

# Cross-Modal Face Matching: Beyond Viewed Sketches

Shuxin Ouyang<sup>1</sup>(✉), Timothy Hospedales<sup>2</sup>, Yi-Zhe Song<sup>2</sup>, and Xueming Li<sup>1</sup>

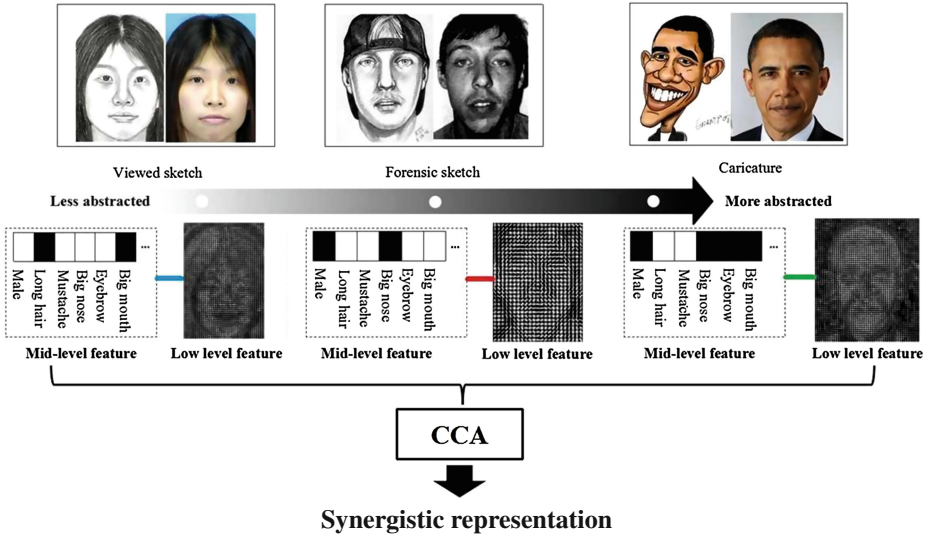
<sup>1</sup> Beijing University of Posts and Telecommunications, Beijing, China  
ouyangshuxin@gmail.com

<sup>2</sup> School of Electronic Engineering and Computer Science,  
Queen Mary University of London, London E1 4NS, UK

**Abstract.** Matching face images across different modalities is a challenging open problem for various reasons, notably feature heterogeneity, and particularly in the case of sketch recognition – abstraction, exaggeration and distortion. Existing studies have attempted to address this task by engineering invariant features, or learning a common subspace between the modalities. In this paper, we take a different approach and explore learning a mid-level representation within each domain that allows faces in each modality to be compared in a domain invariant way. In particular, we investigate sketch-photo face matching and go beyond the well-studied viewed sketches to tackle forensic sketches and caricatures where representations are often symbolic. We approach this by learning a facial attribute model independently in each domain that represents faces in terms of semantic properties. This representation is thus more invariant to heterogeneity, distortions and robust to mis-alignment. Our intermediate level attribute representation is then integrated synergistically with the original low-level features using CCA. Our framework shows impressive results on cross-modal matching tasks using forensic sketches, and even more challenging caricature sketches. Furthermore, we create a new dataset with  $\approx 59,000$  attribute annotations for evaluation and to facilitate future research.

## 1 Introduction

Cross-modal face recognition is an increasingly important research area that encompasses matching face images between different modalities: sketch, photo, infra-red, low/high resolution, 2D/3D and so on. Among all these, facial sketch based face recognition is perhaps the most important and the most well studied, due to its crucial role in assisting law enforcement. Facial sketches can be classified into three categories according to abstraction/deformation level compared to a corresponding photo: *viewed sketch*, *forensic sketch* and *caricature*, as shown in Fig. 1. Most existing studies have considered viewed sketches, which are drawn by artists while looking at a photo. This is the easiest (and most impractical) cross-modal task because the images are sufficiently similar and well aligned that extracting any grayscale descriptor from both is near sufficient



**Fig. 1.** Illustration of sketch abstraction level (top) and pipeline overview (below).

to bridge the cross-modal gap. As a result cross-modal matching rates for viewed sketch are saturated at near-perfect [1–7]. Therefore research focus has moved onto forensic sketches [1] and beyond (caricature) [8].

Contrast to viewed sketches, matching forensic sketches or caricatures to photos is significantly more challenging due to greater cross-modal gap. For forensic sketches, the witness may not exactly remember the appearance of a suspect – omitting, hallucinating or distorting individual details – or may not be able to communicate the visual memory clearly. As a result forensic sketches are often inaccurate and incomplete. In the case of caricatures, the sketch is a purposely exaggerated and distorted version of the original face. In both cases, the cross-modal gap is created by mismatch due to various factors: (i) feature heterogeneity, (ii) missing or additional facial details, (iii) distorted macro or micro proportions – which in turn affects alignment in a way that rigid registration cannot rectify. Despite these challenges, if the sketch subject is known to a human, they have no trouble identifying either forensic or caricature sketches. We are therefore motivated to study both caricature and forensic sketches, as contributions to matching caricature sketches will reflect robustness to the most challenging forensic sketch or other cross-modal recognition tasks.

