Pattern Recognition 114 (2021) 107815

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Discriminative shared transform learning for sketch to image matching

Shruti Nagpal[a], Maneet Singh[a], Richa Singh[b], Mayank Vatsa[b],*

[a] IIIT-Delhi, New Delhi, India
[b] IIT Jodhpur, India

ARTICLE INFO

Article history:
Received 20 June 2020
Revised 30 December 2020
Accepted 31 December 2020
Available online 12 January 2021

Keywords:
Face recognition
Sketch to digital image matching
sketch based image retrieval
Caricature face recognition
Transform learning

ABSTRACT

Sketch to digital image matching refers to the problem of matching a sketch image (often drawn by hand or created by a software) against a gallery of digital images (captured via an acquisition device such as a digital camera). Automated sketch to digital image matching has applicability in several day to day tasks such as similar object image retrieval, forensic sketch matching in law enforcement scenarios, or profile linking using caricature face images on social media. As opposed to the digital images, sketch images are generally edge-drawings containing limited (or no) textural or colour based information. Further, there is no single technique for sketch generation, which often results in varying artistic or software styles, along with the interpretation bias of the individual creating the sketch. Beyond the variations observed across the two domains (sketch and digital image), automated sketch to digital image matching is further marred by the challenge of limited training data and wide intra-class variability. In order to address the above problems, this research proposes a novel *Discriminative Shared Transform Learning (DSTL)* algorithm for sketch to digital image matching. DSTL learns a shared transform for data belonging to the two domains, while modeling the class variations, resulting in discriminative feature learning. Two models have been presented under the proposed DSTL algorithm: (i) Contractive Model (C-Model) and (ii) Divergent Model (D-Model), which have been formulated with different supervision constraints. Experimental analysis on seven datasets for three case studies of sketch to digital image matching demonstrate the efficacy of the proposed approach, highlighting the importance of each component, its input-agnostic behavior, and improved matching performance.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

Sketch to digital image matching is a special kind of cross-domain matching task, where given a query instance of domain *A* (sketch), samples of the same category are retrieved from a different domain *B* (digital image) (as illustrated in Fig. 1). Sketch-to-digital image matching finds applicability in important real-world applications such as surveillance by means of forensic sketch face recognition [1], and social media profile linking or person identification via caricature recognition [2]. With the increasing usage of touchscreens and e-commerce, sketch based image retrieval has become an integral part of image retrieval tasks as well [3–5]. The given problem suffers from the availability of limited training data, and possesses a unique challenge where the intra-domain similarity often overpowers the intra-class similarity. For example, a sample appears more similar to another class's sample from the same domain, as compared to its own class's sample from a different domain.

In order to address the cross-domain matching task of sketch to image matching, researchers have proposed several algorithms to model the inter-domain variations [6]. Most of the existing techniques have either focused on learning an explicit transformation function from one domain to another, or on learning a shared space for two domains. The first approach involves learning a transformation function between the two domains [7,8], however, it often requires large amount of data for modeling the inter-domain variations. For instance, in face recognition applications such as sketch to digital image matching, a learned transformation should be generalizable to new identities seen only during testing. The second approach explicitly models the inter-domain variations by learning a common space for data belonging to two domains [9–11]. Algorithms belonging to this category learn representations in a shared space, thus minimizing the cross-domain variations. This is often achieved by utilizing *representation learning*, which is able to encode effective features for the input data. Beyond cross-domain variations, the problem of sketch to image matching is further exacerbated with the availability of limited training samples.

* Corresponding author.
E-mail addresses: shrutin@iiitd.ac.in (S. Nagpal), maneets@iiitd.ac.in (M. Singh), richa@iitj.ac.in (R. Singh), mvatsa@iitj.ac.in (M. Vatsa).

https://doi.org/10.1016/j.patcog.2021.107815
0031-3203/© 2021 Elsevier Ltd. All rights reserved.

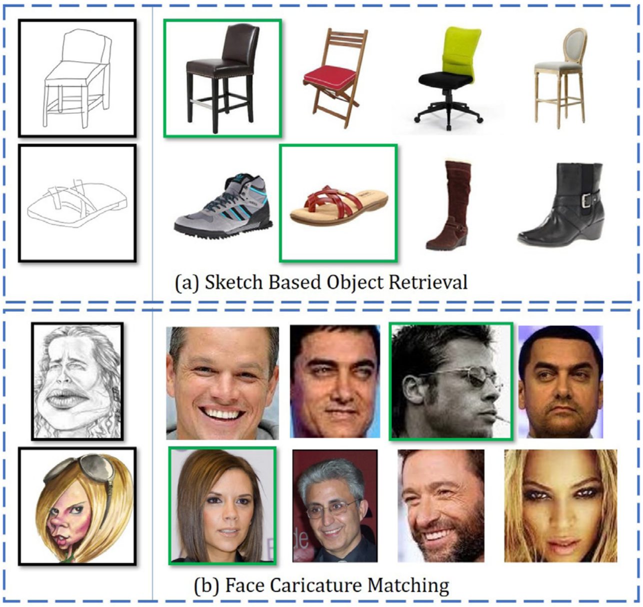

(a) Sketch Based Object Retrieval
(b) Face Caricature Matching

Fig. 1. Sample applications of cross-domain matching demonstrating variation in the information content. Query from domain *A* (i.e. sketch domain) is matched against a database of another domain *B* (i.e. image domain).

