. 2008. 2D-3D face matching using CCA. In *Automatic Face Gesture Recognition*. 1–6.
- Dong Yi, Rong Liu, RuFeng Chu, Zhen Lei, and StanZ. Li. 2007. Face Matching Between Near Infrared and Visible Light Images. In *Advances in Biometrics*. 523–530.
- P.C. Yuen and C. H. Man. 2007. Human Face Image Searching System Using Sketches. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans(TSMC)* (2007), 493–504.

- Baochang Zhang, Lei Zhang, David Zhang, and Linlin Shen. 2010. Directional binary code with application to PolyU near-infrared face database. *Pattern Recognition Letters* (2010), 2337–2344.
- Di Zhang and Jiazhong He. 2010. Face super-resolution reconstruction and recognition from low-resolution image sequences. In *International Conference on Computer Engineering and Technology (ICCET)*. 620–624.
- Wei Zhang, Xiaogang Wang, and Xiaoou Tang. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*. 513–520.
- Xiaozheng Zhang and Yongsheng Gao. 2009. Face Recognition Across Pose: A Review. *PR* (2009), 2876–2896.
- Yong Zhang, C. McCullough, J.R. Sullins, and C.R. Ross. 2010. 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.
- W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. 2003. Face Recognition: A Literature Survey. *Journal ACM Computing Surveys (CSUR)* (2003), 399–458.
- Juanjuan Zhong, Xinbo Gao, and Chunna Tian. 2007. Face Sketch Synthesis using E-HMM and Selective Ensemble. In *Acoustics, Speech and Signal Processing (ICASSP)*. 485–488.
- Changtao Zhou, Zhiwei Zhang, Dong Yi, Zhen Lei, and S.Z. Li. 2011. Low-resolution face recognition via Simultaneous Discriminant Analysis. In *The International Joint Conference on Biometrics (IJCB)*. 1–6.
- Jun-Yong Zhu, Wei-Shi Zheng, and Jian-Huang Lai. 2013b. Logarithm Gradient Histogram: A general illumination invariant descriptor for face recognition. In *Automatic Face and Gesture Recognition (FG)*. 1–8.
- Jun-Yong Zhu, Wei-Shi Zheng, Jian-Huang Lai, and S.Z. Li. 2014. Matching NIR Face to VIS Face Using Transduction. *IEEE Transactions on Information Forensics and Security* (2014), 501–514.
- Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2013a. Deep Learning Identity-Preserving Face Space. In *ICCV*. 113–120.
- W.W.W. Zou and P.C. Yuen. 2012. Very Low Resolution Face Recognition Problem. *TIP* (2012), 327–340.
- Wilman W.W. Zou and Pong C. Yuen. 2010. Learning the Relationship Between High and Low Resolution Images in Kernel Space for Face Super Resolution. In *International Conference on Pattern Recognition (ICPR)*. 1152–1155.
- Xuan Zou, Kittler J, and Messer K. 2007. Illumination Invariant Face Recognition: A Survey. In *Biometrics: Theory, Applications, and Systems*. 1–8.