In this paper, we aim to address the highlighted challenges in cross-modality matching of forensic sketches and caricatures to photos, by constructing a mid-level attribute representation of each facial modality. The idea is that this representation can be learned independently within each modality (thus completely avoiding any cross-modality challenge); but once learned, it is largely invariant to the cross-modal gap. That is, neither feature heterogeneity, nor non-linear cross-modal distortion affect this representation. Specifically, we train a bank of facial attribute detectors to produce low-dimensional semantic representation

within each modality. Finally, although the attribute representation is invariant to the cross-modal gap, it does lose some detailed information encoded by the low-level features. We therefore develop a robust synergistic representation that encodes the best of both attributes and low-level features by learning a CCA subspace that correlates the two. The result outperforms feature-based face-matching techniques, as well as state of the art cross-modal matching techniques that focus on learning a mapping between low-level features without first building an invariant mid-level representation. Moreover, a new dataset combining common forensic [9] and caricature datasets [8] were annotated ( $\approx 59,000$  annotations in total) to learn and evaluate the proposed cross-modal face representation.

The remaining parts of this paper will be organised as follows: related works will be discussed in Sect. 2; the technical methodology to bridge large cross-modal gap, that is to say, the way to matching forensic sketches and caricatures is discussed in Sect. 3; all the experiments and analysis are shown in Sect. 4; and Sect. 5 details our attribute dataset before conclusions in Sect. 6.

## 2 Related Work

### 2.1 Sketch-Based Face Recognition

As we have discussed, sketch-based face recognition can be classified based on the type of sketch used as the probe: viewed, forensic and caricature-based. In each case, strategies to bridge the cross-modal gap broadly break down into four categories: (i) those that learn a cross-modal mapping to synthesise one modality from the other, and then perform within-modality matching [10,11], (ii) those that learn a common subspace where the two modalities are more comparable [12], (iii) those that learn discriminative models to maximise matching accuracy [1,13], and (iv) those that engineer features which are simultaneously invariant to the details of each modality, while being variant to person identity [4,14].

**Viewed Sketches.** Viewed sketches are the simplest type of sketch to match against facial photos because incorrect details and distortion are minimal. This is the most extensively studied type of heterogeneous face recognition. Studies taking synthesis strategies have used eigen-transform [10] and MRF [11] optimisation to map photos into sketches before within-modal matching. Alternative studies have used PLS [12] to synthesize a common subspace where the modalities are more comparable. Meanwhile, others have engineered new invariant descriptors, including histogram of averaged oriented gradients [14] and local radon binary patterns [4]. Recognition rates on viewed sketch benchmarks has saturated, reaching 100% [14], thus research has moved on the more challenging and realistic setting of forensic sketches.

**Forensic Sketches.** One of the earliest studies to discuss automated matching forensic sketches with photos was [15]. Uhl and Lobo’s study [15] proposed a theory, and the first simple method for matching a police artist sketch to a set of photographs. It highlighted the complexity and difficulties in forensic

sketch based face recognition. One of the first major studies on forensic sketches was [1], which combined feature engineering (SIFT and LBP) with a discriminative (LFDA) method to learn a weighting that maximized identification accuracy. Later studies such as [13] improved these results, again combining feature engineering (Weber and Wavelet descriptors) plus discriminative learning (genetic algorithms) strategy to maximize matching accuracy; while [16] followed up also with feature engineering (LBP) and discriminative learning (RS-LDA).

All these strategies to bridge the cross-modal gap can largely address the feature heterogeneity problem, but the more fundamental problems of missing/additional details, and non-linear heteroskedastic distortion remain outstanding. Abstraction and distortion effects mean that any particular patch in a facial sketch image does not necessarily *correspond* to the associated patch in a facial sketch, an intrinsic problem that existing studies do not address. In this paper we avoid this issue by transforming both sketch and photo images into a mid-level semantic representation that does not depend on alignment or ability to find a patch correspondence, and is highly robust to missing/additional details.

**Caricatures.** An even more extreme cross-domain gap than in photo-forensic sketch is created by caricature-based matching. The extreme deformation and abstraction of caricatures seriously challenge all existing strategies: feature engineering methods as well as cross-domain mapping and synthesis methods are hamstrung by the impossibility of establishing patch correspondence, and mismatch of details. The main study so far addressing caricature-based matching is by Klare et al. [8]. This study proposed a semi-automated system to match caricatures to photographs based on *manually* specified facial properties for each image. However, how to *automatically* extract facial attributes is unaddressed. We address this question here, as well as how best to synergistically integrate the extracted attributes with low-level features.

## 2.2 Cross-Modal Mapping

Learning cross-modal mappings is quite widely studied, as it is of broader relevance [17, 18] than face recognition. Common approaches include using partial least squares (PLS) [12], Canonical correlation analysis (CCA) [17–19], or sparse coding [20] to map both modalities to common representation. These methods have all also been applied to cross-modal face recognition with some success. Nevertheless, in each case a fundamental assumption remains that a single linear mapping should relate the two domains. Clearly in the case of forensic sketches and caricatures with non-linear deformations, abstraction, and missing details, the assumptions of a single mapping between all sketches and all images, and that the mapping should be linear, are not met. In this paper we therefore focus on learning a semantic attribute representation, which maps low-level features to a mid-level semantic representation that is invariant to the domain gap [21] and alignment errors. Since the low-level feature to attribute transformation can be non-linear, overall this means that – unlike existing approaches – the learnable sketch-photo mapping can also be non-linear.

