Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition

Wei Zhang[1]
[1]Department of Information Engineering
The Chinese University of Hong Kong
zw009@ie.cuhk.edu.hk

Xiaogang Wang[2,3]
[2]Department of Electronic Engineering
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

Xiaoou Tang[1,3]
[3]Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, China
xtang@ie.cuhk.edu.hk

Abstract

Automatic face photo-sketch recognition has important applications for law enforcement. Recent research has focused on transforming photos and sketches into the same modality for matching or developing advanced classification algorithms to reduce the modality gap between features extracted from photos and sketches. In this paper, we propose a new inter-modality face recognition approach by reducing the modality gap at the feature extraction stage. A new face descriptor based on coupled information-theoretic encoding is used to capture discriminative local face structures and to effectively match photos and sketches. Guided by maximizing the mutual information between photos and sketches in the quantized feature spaces, the coupled encoding is achieved by the proposed coupled information-theoretic projection tree, which is extended to the randomized forest to further boost the performance. We create the largest face sketch database including sketches of $1,194$ *people from the FERET database. Experiments on this large scale dataset show that our approach significantly outperforms the state-of-the-art methods.*

1. Introduction

Face photo-sketch recognition is to match a face sketch drawn by an artist to one of many face photos in the database. In law enforcement, it is desired to automatically search photos from police mug-shot databases using a sketch drawing when the photo of a suspect is not available. This application leads to a number of studies on this topic [26, 27, 28, 31, 9, 14, 6]. Photo-sketch generation and recognition are also useful in digital entertainment industry.

The major challenge of face photo-sketch recognition is to match images in different modalities. Sketches are a concise representation of human faces, often containing shape exaggeration and having different textures than photos. It is infeasible to directly apply face photo recognition algorithms. Recently, great progress has been made in two directions. The first family of approaches [27, 18, 31] focused on the *preprocessing* stage and synthesized a pseudo-photo from the query sketch or pseudo-sketches from the gallery photos to transform inter-modality face recognition into intra-modality face recognition. Face photo/sketch synthesis is actually a harder problem than recognition. Imperfect synthesis results significantly degrade the recognition performance. The second family of approaches [17, 15, 14] focused on the *classification* stage and tried to design advanced classifiers to reduce the modality gap between features extracted from photos and sketches. If the inter-modality difference between the extracted features is large, the discriminative power of the classifiers will be reduced.

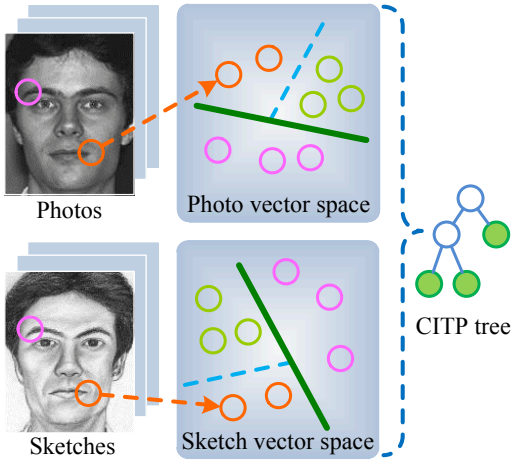


Figure 1. A CITP tree with three levels for illustration purpose. The local structures of photos and sketches are sampled and coupled encoded via the CITP tree. Each leaf node of the CITP tree corresponds to a cell in the photo vector space and in the sketch vector space. The sampled vectors in the same cell are assigned the same code, so that different local structures have different codes and the same structures in different modalities have the same code.

In this paper, we propose a new approach of reducing the modality gap at the *feature extraction* stage. A new face descriptor is designed by the coupled information-theoretic encoding, which quantizes the local structures of face photos and sketches into discrete codes. In order to effectively match photos and sketches, it requires that the extracted

codes are uniformly distributed across different subjects, which leads to high discriminative power, and that the codes of the same subject's photo and sketch are highly correlated, which leads to small inter-modality gap. These requirements can be well captured under the criterion of maximizing the mutual information between photos and sketches in the quantized feature spaces. The coupled encoding is achieved by the proposed randomized coupled information-theoretic projection forest, which is learned with the *maximum mutual information* (MMI) criterion.

Another contribution of this work is to release CUHK Face Sketch FERET Database (CUFSF)[1], a large scale face sketch database. It includes the sketches of $1,194$ people from the FERET database [22]. Wang and Tang [31] published the CUFS database with sketches of 606 people. The sketches in the CUFS database had less shape distortion. The new database is not only larger in size but also more challenging because its sketches have more shape exaggeration and thus are closer to practical applications. Experiments on this large scale dataset show that our approach significantly outperforms the state-of-the-art methods.

1.1. Related work

To synthesize pseudo photos (sketches) from sketches (photos), Tang and Wang [27] proposed to apply the eigen-transform globally. Another global approach proposed by Gao et al. [9] was based on the embedded hidden Markov model and the selective ensemble strategy. Liu *et al*. [18] proposed patch-based face sketch reconstruction using local linear embedding based mapping. The sketch patches were synthesized independently ignoring the spatial relationship. Wang and Tang [31] used a multiscale Markov random field (MRF) to model the dependency of neighboring sketch patches. Photos and sketches were matched once they were transformed to the same modality.