This research proposes a novel *Discriminative Shared Transform Learning* algorithm for sketch to digital image matching. The proposed algorithm learns domain invariant and discriminative features under supervised constraints. It is able to learn effective features for cross-domain matching with a small number of parameters, and improve the current state-of-the-art performance. The key highlights of this research are:

- This research presents a novel learning algorithm, *Discriminative Shared Transform Learning*, for cross-domain matching, specifically sketch to digital image matching. As part of the proposed technique, two models have also been presented: (i) *Contractive model (C-model)*[1], and (ii) *Divergent model (D-model)*. The efficacy of the proposed algorithm is demonstrated on seven datasets pertaining to three sketch to digital image matching case-studies: (i) caricature face recognition, (ii) sketch based object image retrieval, and (iii) sketch face recognition.
- A shared transform is learned which enables learning of domain-invariant features for data belonging to different domains. It also reduces the number of parameters to be learned, thereby making it a suitable choice for scenarios with limited training data.
- The proposed algorithm is feature agnostic, i.e. it can operate on different features as input in order to achieve enhanced performance. Extensive evaluation of the proposed C-model and D-model demonstrates the effectiveness of the proposed algorithm with different input features. Detailed analysis also highlights the importance of each component of the proposed algorithm.

2. Related work

With increase in automated computer vision applications, problems related to cross-domain matching, specifically sketch to digital image matching, have been gaining attention of the research community. According to the case studies presented in Fig. 1, the related work is divided into two subsections: (i) Sketch based Image Retrieval and (ii) Sketch and Caricature Face Recognition.

Sketch based Image Retrieval (SBIR): SBIR refers to the task of retrieving digital images similar to the given sketch query image pertaining to an object. Since the query and retrieved samples belong to different domains, cross-domain matching is performed. SBIR is relevant in today's context for applications such as online shopping websites, or for searching similar images by sketching the object on our touch screens.

Due to the lack of color and texture information in object sketch images, traditional approaches utilized structural information to extract features for matching. Cao et al. [13] proposed a real time approach based on structure index, termed as EdgelIndex and a corresponding contour based matching algorithm. Similar to the Bag of Words model, Eitz et al. [14] proposed the Bag of Visual Words (BOVW) model for SBIR. BOVW utilizes images as visual words, followed by clustering for generating a codebook, and histogram matching. Hu and Collomosse [15] presented a variant of the popular Histogram of Oriented Gradients (HOG) feature extractor, termed as Gradient Field-HOG (GF-HOG), useful for extracting low-level features from sketch and digital images. A combination of GF-HOG and BOVW was presented for improved SBIR performance.

Recently, some deep learning models have also been proposed for sketch based image retrieval. Qi et al. [16] presented a Siamese based convolutional neural network for SBIR. Yu et al. [17] presented a new dataset of shoes and chairs for fine-grained SBIR. The authors proposed a triplet ranking deep model and a novel pre-training strategy. This was followed by a spatially aware attention module [9], which combines fine semantic information along with coarse information for fine-grained SBIR. Recently, Huang et al. [18] highlighted and addressed the several challenges faced in SBIR. The authors proposed a deep visual semantic descriptor to encode low level and high level features for both the domains, followed by a clustering based re-ranking approach. Zhang et al. [19] presented a technique which dynamically discovers landmarks, which aids in learning the discriminative structural representations. Further, Zhang et al. [20] proposed a Hybrid CNN model for modeling the appearance and shape information for sketch based image retrieval. Sketch based image object retrieval has also been addressed by utilizing pre-trained deep learning models with domain-specific information [10,21,22]. Owing to the large number of parameters, most of the deep learning based models suffer from the major challenge of requiring large amount of labeled training data (either for training or pre-training), and often, labeled data pairs across domains. The data is required to effectively train the model in order to learn efficient representations, often rendering the model unscalable for large number of classes when trained with limited data.

Sketch and Caricature Face Recognition: Ouyang et al. [6] documented the existing techniques for sketch based face recognition [23,24], including the other scenarios of heterogeneous face matching such as infra-red, 3D, and low resolution. For sketch face recognition, past approaches can be divided into feature based [25–27], synthesis based [28], and projection based [29,30]. Feature based approaches focus on extracting domain invariant features from the given data. Synthesis based approaches transform data from one domain to another in order to reduce the inter domain gap, while projection based approaches project data belonging to both the domains onto a common subspace. Since projection based techniques do not transform data from one domain to another, they are often preferred in comparison to synthesis based methods. In the past, researchers have also tried to address the memory gap present in sketch face recognition by modeling the human forgetting process [31]. de Freitas Pereira et al. [32] proposed utilizing the high-level features of deep learning models for learning domain specific

[1] A preliminary version of this formulation has been published in IEEE BTAS 2018 [12].

units. Nagpal et al. [8] proposed deep transform learning for sparse feature extraction, followed by a mapping between the representations of the two domains.