### 2.3 Semantic Attributes

Semantic attributes [22] have gained wide popularity as an effective representation in the broader vision community with application to object recognition [22, 23], person identification [21], and action recognition [24, 25]. However application of attributes to faces [26] or face recognition [27] has been relatively limited, and their potential for bridging the cross-modal gap is not yet explored.

Psychologists have shown that the ability of humans to perform basic-level categorization (e.g. cats vs. dogs) develops well before their ability to perform subordinate-level or fine-grained visual categorization (e.g., species), or in our case, facial attribute detection [28]. Unlike basic-level recognition, even humans have difficulty with recognition of facial attributes. This is due to attributes, such as different types of noses and eyes being quite fine grained discrimination tasks. Models and algorithms designed for basic-level object or image categorization tasks are often unprepared to catch such subtle differences among different facial attributes. In this paper, we alleviate this problem by exploiting an ensemble of classifier regions with various shapes, sizes and locations [29].

### 2.4 Our Contributions

The contributions of the paper are summarized as follows: (1) we release a dataset with  $\approx 59,000$  attribute annotations for the major caricature [8] and forensic photo-sketch datasets [1]; (2) we show how to automatically detect photo/sketch facial attributes as a modality-invariant semantic feature; (3) we show how to synergistically integrate attributes and low-level features for recognition; and (4) we demonstrate the efficacy of our approach on challenging forensic sketch and caricature sketch based recognition.

## 3 Matching Faces Across Modalities

**Problem Setting.** In the cross-modal face recognition problem, we are given a set of photo and sketch face images,  $D^p = \{\mathbf{x}_i^p\}_{i=1}^N$  and  $D^s = \{\mathbf{x}_i^s\}_{i=1}^N$  respectively. Each image is assumed to be represented by a fixed-length  $d$ -dimensional feature vector  $\mathbf{x}$ . The goal is to establish the correct correspondence between the photo set and the sketch set. Feature engineering approaches [4, 14] focus on designing the representation  $\mathbf{x}$  such that each ‘probe’ sketch  $\mathbf{x}^s$  can be matched with its corresponding photo by simple nearest neighbour as in Eq. (1), where  $|\cdot|$  indicates some distance metric such as L1, L2 [1] or  $\mathcal{X}^2$  [13, 14]. Going beyond feature engineering, studies have attempted to learn a mapping to make the modalities more comparable, such that mappings can be learned by synthesizing one modality from the other or discovering a new subspace. This typically results in NN matching in the form of Eq. (2), where the matrices  $W^s$  and/or  $W^p$  are learned, e.g., by CCA [19, 30] or PLS [12]. Alternatively, matrices  $W$  may also be learned by discriminative models [1, 8, 13] to maximize matching rate.

$$i_{NN}^* = \underset{i}{\operatorname{argmin}} |\mathbf{x}^s - \mathbf{x}_i^p| \quad (1)$$

$$i_{map}^* = \underset{i}{\operatorname{argmin}} |W^s \mathbf{x}^s - W^p \mathbf{x}_i^p| \quad (2)$$

$$i_{attr}^* = \underset{i}{\operatorname{argmin}} |\mathbf{a}^s(\mathbf{x}^s) - \mathbf{a}^p(\mathbf{x}_i^p)| \quad (3)$$

In this paper, we will go beyond existing approaches by learning a mid-level semantic attribute representation  $\mathbf{a}$  for each modality. Since the attribute representation  $\mathbf{a}$  is both (1) discriminative by design<sup>1</sup> [21, 22, 27] and (2) modality invariant, this means that NN matching as in Eq. (3) can be more powerful than previous while being robust to the modality gap approaches. In the next section we discuss how to compute a semantic attribute representation  $\mathbf{a}(\cdot)$  for photos and sketches.

### 3.1 Attribute Detection

**Training an Ensemble Classifier for Attribute Detection.** We assume an ontology of  $j = 1 \dots A$  attributes is provided (see Sect. 5 for details of the ontology). Each training image set  $D$  now also contains attribute annotation  $\mathbf{a}$  as well as images,  $D = \{\mathbf{x}_i, \mathbf{a}_i\}_{i=1}^N$ . For each modality and for each attribute  $j$  we train an ensemble classifier  $a_j(\cdot)$  as follows. Given the training data  $D$ , we randomly sample a set of  $M$  windows around the three annotated semantically relevant regions for each attribute. For all  $M$  regions, we then train a support vector machine (SVM) classifier to predict the presence of the current attribute  $j$  in the training set. The randomly sampled regions are evaluated for discriminativeness by their attribute-detection performance on a held out validation set. The top three most discriminative regions  $r = 1, 2, 3$  are then selected for each attribute.