In order to reduce the inter-modality gap at the classification stage, Lin and Tang [17] mapped features from two modalities into a common discriminative space. Lei and Li [15] proposed coupled spectral regression (CSR). CSR was computationally efficient in learning projections to map data from two modalities into a common subspace. Klare *et al*. [14] proposed local feature-based discriminant analysis (LFDA). They used multiple projections to extract a discriminative representation from partitioned vectors of SIFT and LBP features.

There is an extensive literature on descriptor-based face recognition [1, 32, 36], due to its advantages of computational efficiency and relative robustness to illumination and pose variations. They are relevant to our coupled encoding. However, those handcrafted features, such as LBP [1] and SIFT [19], were not designed for inter-modality face recognition. The extracted features from photos and sketches may have large inter-modality variations.

Although information-theoretic concepts were explored in building decision trees and decision forests for vector quantization [2, 21, 23] in the application of object recognition, these algorithms were applied in a single space and did not address the problem of inter-modality matching. With the supervision of object labels, their tree construction processes were much more straightforward than ours.

2. Information-Theoretic Projection Tree

Vector quantization was widely used to create discrete image representations, such as textons [20] and visual words [24], for object recognition and face recognition. Image pixels [5, 23], filter-bank responses [20] or invariant descriptors [24, 33] were computed either sparsely or densely on a training set, and clustered to produce a codebook by algorithms such as k-means, mean shift [12], random projection tree [5, 8, 33] and random forest [21, 23]. Then with the codebook any image could be turned into an encoded representation.

However, to the best of our knowledge, it has not been clear how to apply vector quantization to cross-modality object matching yet. In this section, we present a new coupled information-theoretic projection (CITP) tree for coupled quantization across modalities. We further extend the CITP tree to the randomized CITP tree and forest. For clarity of exposition, we present the method in the photo-sketch recognition scenario.

2.1. Projection Tree

A projection tree [8] partitions a feature space $\mathbb{R}^D$ into cells. It is built in a recursive manner, splitting the data along one projection direction at a time. The succession of splits leads to a binary tree, whose leaves are individual cells in $\mathbb{R}^D$. With a built projection tree, a code is assigned to each test sample $\mathbf{x}$, according to the cell (i.e. leaf node) it belongs to. The sample is simply propagated down the tree, starting from the root node and branching left or right until a leaf node is reached. Each node is associated with a learned binary function $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} - \tau)$. The node propagates $\mathbf{x}$ to its left child if $f(\mathbf{x}) = -1$ and to its right child if $f(\mathbf{x}) = 1$.

2.2. Mutual Information Maximization

Since quantization needs to be done in both the photo space and the sketch space, we extend a projection tree to a coupled projection tree. In a coupled projection tree, vectors sampled from photos and sketches share the same tree structure, but are input to different binary functions $f_p(\mathbf{x}_p)$ and $f_s(\mathbf{x}_s)$ at each node. A vector $\mathbf{x}_p$ sampled from the neighborhood of a photo pixel is quantized with f_p and a vector $\mathbf{x}_s$ sampled from the neighborhood of a sketch pixel is quantized with f_s. Then the sampled photo vectors and

[1] Available at `http://mmlab.ie.cuhk.edu.hk/cufsf/`.

sketch vectors are mapped to the same codebook, but their coding functions represented by the tree are different, denoted by $\mathcal{C}_p$ and $\mathcal{C}_s$, respectively.

To train a coupled projection tree, a set of vector pairs $\mathcal{X} = \{(\mathbf{x}_i^p, \mathbf{x}_i^s), i = 1, ..., N\}$ is prepared, where $\mathbf{x}_i^p, \mathbf{x}_i^s \in \mathbb{R}^D$. In this paper, $\mathbf{x}_i^p$ and $\mathbf{x}_i^s$ are the normalized vectors of sampled gradients around the same location[2] in a photo and a sketch of the same subject, respectively. Denote that $\mathbf{X}^p = [\mathbf{x}_1^p, ..., \mathbf{x}_N^p]$, $\mathbf{X}^s = [\mathbf{x}_1^s, ..., \mathbf{x}_N^s]$. Since $\mathbf{x}_i^p$ and $\mathbf{x}_i^s$ are sampled from the same subject at the same location, it is expected that they are quantized into the same code by the coupled projection tree. In the meanwhile, in order to increase the discriminative power, it is expected that the codes of $\mathbf{X}^p$ and $\mathbf{X}^s$ are uniformly distributed across different subjects. To achieve these goals, our coupled information-theoretic projection (CITP) trees are learned using the *maximum mutual information* (MMI) criterion (see Fig. 2).

Mutual information, which is a symmetric measure to quantify the statistical information shared between two random variables [7], provides a sound indication of the matching quality between coded photo vectors and coded sketch vectors. Formally, the objective function is as follows.[3]

$$I(\mathcal{C}_p(\mathbf{X}^p); \mathcal{C}_s(\mathbf{X}^s)) = H(\mathcal{C}_p(\mathbf{X}^p)) - H(\mathcal{C}_s(\mathbf{X}^p)|\mathcal{C}_p(\mathbf{X}^s)). \quad (1)$$

To increase the discriminative power, the quantization should maximize the entropy $H(\mathcal{C}_p(\mathbf{X}^p))$ so that the samples are nearly uniformly distributed over the codebook. To reduce the inter-modality gap, the quantization should minimize the conditional entropy $H(\mathcal{C}_p(\mathbf{X}^p)|\mathcal{C}_s(\mathbf{X}^s))$.