**Detecting Attributes.** The final evaluation of the classifier ensemble for attribute  $j$  on a test image  $\mathbf{x}^*$  is  $a_j(\mathbf{x}^*) = \sum_r f_{r,j}(\mathbf{x}^*) > 0$ , where  $f_{r,j}(\cdot)$  is the binary SVM classifier for attribute  $j$  trained on region  $r$ . That is, if any classifier in the ensemble predicts the attribute is present, then it is assumed to be present. This strategy has two key advantages: (i) by selecting relevant regions for each attribute it performs feature selection to focus on relevant sub windows thus increasing detection accuracy, (ii) by exploiting an ensemble of regions it is less sensitive to alignment or deformation, typical variations of these types will trigger at least one of the classifiers in the ensemble. Given the trained classifiers for each attribute, the  $A$  dimensional attribute representation for an sketch or photo  $\mathbf{x}$  is represented by stacking them as  $\mathbf{a}(\mathbf{x}) = [a_1(\mathbf{x}), \dots, a_A(\mathbf{x})]$ .

<sup>1</sup> Attributes are chosen to be properties that differentiate groups of the population, such as male/female, asian/white, young/old – thus an  $A$ -length attribute code can potentially differentiate  $2^A$  people, providing a highly discriminative representation.

### 3.2 Learning a Synergistic Low+Mid Level Representation

The attribute representation derived in the previous section is robust and discriminative, but the original low-level features still retain some complementary information. A simple method to exploit both could be early fusion (concatenation  $[\mathbf{x}, \mathbf{a}(\mathbf{x})]$ ), or score level fusion of the similarities obtained by each representation. As a significantly better alternative, we use canonical correlation analysis (CCA) to learn an embedding space that synergistically exploits both representations.

**CCA For Representation Learning.** Specifically, assuming that we have  $N$  images in total. Let  $X_x$  be the  $N \times d$  dimensional matrix stacking the low-level feature representations  $\mathbf{x}$  for all images and  $X_a$  is a  $N \times A$  dimensional matrix stacking the attribute representations  $\mathbf{a}(\mathbf{x})$  for all images, then we find the projection matrices  $W_x$  and  $W_a$  such that:

$$\begin{aligned} \min_{W_x, W_a} & \|X_a W_a - X_x W_x\|_F^2 \\ \text{subject to } & W_a^T \Sigma_{ax} W_x = I, w_{ak}^T \Sigma_{ax} w_{xl} = 0, \\ & k, l = 1, \dots, c \end{aligned} \quad (4)$$

where  $\Sigma_{ax}$  is the covariance between  $X_a$  and  $X_x$  and  $w_{ak}$  is the  $k$ th column of  $W_a$ , and  $c$  is the dimensionality of the desired CCA subspace. To solve this optimization, we use the efficient generalized eigenvalue method of [18].

Note that this is a somewhat different use of CCA to some previous studies that used it to map across facial image domains [19, 20, 30]. Instead we use CCA to construct an embedding [18] to constructively fuse attribute and low-level feature representations.

**Using the Representation.** In order to obtain the semantic embedding for a test image  $\mathbf{x}^*$ , we first obtain its estimated attributes  $\mathbf{a}(\mathbf{x})$ . Then we project both the original and semantic views of the image into the embedding space:  $\mathbf{x}W_x$  and  $\mathbf{a}(\mathbf{x})W_a$ . Finally, we concatenate both views to give the final  $2c$  dimensional representation:  $R(\mathbf{x}) = [\mathbf{x}W_x, \mathbf{a}(\mathbf{x})W_a]$ . Once our new robust and domain invariant representation is obtained for sketch and photo images, matching a sketch  $\mathbf{x}^s$  against a photo dataset  $D = \{\mathbf{x}_i^p\}_i^N = 1$  is then performed by nearest neighbor with L2 distance,

$$i^* = \underset{i}{\operatorname{argmin}} |R^s(\mathbf{x}^s) - R^p(\mathbf{x}_i^p)| \quad (5)$$

Note that the representation  $R$  in Eq. (5) is indexed by (s)ketch or (p)hoto because the semantic attribute model  $\mathbf{a}(\cdot)$  is independently trained for each modality, although the CCA mapping is shared.

## 4 Experimental Results

### 4.1 Datasets and Settings

**Datasets:** We evaluate our algorithm on two challenging datasets for photo-sketch face matching forensic [9] and caricature dataset [8]. For the forensic

dataset we have 196 pairs of 200\*160 pixel resolution face and photo images. For the caricature dataset, we have 207 pairs of caricatures and photographs of highly variable size. To obtain matching results, we perform 3-fold cross-validation, splitting the data into 2/3s training, and test on the held out 1/3. Within the training set we use 4-fold cross-validation to both train the attribute representation (Sect. 3.1) and optimize dimensionality ( $c = 250$  for both datasets) of the CCA subspace (Sect. 3.2).

**Low-Level Features:** For low-level feature representation, we densely extract histogram of gradients (HOG) and local binary patterns (LBP) on each image on a  $16 \times 16$  grid, resulting in 4030 and 6960 dimensional descriptors respectively. We then use PCA to reduce the dimension of each to 350 and concatenate the result, producing a  $d = 700$  dimensional descriptor for each image.

**Training Attribute Detectors:** Using the 73 attribute methodology defined in Sect. 5, and the training procedure in Sect. 3.1, we produce a 71 dimensional binary attribute vector for each image in the caricature dataset, and a 53 dimensional binary attribute vector for each image in the forensic dataset<sup>2</sup>.

**Baselines:** We compare our method with the following four variants of our method: (i) use only HOG and LBP (**LLF**); (ii) use only the attribute representation in nearest neighbor matching (**Attribute**); (iii) use low-level features and attributes together with simple early (feature) level fusion (**Attribute+LLF**); (iv) our full method, using low-level features and attributes together through synergistic CCA space (Cross-modal Matching by Facial Attributes, **CMMFA**).

Additionally, we compare the following two previous state of the art approaches: (i) low-level features engineered for photo-caricature recognition followed by NN matching [8] (**Klare**); (ii) state of the art learned cross-modal mapping, learned based on our LLFs, followed by NN matching [19] (**CFS**).

**Table 1.** Attribute recognition results for caricature and forensic datasets, comparing our ensemble attribute classifier with flat-model ones (acc. is for average accuracy).

Dataset	Classifier	Acc. (sketch)	Acc. (photo)
Caricature	Flat-model	53.95 %	55.15 %
Forensic	Flat-model	56.23 %	54.43 %
<i><b>Caricature</b></i>	Ensemble	69.15 %	70.24 %
<i><b>Forensic</b></i>	Ensemble	65.19 %	65.28 %

<sup>2</sup> Both datasets (especially forensic) exhibits some degrees of lower diversity of attributes, so some attributes are always on or off rendering them meaningless for representation, so these are excluded for convenience.



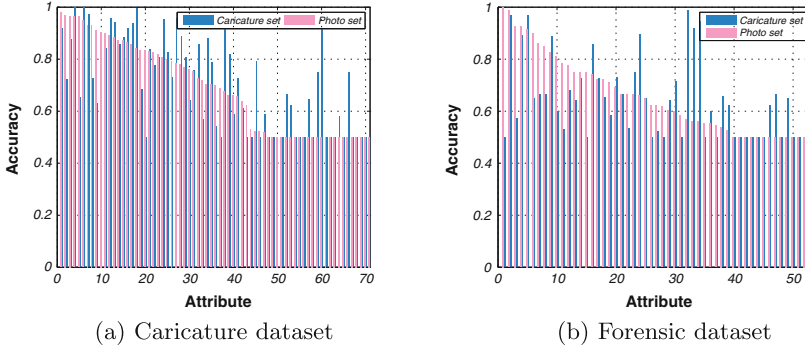


Fig. 2. Breakdown of per-attribute detection performance.

## 4.2 Attribute Detection



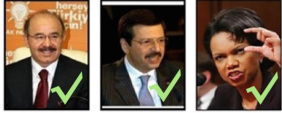


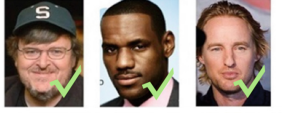




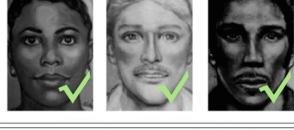


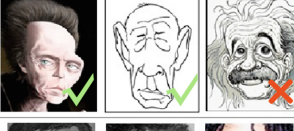




In this section, we first evaluate the performance of our automated attribute detection procedure (Sect. 3.1). In Table 1, we offer the average attribute detection accuracy for each of our datasets, and performance comparison between the proposed ensemble attribute detector and a flat-model variant where SVM classifiers are trained on whole images instead. Although many attributes are quite subtle (Sect. 5), the average accuracies in the range 65–70% clearly demonstrate that many of them can be reasonably reliably detected, especially when compared with flat-model performance (53–56%). Table 2 reports the top 5 most accurate attributes for each modality and dataset. The top 4 rows of Fig. 3 illustrate attribute detection results for the 1st ranked attributes (shown schematically) in each dataset/modality, the bottom 2 rows show failure examples of attribute detection (denoted by red cross), i.e., when automatic attribute detection disagrees with human annotated ground-truth.

To further investigate how these averages break down, we plot the per-attribute accuracy in Fig. 2 sorted by photo set accuracy. Clearly while there are some attributes which are too subtle to be reliably detected (some attributes at 50% accuracy, e.g., slanted and sleepy eyes), others can be detected with near perfect accuracy (plots peak at around 100% accuracy). Interestingly, while there is a general correlation of attribute reliability between datasets, it is relatively weak, so some of the best photo attributes don't work on sketch and vice-versa.

## 4.3 Matching Across Modalities

Given the attribute encoding as evaluated in the previous section, we next address the final goal of cross-modal face matching.