2.3. Tree Construction with MMI

Similar to random projection tree [8], the CITP tree is also built top down recursively. However, it is different in that the CITP tree is not a balanced binary tree, i.e. the leaf nodes are at different levels. So the tree building process consists of searching for both the best tree structure and the optimal parameters at each node.

Tree structure searching. We adopt a greedy algorithm to build the tree structure. At each iteration, we search the node whose splitting can maximize the mutual information between the codes of sampled photo and sketch vectors. The mutual information, given in Eqn. (1), can be easily approximated in a nonparametric way. All the sampled photo and sketch vectors in the training set are quantized into codes with the current tree after splitting the candidate node, and the joint distribution of photo and sketch codes is computed to estimate the mutual information. A toy example is shown in Fig. 2.

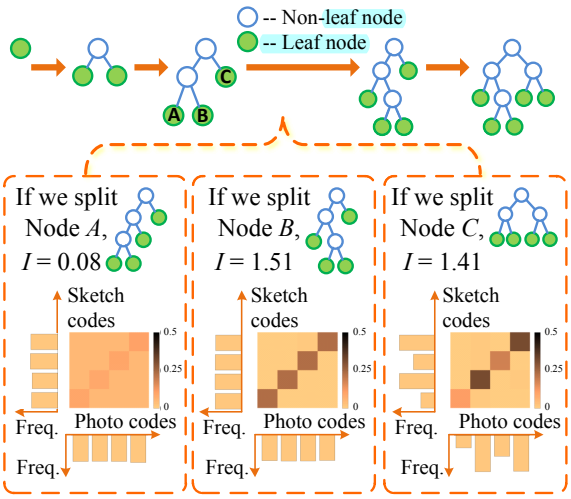


Figure 2. An illustration of tree construction with MMI. In each step, all current leaf nodes are tested and the one with the maximum mutual information is selected to split. For a leaf node, we try to split it and obtain a tree to encode photo vectors and sketch vectors. The selected leaf node should satisfy: (1) the codes are uniformly distributed; (2) the codes of photo vectors and corresponding sketch vectors are highly correlated. These requirements can be well captured under the MMI criterion. In this example, if we split node A, requirement (2) will not be satisfied, and if we split node C, requirement (1) will not be satisfied. The corresponding mutual information I of both are relatively small. So node B with the maximum mutual information is selected. The histograms and joint histograms of photo and sketch codes are visualized. In joint histograms, the colors represent the joint probability densities.

Node parameter searching. It is critical to search for optimal parameters of binary functions $f_p(\mathbf{x}_p)$ and $f_s(\mathbf{x}_s)$ to determine how to split the node. Formally, we aim at finding projection vectors $\mathbf{w}_p, \mathbf{w}_s$ and thresholds τ_p, τ_s for node k[4], such that

$$\begin{aligned} y_i^p &= \mathbf{w}_p^T \mathbf{x}_i^p - \tau_p, \quad \hat{y}_i^p = \text{sign}(y_i^p), \\ y_i^s &= \mathbf{w}_s^T \mathbf{x}_i^s - \tau_s, \quad \hat{y}_i^s = \text{sign}(y_i^s). \end{aligned} \quad (2)$$

Then a binary value $\hat{y}_i^p$ (or $\hat{y}_i^s$) is assigned to each vector $\mathbf{x}_i^p$ (or $\mathbf{x}_i^s$), to split the training data into two subsets and propagate them to the two child nodes. The node propagates a training vector pair $(\mathbf{x}_i^p, \mathbf{x}_i^s)$ to its children only if the binary values $\hat{y}_i^p$ and $\hat{y}_i^s$ are the same. Otherwise, the vector pair is treated as an outlier and discarded.

Suppose that the input of a node k is a set of vector pairs $\mathcal{X}_k = \{(\mathbf{x}_{k_i}^p, \mathbf{x}_{k_i}^s), 1 \le i \le N_k\}$. Denote that $\mathbf{X}_k^p = [\mathbf{x}_{k_1}^p, ..., \mathbf{x}_{k_{N_k}}^p]$, $\mathbf{X}_k^s = [\mathbf{x}_{k_1}^s, ..., \mathbf{x}_{k_{N_k}}^s]$, $\mathbf{Y}_k^p = [y_{k_1}^p, ..., y_{k_{N_k}}^p]$, $\mathbf{Y}_k^s = [y_{k_1}^s, ..., y_{k_{N_k}}^s]$, $\hat{\mathbf{Y}}_k^p = [\hat{y}_{k_1}^p, ..., \hat{y}_{k_{N_k}}^p]$

[2] We sample the gradients (i.e. the first-order derivatives in the horizontal and vertical directions) I_u and I_v for an image I. Please refer to Section 3 for details.

[3] The mutual information is originally defined between two random variables $\mathcal{C}_p(\mathbf{x}_i^p)$ and $\mathcal{C}_s(\mathbf{x}_i^s)$. We use the empirical mutual information estimated on the training set throughout this paper.