**Caricature Dataset.** The results for cross-modal face matching on the caricature dataset are shown in Fig. 4(a) and Table 3. For the caricature dataset our attribute encoding is significantly better than any of the LLF based approaches (Table 3, *Attribute versus HOG/LBP*). This because the cross-modal gap for the

Attribute	Caricature and forensic sketches	Photographs
 Teeth		
 Mustache		
 Cheek		
 Face(3)		
 Nose to mouth		
 Eyebrows(1)		

**Fig. 3.** Illustration of detections for the best performing attributes in Table 2 (top 4 rows) and 2 other average performing attributes (bottom 2 rows), Left: Schematic illustration of query attribute, middle and right: pairs of sketch/caricature (middle) and photograph (right) of the same identity (green tick for successful attribute detection, red cross otherwise) (Colour figure online).

caricature dataset is the most extreme, so low-level features cannot be effectively compared. CFS improves the results somewhat compared to LLFs, but due to the extreme gap between the domains, it offers limited improvement (Table 3, *CFS versus HOG/LBP*). For context, we also show the matching results obtained using the ground-truth attributes in the bottom row. Interestingly this is only a few percent above that obtained by using our inferred attributes, suggesting that we are already capturing most of the value in the current attribute methodology (Table 3, *Ground-truth attribute versus Attribute*).

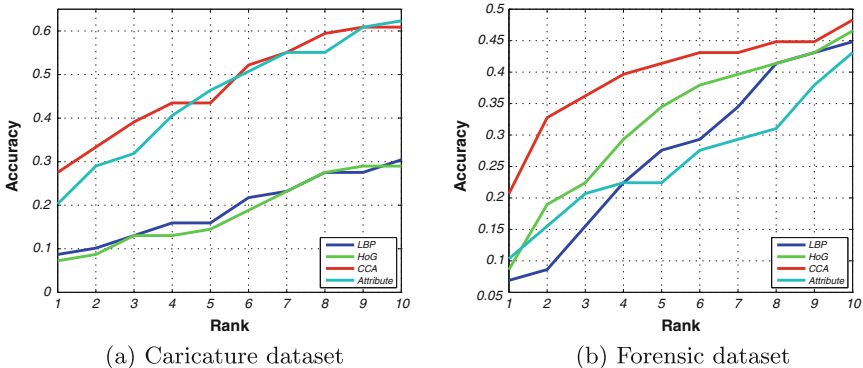
With regards to strategies for combining attributes and LLFs, vanilla concatenation actually worsens the results somewhat compared to attributes alone (Table 3, *Attribute versus Attribute+LLF*). This is understandable, because the

**Table 2.** Top 5 attributes for caricature and forensic datasets

Caricature dataset			Forensic dataset		
Domain	Attribute	Accuracy	Domain	Attribute	Accuracy
Photo	Teeth	97.98 %	Photo	Cheeks(5)	100.00 %
	Cheeks(3)	96.96 %		No beard	98.68 %
	Glasses	96.43 %		Small forehead	92.86 %
	Small beard	94.87 %		Eyebrow(2)	92.50 %
	Big chin	93.12 %		No moustache	91.67 %
Caricature	Small mustache	100.00 %	Forensic	Face(3)	98.84 %
	Forehead(1)	100.00 %		No beard	96.97 %
	Square face	99.45 %		No moustache	96.67 %
	Big moustache	97.46 %		Thick eyebrow	94.57 %
	Big mouth	95.83 %		Small mouth	91.94 %

attributes are much stronger than the LLFs. In contrast, combining them in our CCA framework achieves the best result of all, 27.54 % at Rank 1. Finally, we compare with the results based on engineered image features reported in [8]. The features from [8] slightly outperform our LLFs alone. However our entire framework outperforms [8] by a noticeable margin.

We note that using everything together, [8]’s final result only slightly outperforms our CMMFA. However, this is using the critical but unrealistic cue of manually annotated ground-truth attributes at test time, which makes this approach not meaningful for a practical scenario that should be fully automated (Table 3, *Klare versus CMMFA*). In contrast, our CMMFA is computed based on image features alone without manual intervention.

**Fig. 4.** CMC curves for cross-modality matching. Ranks = 1:10.

**Table 3.** Caricature dataset: comparison of all methods.

Methods	Rank1	Rank5
Dense HOG	7.25 %	14.49 %
Dense LBP	8.60 %	15.94 %
CFS [19] HOG+LBP	13.45 %	
Attribute	20.29 %	46.38 %
Attribute+LLF	18.84 %	46.38 %
<b><i>CMMFA (Attribute+LLF+CCA)</i></b>	27.54 %	43.48 %
Klare et al. (image only) [8]	12.10 %	52.10 %
Klare et al. (method fusion and manual attributes) [8] <sup>a</sup>	32.30 %	–
Ground-truth attribute	23.19 %	52.17 %

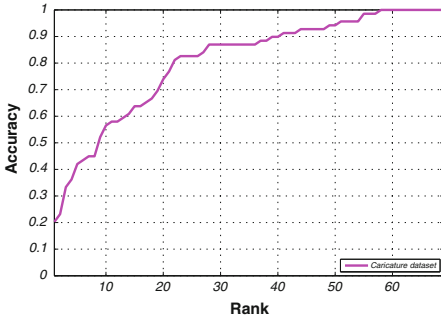
<sup>a</sup> Not directly comparable due to use of manual intervention.