[4] We omit index k of the parameters, for conciseness.

and $\widehat{\mathbf{Y}}_k^s = [\widehat{y}_{k_1}^s, ..., \widehat{y}_{k_{N_k}}^s]$. The node is split according to the MMI criterion, i.e. maximizing

$$I(\widehat{\mathbf{Y}}_k^p; \widehat{\mathbf{Y}}_k^s) = H(\widehat{\mathbf{Y}}_k^p) + H(\widehat{\mathbf{Y}}_k^s) - H(\widehat{\mathbf{Y}}_k^p, \widehat{\mathbf{Y}}_k^s). \quad (3)$$

Instead of solving the above maximization problem directly, an approximate objective $I(\mathbf{Y}_k^p; \mathbf{Y}_k^s)$ is maximized first. Through maximizing $I(\mathbf{Y}_k^p; \mathbf{Y}_k^s)$, $\mathbf{w}_p$ and $\mathbf{w}_s$ are estimated without considering τ_p and τ_s. Assume that $y_{k_i}^p$ and $y_{k_i}^s$ are jointly Gaussian distributed. The entropy of a jointly Gaussian random vector $\mathbf{g}$ is $\frac{1}{2}\ln[\det(\mathbf{\Sigma_g})] + \text{const}$ [7], where $\mathbf{\Sigma_g}$ is the covariance matrix of $\mathbf{g}$. Following this, the mutual information can be rewritten in a simple form

$$I(\mathbf{Y}_k^p; \mathbf{Y}_k^s) = \frac{1}{2}\ln\left(\frac{\det(\Sigma_k^p)\det(\Sigma_k^s)}{\det(\mathbf{\Sigma}_k)}\right) + \text{const}, \quad (4)$$

where Σ_k^p, Σ_k^s and $\mathbf{\Sigma}_k$ are the covariance of $\mathbf{Y}_k^p$, $\mathbf{Y}_k^s$ and $[(\mathbf{Y}_k^p)^T, (\mathbf{Y}_k^s)^T]^T$, respectively. According to Eqn (2),

$$\Sigma_k^p = \mathbf{w}_p^T \mathbf{C}_k^p \mathbf{w}_p,\ \Sigma_k^s = \mathbf{w}_s^T \mathbf{C}_k^s \mathbf{w}_s,$$
$$\mathbf{\Sigma}_k = \begin{bmatrix} \mathbf{w}_p^T \mathbf{C}_k^p \mathbf{w}_p & \mathbf{w}_p^T \mathbf{C}_k^{p,s} \mathbf{w}_s \\ \left(\mathbf{w}_p^T \mathbf{C}_k^{p,s} \mathbf{w}_s\right)^T & \mathbf{w}_s^T \mathbf{C}_k^s \mathbf{w}_s \end{bmatrix}, \quad (5)$$

where $\mathbf{C}_k^p$ and $\mathbf{C}_k^s$ are the covariance matrix of $\mathbf{X}_k^p$, $\mathbf{X}_k^s$, respectively, and $\mathbf{C}_k^{p,s}$ is the covariance matrix between $\mathbf{X}_k^p$ and $\mathbf{X}_k^s$.

Substituting Eqn. (5) into Eqn. (4), we find the equivalence between maximizing (4) and the Canonical Correlation Analysis (CCA) model

$$\max_{\mathbf{w}_p, \mathbf{w}_s} \frac{\mathbf{w}_p^T \mathbf{C}_k^{p,s} \mathbf{w}_s}{\sqrt{\mathbf{w}_p^T \mathbf{C}_k^p \mathbf{w}_p \mathbf{w}_s^T \mathbf{C}_k^s \mathbf{w}_s}}. \quad (6)$$

So the optimal $\mathbf{w}_p$ and $\mathbf{w}_s$ are obtained by solving CCA (details are given later). CCA is found with good trade-off between the scalability and performance, when the input set is usually of a large size (about 2.5 million sample pairs in our experiments).

To estimate the thresholds τ_p and τ_s, we use brute-force search to maximize (3) in the region $(\tau_p, \tau_s) \in [\hat{\mu}^p - \hat{\sigma}^p, \hat{\mu}^p + \hat{\sigma}^p] \times [\hat{\mu}^s - \hat{\sigma}^s, \hat{\mu}^s + \hat{\sigma}^s]$, where $\hat{\mu}^p = \text{median}_i(y_i^p)$ and $\hat{\sigma}^p = \text{median}_i(|y_i^p - \hat{\mu}^p|)$ are the median and median of absolute deviation of y_i^p, respectively, and $\hat{\mu}^s$ and $\hat{\sigma}^s$ are the median and median of absolute deviation of y_i^s, respectively.

Canonical Correlation Analysis. CCA was introduced by Hotelling for correlating linear relationships between two sets of vectors [10]. It was used in some computer vision applications [34, 13, 25]. However, it has not been explored as a component of a vector quantization algorithm. Blaschko and Lampert [4] proposed an algorithm for spectral clustering with paired data based on kernel CCA. However, this method is not appropriate for quantization, as the kernel trick causes high computational and memory cost due to the very large size of the training set, and the nearest centroid assignment may be unstable (there is no hard constraint to require a pair of vectors in the same cluster).

To solve CCA in (6), let

$$\mathbf{S}_m = \begin{bmatrix} \mathbf{0} & \mathbf{C}_k^{p,s} \\ (\mathbf{C}_k^{p,s})^T & \mathbf{0} \end{bmatrix}, \mathbf{S}_n = \begin{bmatrix} \mathbf{C}_k^p & \mathbf{C}_k^{p,s} \\ (\mathbf{C}_k^{p,s})^T & \mathbf{C}_k^s \end{bmatrix},$$

and then $\mathbf{w} = [\mathbf{w}_p^T, \mathbf{w}_s^T]^T$ can be solved as the eigenvector associated with the largest eigenvalue of the generalized eigenvalue problem $\mathbf{S}_m\mathbf{w} = \lambda(\mathbf{S}_n + \varepsilon\mathbf{I})\mathbf{w}$, where ε is a small positive number for regularization.

The whole algorithm for building a CITP tree is summarized as Algorithm 1.

Algorithm 1 Algorithm of building a CITP Tree

1: **Input**: a set of vector pairs $\mathcal{X} = \{(\mathbf{x}_i^p, \mathbf{x}_i^s), i = 1, ..., N\}$, where $\mathbf{x}_i^p, \mathbf{x}_i^s \in \mathbb{R}^D$, and the expected number of codes (i.e. leaf nodes) n_L.
2: Create an empty set $\mathcal{S}$, and add the root node to $\mathcal{S}$.
3: **repeat**
4: **for** each node k in $\mathcal{S}$ and its associated vector set $\mathcal{X}_k$ **do**
5: Compute the possible node splitting:
(i) Generate projection vectors $\mathbf{w}_p, \mathbf{w}_s$ and thresholds τ_p, τ_s with $\mathcal{X}_k$;
(ii) For its left child L and right child R,
$\mathcal{X}_L \leftarrow \{(\mathbf{x}_i^p, \mathbf{x}_i^s) | \mathbf{w}_p^T \mathbf{x}_i^p \le \tau_p, \mathbf{w}_s^T \mathbf{x}_i^s \le \tau_s\}$,
$\mathcal{X}_R \leftarrow \{(\mathbf{x}_i^p, \mathbf{x}_i^s) | \mathbf{w}_p^T \mathbf{x}_i^p > \tau_p, \mathbf{w}_s^T \mathbf{x}_i^s > \tau_s\}$,
$(\mathcal{X}_L \subset \mathcal{X}_k, \mathcal{X}_R \subset \mathcal{X}_k)$;
6: **end for**
7: Select the best node splitting with the maximum mutual information in Eqn. (1);
8: Split the node, remove the node from $\mathcal{S}$ and add its child nodes to $\mathcal{S}$;
9: **until** the number of leaf nodes is n_L.
10: **Output**: the CITP tree with projection vectors and thresholds at each node.

2.4. Randomized CITP Forest

Randomization is an effective way to create an ensemble of trees to boost the performance of tree structured algorithms [21, 23, 33]. The randomized counterpart of the CITP tree includes two modifications on node splitting as follows.

Randomization in sub-vector choice. At each node, we randomly sample α percent (empirically $\alpha = 80$) of the element indices of the sampled vectors, i.e. use a sub-vector of each sampled vector, to learn the projections. To improve the strength of generated trees, the random choice is repeated for 10 times empirically at each node, and the one

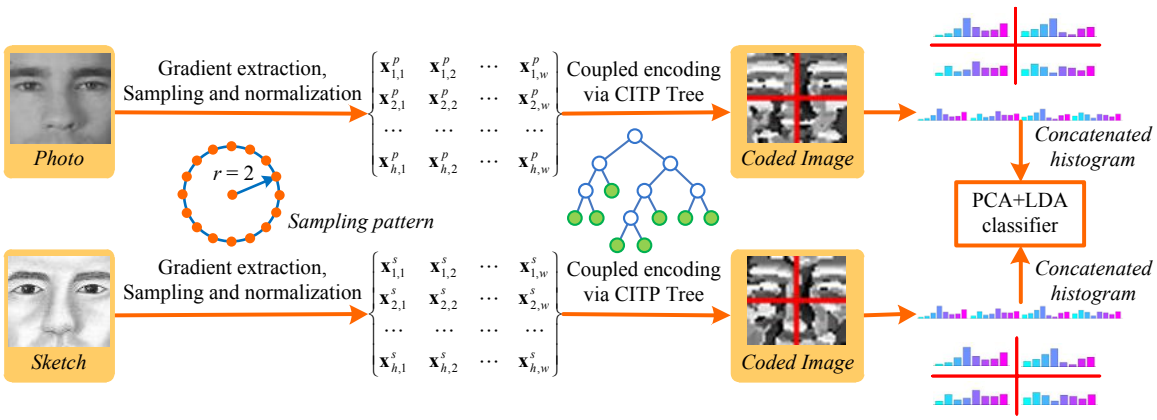


Figure 3. The pipeline of extracting CITE descriptors.

with the maximum mutual information in Eqn. (3) is selected. The randomization at each node results in randomized trees with different tree structures and utilizing different information from the training data. Therefore, the randomized trees are more complementary.

Randomization in parameter selection. The eigenvectors associated with the first d largest eigenvalues in the CCA model are first selected. Then a set of n vectors are generated by randomly linearly combining the d selected eigenvectors.[5] According to the MMI criterion in Eqn. (3), the best one is selected from the set of n random vectors and used as the projection vectors $\mathbf{w}_p$ and $\mathbf{w}_s$. In our experiments, we choose $d = 3$ and $n = 20$.

The creation of a random ensemble of diverse trees can significantly improve the performance over a single tree, which is verified by our experiments.

3. Coupled Encoding Based Descriptor

In this section, we introduce our coupled information-theoretic encoding (CITE) based descriptor. With a CITP tree, a photo or a sketch can be converted into an image of discrete codes. The CITE descriptor is a collection of region-based histograms of the "code" image. The pipeline of photo-sketch recognition using a single CITP tree is shown in Fig. 3. The details are given as follows.