**Forensic Dataset.** The results for the forensic dataset are shown in Fig. 4(b) and Table 4. In this case our attribute encoding still outperforms LLF based approaches (Table 4, *Attribute versus HOG/LBP*), despite the fewer and weaker attributes in this case. CFS now improves the LLF results more significantly as expected since the cross-modal gap is more straightforward to model (Table 4, *CFS versus HOG/LBP*). However our full method still outperforms CFS (Table 4, *CMMFA versus CFS*).

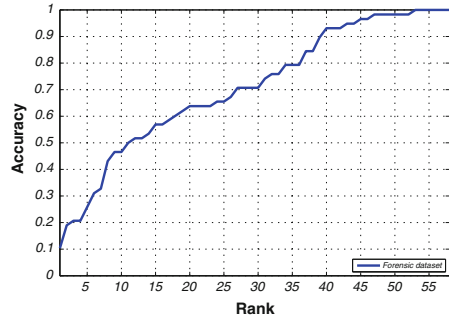
Our CMMFA performance is slightly weaker on the forensic than the caricature dataset. This is somewhat surprising, because the caricature dataset might be considered ‘harder’ due to the bigger cross-modal gap. However, it is understandable because there are about 20 facial attributes which do not occur (or occur infrequently) in the forensic set, thus resulting in fewer working attribute detectors (Fig. 2); and because the process of caricature sketching often actually involves exaggerating facial attributes, thus making them easier to detect.

#### 4.4 Attribute Description Search

As a final application of potential relevance to forensic search, we consider querying a mugshot database solely by attribute description. This is interesting and potentially useful for law enforcement, especially in situations where a trained forensic sketch expert is unavailable. In this application scenario, a witness would select all the attributes they recall from the full attribute list. A mugshot-database can then be queried directly by the attribute representation. We simulate this experiment by querying each person  $j$ ’s ground-truth attribute  $\mathbf{a}_j^{gt}$  against the database of estimated attributes for the mugshots  $\{\mathbf{a}(\mathbf{x}_i^p)\}_{i=1}^N$ ,  $i^* = \arg\min_i |\mathbf{a}_j^{gt} - \mathbf{a}(\mathbf{x}_i^p)|$ . With this setting, we achieve average of 10.3 % rank 1 accuracy for the forensic dataset, and 20.7 % rank 1 accuracy for the caricature dataset. Full CMC curves are shown in Fig. 5.



(a) Caricature dataset



(b) Forensic dataset

**Fig. 5.** CMC curves for attribute description search.**Table 4.** Forensic dataset: comparison of all methods.

Methods	Rank1	Rank5
Dense HOG	8.60 %	34.48 %
Dense LBP	6.90 %	27.59 %
CFS [19] HOG+LBP	19.12 %	
Attribute	10.34 %	22.41 %
Attribute+LLF	18.97 %	36.21 %
<b><i>CMMFA (Attribute+LLF+CCA)</i></b>	20.69 %	41.38 %
Ground-truth attribute	15.52 %	44.83 %

## 5 Attribute Dataset

In this section we describe the dataset that was created in this study. Future studies comparing accuracies on this dataset should follow the protocol detailed in Sect. 4. We build our attribute dataset by annotating the caricature dataset<sup>3</sup> [8] and forensic dataset [9].

**Caricature Dataset:** The dataset consists of pairs of a caricature sketch and a corresponding facial photograph from 207 subjects. Two sources were used to collect these images. The first was through contacts with various artists who drew the caricatures. And the second source of caricature images was from Google Image searches. When selecting face photographs, care was taken to find images that had minimal variations in pose, illumination and expression, however, those images are hard to find. So, many of the factors still persist [8].

**Forensic Dataset:** The dataset consists of pairs of a forensic sketch and a corresponding mugshot from 196 subjects. Forensic sketches are drawn by a

<sup>3</sup> Reference [8] did not release their attributes. Our attributes and corresponding annotations are available at <http://www.eecs.qmul.ac.uk/~yzs/heteroface/>.

sketch artist from the description of an eyewitness based on his/her recollection of the crime scene. Two sources were used to collect these images. The first was through contact with various artists who drew the forensic sketches: Lois Gibson and Karen Taylor. The second was from Google Image searches [9].

**Attribute Annotation.** In our attribute dataset, each of the images (caricature, forensic sketch and photograph) is labeled with a set of facial attributes (categorical facial features). We start with the 63 facial attributes proposed by Klare et al [8], and add 10 additional attributes for a total of 73 attributes. Those 10 additional attributes include: wrinkles, glasses, ties, teeth, cheeks, black/while/asian, blonde hair and gender.

Each image (caricature, forensic sketch and photograph) was annotated for these 73 attributes. Each annotator labeled the entire set of image pairs with 3–4 facial attributes, being asked to label a single image with a single attribute at a time. Thus the annotator was shown an image of a caricature, a forensic sketch or a photograph, and the current attribute being labeled. If the attribute is present, then they label the image with ‘1’, otherwise ‘0’. In total, we provided 58,838 labels on the 806 images. For each attribute (not each image), annotators are also asked to provide an estimate of three salient regions for that attribute, which were used to guide random sampling for attribute detection (Sect. 3.1).

## 6 Summary

In this work, going beyond viewed sketches, we address the challenging task of cross-modal face recognition for forensic and caricature sketches. To deal with the cross-modal gap due to heterogeneity, abstraction and distortion, we constructed an intermediate level attribute representation within each modality. To address the challenge of automated attribute detection within each image we introduce an ensemble of attribute detectors. Crucially, our semantic attribute representation is invariant to the details of the modality, and thus can be more directly compared across modalities than pixels or low-level features. Finally, we created a synergistic representation to integrate the semantic and low-level feature representations by learning an embedding subspace using CCA. As a result we are able to outperform several state of the art cross-domain mapping methods for both challenging datasets. We believe this is the first use of fully automated facial attribute analysis to improve cross-modal recognition.

Promising avenues for future research include integrating features at an abstraction level between pixels and attributes (e.g., facial interest points) along with our current framework of attributes and low-level image features. We also plan to investigate reasoning about attribute correlation; and extending our framework to apply to other modalities such as infra-red as well as sketch.

## References

1. Klare, B., Li, Z., Jain, A.: Matching forensic sketches to mug shot photos. In: TPAMI, pp. 639–646 (2011)

2. Khan, Z., Hu, Y., Mian, A.: Facial self similarity for sketch to photo matching. In: Digital Image Computing Techniques and Applications (DICTA), pp. 1–7 (2012)
3. Kiani Galoogahi, H., Sim, T.: Face photo retrieval by sketch example. In: the 20th ACM International Conference on Multimedia, pp. 949–952 (2012)
4. Galoogahi, H., Sim, T.: Face sketch recognition by local radon binary pattern lrbp. In: ICIP, pp. 1837–1840 (2012)
5. Pramanik, S., Bhattacharjee, D.: Geometric feature based face-sketch recognition. In: Pattern Recognition, Informatics and Medical Engineering (PRIME), pp. 409–415 (2012)
6. Bhatt, H.S., Bharadwaj, S., Singh, R., Vatsa, M.: On matching sketches with digital face images. In: Biometrics: Theory Applications and Systems, pp. 1–7 (2010)
7. Choi, J., Sharma, A., Jacobs, D., Davis, L.: Data insufficiency in sketch versus photo face recognition. In: CVPR, pp. 1–8 (2012)
8. Klare, B., Bucak, S., Jain, A., Akgul, T.: Towards automated caricature recognition. In: The 5th IAPR International Conference on Biometrics Compendium, pp. 139–146 (2012)
9. Bhatt, H.S., Bharadwaj, S., Singh, R., Vatsa, M.: Memetic approach for matching sketches with digital face images. Indraprastha Institute of Information Technology Delhi, pp. 1–8 (2012)
10. Tang, X., Wang, X.: Face photo recognition using sketch. In: ICIP, pp. 257–260 (2002)
11. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. TPAMI **31**, 1955–1967 (2009)
12. Sharma, A., Jacobs, D.W.: Bypassing synthesis PLS for face recognition with pose, low-resolution and sketch. In: CVPR, pp. 593–600 (2011)
13. Bhatt, H., Bharadwaj, S., Singh, R., Vatsa, M.: Memetically optimized mcwld for matching sketches with digital face images. IEEE Trans. Inf. Forensics Secur. **7**, 1522–1535 (2012)
14. Galoogahi, H., Sim, T.: Inter-modality face sketch recognition. In: ICME (2012)
15. Uhl, R.G., Jr., da Vitoria Lobo, N.: A framework for recognizing a facial image from a police sketch. In: CVPR, pp. 586–593 (1996)
16. Bonnen, K., Klare, B., Jain, A.: Component-based representation in automated face recognition. IEEE Trans. Inf. Forensics Secur. **8**, 239–253 (2013)
17. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the international conference on Multimedia, pp. 251–260 (2010)
18. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV **106**, 210–233 (2014)
19. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: ICCV, pp. 2088–2095 (2013)
20. Huang, D.A., Wang, Y.C.F.: Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In: ICCV (2013)
21. Layne, R., Hospedales, T.M., Gong, S.: Person re-identification by attributes. In: BMVC, pp. 1–11 (2012)
22. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR, pp. 951–958 (2009)
23. Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z., Gong, S.: Transductive multi-view embedding for zero-shot recognition and annotation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 584–599. Springer, Heidelberg (2014)

24. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR, pp. 3337–3344 (2011)
25. Fu, Y., Hospedales, T., Xiang, T., Gong, S.: Learning multimodal latent attributes. TPAMI **36**, 303–316 (2014)
26. Luo, P., Wang, X., Tang, X.: A deep sum-product architecture for robust facial attributes analysis. In: ICCV, pp. 2864–2871 (2013)
27. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV, pp. 365–372 (2009)
28. Johnson, K.E.: Effects of knowledge and development on subordinate level categorization. *Cognitive Dev.* **13**, 515–545 (1998)
29. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR, pp. 1577–1584 (2011)
30. Yi, D., Liu, R., Chu, R., Lei, Z., Li, S.Z.: Face matching between near infrared and visible light images. In: Lee, S.-W., Li, S.Z. (eds.) *Advances in Biometrics*. LNCS, vol. 4642. Springer, Heidelberg (2007)