Preprocessing. The same geometric rectification and photometric rectification are applied to all the photos and sketches. With affine transform, the images are cropped to 80×64, and the two eye centers and the mouth center of all the face images are at fixed positions. Then both the photo and sketch images are processed with a Difference-of-Gaussians (DoG) filter [11] to remove both high-frequency and low-frequency illumination variations. Empirical investigations show that $(\sigma_1, \sigma_2) = (1, 2)$ is the best in our experiments.

[5]The eigenvectors are orthogonalized with Gram-Schmidt orthogonalization and normalized with L_2-norm.

Sampling and normalization. At each pixel, its neighboring pixels are sampled in a certain pattern to form a vector. A sampling pattern is a combination of one or several rings and the pixel itself. On a ring with radius r, $8r$ pixels are sampled evenly. Fig. 3 shows the sampling pattern of $r = 2$. We denote a CITE descriptor by a sampling pattern with rings of radius $r_1, ..., r_s$ as $\text{CITE}_{r_1,...,r_s}$.

We find that sampling the gradients I_u and I_v results in a better descriptor than sampling the intensities [5]. The gradient domain explicitly reflects relationships between neighboring pixels. Therefore, it has more discriminating power to discover key facial features than the intensity domain. In addition, the similarity between photos and sketches are easier to compare in the gradient domain than intensity domain [35].

After the sampling, each sampled vector is normalized such that its L_2-norm is unit.

Coupled Information-Theoretic Encoding. In the encoding step, the sampled vectors are turned into discrete codes using the proposed CITP tree (Section 2). Then each pixel has a code and the input image is converted into a "code" image. The vectors sampled from photos and sketches for training CITP tree are paired according to the facial landmarks detected by a state-of-the-art alignment algorithm [16].[6] Specifically, a pixel in the sketch image finds its counterpart in the photo image using a simple warping based on the landmarks. Note that the pairing is performed after sampling so that local structures are not deformed by the warping.

CITE Descriptor. The image is divided into 7×5 local regions with equal size, and a histogram of the codes is computed in each region. Then the local histograms are concatenated to form a histogram representation of the image, i.e. the CITE descriptor.

[6]According to our observation, a general face alignment algorithm trained on commonly used face photo data sets is actually also effective for sketch alignment. We did not separately train a face alignment algorithm for sketches.

Figure 4. Examples of photos from the CUFSF database and corresponding sketches drawn by the artist.

Classifier. We use a simple PCA+LDA classifier[7] [3, 29] to compute the dissimilarity between a photo and a sketch. By learning a linear projection matrix on the training set, it projects CITE descriptors into a low-dimensional space. Note that the descriptors are centered, i.e. the mean of the training CITE descriptors is subtracted from them. Then each projected CITE descriptor is normalized to a unit L_2-norm and the Euclidean distance between the normalized low-dimensional representation of a photo and a sketch is computed as their dissimilarity.

Fusion. We use a linear SVM to fuse dissimilarities by different CITE descriptors. The different CITE descriptors can be obtained by running the randomized CITP tree algorithm repeatedly. To train the one-class SVM, we select all the intrapersonal pairs and the same number of interpersonal pairs with smallest dissimilarities.

4. Experiments

In this section, we study the performance of our CITE descriptors and CITP trees on face photo-sketch recognition task. We first compare the performance of our CITE descriptor, with a single sampling pattern and single tree, to popular facial features, including LBP [1] and SIFT [19]. The classifier is not used in this part to clearly show their difference. Then we investigate the effect of various free parameters on the performance of the system. Finally we show that our method is superior to the state-of-the-art.

Datasets. The CUHK Face Sketch FERET Database (CUFSF) is used for the experiments. There are $1,194$ people with lighting variations in the set. Each person has a photo and a sketch with shape exaggeration drawn by an artist. Some examples are shown in Fig. 4. The CUFS database [31] is also used as a benchmark. This dataset consists of 606 persons, each of which has a photo-sketch pair. The sketches were drawn without exaggeration by an artist when viewing the photo.

On the CUFSF database, 500 persons are randomly selected as the training set, and the remaining 694 persons form the testing set. On the CUFS database, 306 persons are in the training set and the other 300 persons are in the testing set.

Evaluation metrics. The performance is reported as Verification Rates (VR) at 0.1% False Acceptance Rate (FAR) and Receiving Operator Characteristic (ROC) curves.

[7] A small regularization parameter is added to the diagonal elements of the within-class matrix of LDA to avoid singularity.

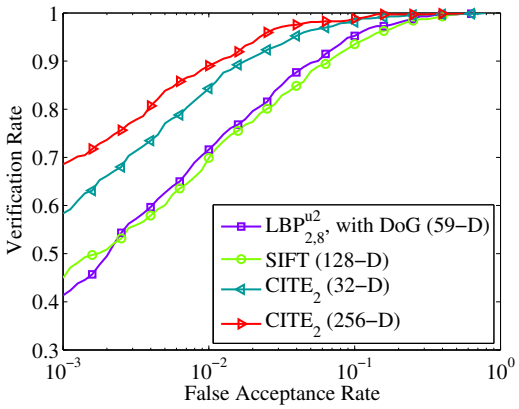


Figure 5. Comparison between $CITE_2$ (single CITP tree), LBP and SIFT. The dissimilarity between a photo and a sketch is computed as the distance between descriptors extracted on them. The χ^2 distance [1] is used for LBP and $CITE_2$, and Euclidean distance is used for SIFT. For simplicity, we give the length of a local histogram for each descriptor, instead of the length of the whole descriptor, in brackets.

4.1. Descriptor Comparison

We compare our descriptor with LBP [1] and SIFT [19]. The LBP is computed based on sampling points on a circle. We explore different numbers of sampling points and different radiuses. We find that the LBP descriptors extracted from DoG filtered images perform better than from original images. The 128-dimensional SIFT has 4×4 spatial bins of the same size and 8 orientation bins evenly spaced over $0^\circ - 360^\circ$. The vote of a pixel to the histogram is weighted by its gradient magnitude and a Gaussian window with parameter σ centered at the center of the region. We explore different sizes of the region and different σ. For our CITE descriptor, we use the sampling pattern of a single ring with $r = 2$ as shown in Fig. 3. We test on different numbers of leaf nodes (i.e. different sizes of a local histogram).

The ROC curves are shown in Fig. 5. Even 32-dimensional $CITE_2$ (please refer to Section 3 for this notation) significantly outperforms the 59-dimensional LBP and 128-dimensional SIFT. The 256-dimensional $CITE_2$ (68.58%) beats the best results of LBP (41.35%) and SIFT (44.96%) by 20% on VR at 0.1% FAR.

4.2. Parameter Exploration

We investigate the effect of various free parameters on the performance of the system, including the number of leaf nodes, the projected dimension by PCA+LDA, the size of randomized forest and the effect of using different sampling patterns. We fix the other factors when investigating one parameter.

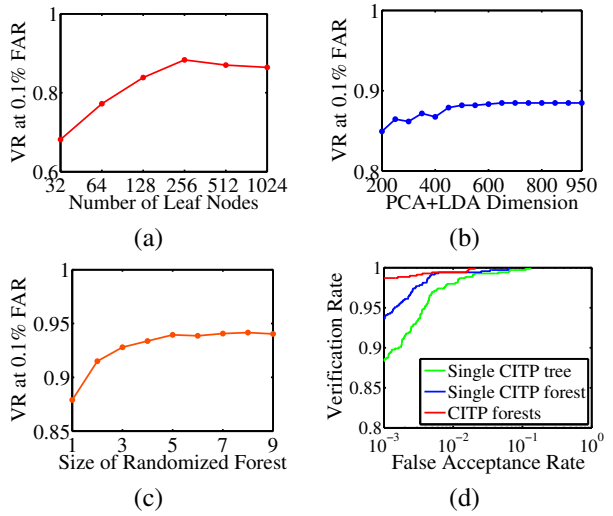


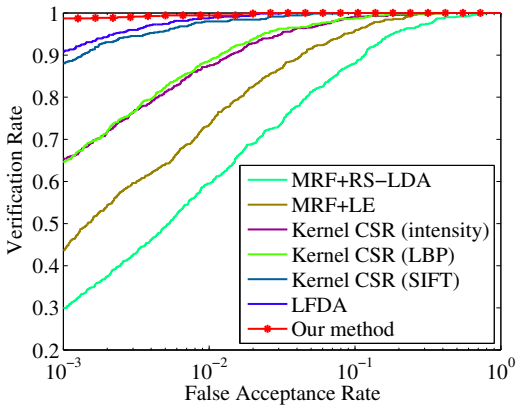
Figure 6. VR at 0.1% FAR vs. (a) number of leaf nodes; (b) PCA+LDA dimension; (c) size of randomized forest; (d) comparison of ensemble of forests with different sampling patterns and the forest with a single sampling pattern. In (a)–(c), The descriptor is CITE_2. In (a), the descriptors are compressed to 600 dimensional using PCA+LDA, and a single CITP tree is used. In (b), we use 256 leaf nodes and a single CITP tree. In (c) and (d), we use 256 leaf nodes and 600 PCA+LDA dimensions.

Number of Leaf Nodes. We compare the effect of using different numbers of leaf nodes in a CITP tree. The number is extensively studied from 32 (2^5) to 1024 (2^{10}). As shown in Fig. 6(a), the VR initially increases, and does not increase when the number is larger than 256. Due to small performance gain and high computational cost of a large leaf node number, we choose 256 leaf nodes as our default setting.

PCA+LDA Dimension. The reduced dimension is an important parameter of PCA+LDA. The VR has a fairly large stable region and varies less than 1% from 500 to 950 (see Fig. 6(b)). We choose 600 PCA+LDA dimensions in our final system.

Size of Randomized Forest. We vary the number of randomized trees in the CITP forest from 1 to 9. Fig. 6(c) shows that increasing the number of trees from 1 to 5 increases the VR from 87.90% to 93.95%, with little improvement beyond this. Hence, we fix the number of randomized trees in a CITP forest to be 5.

Ensemble of Randomized Forests with Different Sampling Patterns. Although the performance increases slowly when the number of randomized trees is more than 5, using ensemble of randomized forests with different sampling patterns can further boost the performance. Different sampling patterns can capture rich information across multiple scales. Fig. 6(d) shows that using five sampling patterns improves the VR at 0.1% FAR from 93.95% to 98.70%.



VR at 0.1% FAR		
MRF+RS-LDA	MRF+LE	LFDA
29.54%	43.66%	90.78%
Kernel CSR (LBP)	Kernel CSR (SIFT)	Ours
64.55%	88.18%	98.70%

Figure 7. Comparison of the state-of-the-art approaches and our method on the CUFSF database. ROC curves and VR at 0.1% FAR are shown.

4.3. Experiments on Benchmarks

We compare our algorithm with the following state-of-the-art approaches on the CUFSF database. The algorithms are tuned to the best settings according to their paper.

- MRF-based synthesis [31]. Pseudo photos are synthesized from query sketches, and random sampling LDA (RS-LDA) [30] is used to match them to gallery photos. In addition, we test LE [5] on matching pseudo photos and gallery photos.
- Kernel CSR [15]. The CSR model is trained to seek for a common discriminative subspace, based on intensities, LBP and SIFT feature vectors separately.
- LFDA [14]. It fuses the LBP features with four different radiuses and the SIFT features with a discriminative model. For each feature, multiple projection vectors are learnt.

Fig. 7 shows that our method significantly outperforms the state-of-the-art approaches. MRF-based synthesis requires that there is no significant shape distortion between photos and sketches in the training set, and also that training photos are taken under similar lighting conditions. This method does not work well in this new data set because the drawing style of the artist involves large shape exaggeration and the photos in the FERET database are taken under different lightings with large variations. Therefore, the pseudo photos by MRF-based synthesis have artifacts such as distortions. Such artifacts degrade the performance of state-of-the-art face photo recognition algorithms including RS-LDA and LE. The results of Kernel CSR on different

Table 1. Rank-1 recognition rates on the CUFS database. The recognition rates are averaged over five random splits of 306 training persons and 300 testing persons. We test our method with the same configuration of training and testing splits as [31, 14].

MRF+RS-LDA [31]	LFDA [14]	Ours
96.30%	99.47%	99.87%

features verify that the inappropriate selection of features will reduce the discriminative power of the classifier. SIFT features have better results than LBP on the photo-sketch recognition task. LFDA achieves a good result by fusing five different kinds of features with two different spatial partitions. However, its error rate (9.22%) is much higher than ours (1.30%) for 0.1% FAR.

Our method also has superior performance on the CUFS database, a standard benchmark for face photo-sketch recognition, as shown in Table 1. Apparently, this dataset is now an easy one for the state-of-the-art methods.

5. Conclusions

We proposed a coupled information-theoretic encoding based descriptor for face photo-sketch recognition. We introduced coupled information-theoretic projection forest to maximize the mutual information between the encoded photo and encoded sketch of the same subject. Our system significantly outperforms the state-of-the-art approaches. In the future work, we would like to further investigate the system with more cross-modality recognition problems.

Acknowledgements

The authors would like to thank many previous members of multimedia lab in CUHK for their contributions to the CUFSF database, and Zhimin Cao for valuable discussion on learning-based descriptor.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE TPAMI*, 28(12):2037, 2006. 514, 518

[2] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9(7):1545–1588, 1997. 514

[3] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 2002. 518

[4] M. Blaschko and C. Lampert. Correlational spectral clustering. In *CVPR*, 2008. 516

[5] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, 2010. 514, 517, 519

[6] L. Chang, M. Zhou, Y. Han, and X. Deng. Face sketch synthesis via sparse representation. In *ICPR*, 2010. 513

[7] T. Cover and J. Thomas. *Elements of information theory*. John Wiley and Sons, 2006. 515, 516

[8] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *NIPS*, 2007. 514, 515

[9] X. Gao, J. Zhong, J. Li, and C. Tian. Face sketch synthesis algorithm based on E-HMM and selective ensemble. *IEEE TCSVT*, 18(4):487–496, 2008. 513, 514

[10] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321, 1936. 516

[11] G. Hua and A. Akbarzadeh. A robust elastic and partial matching metric for face recognition. In *ICCV*, 2009. 517

[12] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005. 514

[13] T. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE TPAMI*, pages 1415–1428, 2008. 516

[14] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mugshot photos. *IEEE TPAMI*, 33(3):639–646, 2011. 513, 514, 519, 520

[15] Z. Lei and S. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*, 2009. 513, 514, 519

[16] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *ECCV*, 2008. 517

[17] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, 2006. 513, 514

[18] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *CVPR*, 2005. 513, 514

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 514, 518

[20] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, 2001. 514

[21] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, 2007. 514, 516

[22] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE TPAMI*, 22(10):1090–1104, 2002. 514

[23] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 514, 516

[24] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, 2003. 514

[25] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 516

[26] X. Tang and X. Wang. Face photo recognition using sketch. In *ICIP*, 2002. 513

[27] X. Tang and X. Wang. Face sketch synthesis and recognition. In *ICCV*, 2003. 513, 514

[28] X. Tang and X. Wang. Face sketch recognition. *IEEE TCSVT*, 14(1):50–57, 2004. 513

[29] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE TPAMI*, 26(9):1222–1228, 2004. 518

[30] X. Wang and X. Tang. Random sampling for subspace face recognition. *IJCV*, 70(1):91–104, 2006. 519

[31] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 31(11):1955–1967, 2009. 513, 514, 518, 519, 520

[32] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*, 2008. 514

[33] J. Wright and G. Hua. Implicit elastic matching with random projections for pose-variant face recognition. In *CVPR*, 2009. 514, 516

[34] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li. Face matching between near infrared and visible light images. *Advances in Biometrics*, pages 523–530, 2007. 516

[35] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *ECCV*, 2010. 517

[36] J. Zou, Q. Ji, and G. Nagy. A comparative study of local matching approach for face recognition. *IEEE TIP*, 16(10):2617–2628, 2007. 514