Limited research has been performed to automate the process of face caricature matching. Klare et al. [34] utilized 25 qualitative features describing the face features such as the face shape, nose shape, and hair type to encode the image and caricature faces. The authors also created the first publicly available dataset for caricature face recognition. Takayama et al. [35] proposed using a similarity vector based on skin color, hair type, and hair quantity for face caricature recognition. Recently, Shi et al. [36] proposed a caricature generation technique for creating a caricature from a given digital face image. It is important to note that most of the existing automated caricature face recognition approaches have primarily focused on using annotated qualitative features. This results in the requirement of annotated information for the testing data, which often limits the utility of the recognition algorithm.

Despite the recent advances in sketch to digital image matching, the performance achieved by the state-of-the-art techniques presents a need for further improvement, especially for real world scenarios (eg. 19.44% rank-10 performance on the IIIT-D Forensic sketch dataset [26]). Most of the existing approaches focus on extracting hand-crafted features, useful for particular domains; utilize task-specific knowledge (eg. incorporating facial landmarks for face retrieval); or contain large number of trainable parameters requiring ample amount of paired training data from different domains (eg. deep learning based techniques). Recent representation learning techniques transform data from one domain to another or project data onto a common subspace. As mentioned previously, learning a generalized transformation function is often a challenging task, requiring large amount of training data, while projections onto a common subspace may often result in the loss of useful discriminative information.

3. Preliminaries: transform learning

This research focuses on the transform learning (TL) paradigm [37] of representation learning for cross-domain matching. Transform learning is an unsupervised learning approach which learns a *transform matrix* and corresponding features for the given data. Its ability to learn sparse features under the applied constraints has resulted in high performance for several tasks such as image classification, denoising, deblurring, and face recognition [8,38–40]. Transform learning is the analysis equivalent of the dictionary learning algorithm [41]. It analyzes the input ($\mathbf{X}$) and learns a transform ($\mathbf{T}$) to produce the corresponding sparse coefficients/representations $\mathbf{Z}$. A condition for sparsity is enforced to promote sparse features. Mathematically, transform learning is expressed as:

$$\min_{\mathbf{T},\mathbf{Z}} \|\mathbf{TX}-\mathbf{Z}\|_F^2 + \lambda\left(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\right)\ s.t.\|\mathbf{Z}\|_0 \le \tau \quad (1)$$

where, $\|\mathbf{T}\|_F^2$ corresponds to the ℓ_2-norm of the transform matrix, and the second regularization term is the log-determinant [42] which imposes full rank on the learned transform to prevent degenerate solutions. λ and ϵ correspond to the weights given to the regularizers. Given n input images of dimension $m \times m$, $\mathbf{X}$ contains the images stacked in a column-wise manner, such that each image corresponds to a column vector. This results in a matrix of dimension $m^2 \times n$. $\mathbf{T}$ and $\mathbf{Z}$ are of dimensions $t \times m^2$ and $t \times n$, respectively, where t corresponds to the dimension of the learned representation. In order to optimize the above equation, an alternating minimization approach is followed [39,43], which iterates over the following two steps:

$$\mathbf{Z} \leftarrow \min_{\mathbf{Z}} \|\mathbf{TX}-\mathbf{Z}\|_F^2,\ s.t.\ \|\mathbf{Z}\|_0 \le \tau \quad (2)$$

$$\mathbf{T} \leftarrow \min_{\mathbf{T}} \|\mathbf{TX}-\mathbf{Z}\|_F^2 + \lambda\left(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\right) \quad (3)$$

Alternating minimization updates one variable while keeping the remaining constant, thereby resulting in a two step update for transform learning. The representations (or coefficients) in Eq. (2) can be updated via any sparse recovery algorithm, such as the Orthogonal Matching Pursuit [44]. The transform matrix is obtained by solving Eq. (3) using a three step approach given by Ravishankar et al. [40]:

$$\mathbf{XX^T} + \lambda\epsilon\mathbf{I} = \mathbf{LL^T};\ \mathbf{L^{-1}XZ^T} = \mathbf{USV^T} \quad (4)$$

$$\mathbf{T} = \mathbf{0.5V}\left(\mathbf{S} + (\mathbf{S^2} + \mathbf{2\lambda I})^{\frac{1}{2}}\right)\mathbf{U^TL^{-1}} \quad (5)$$

The first step corresponds to a Cholesky decomposition on the input data, which is followed by performing a full Singular Value Decomposition. Finally, Eq. (5) gives the update step for the transform matrix $\mathbf{T}$. In-depth proof of convergence and analysis of the above solution can be found in [43].

4. Proposed discriminative shared transform learning

Traditionally, transform learning is not directly applicable to cross-domain matching. In order to use the vanilla transform learning methods for heterogeneous matching, separate transforms are learned for the two domains, followed by matching via a classifier or distance metric (Fig. 2(a)). In the literature, transform learning has been extended for cross-domain matching by incorporating a mapping matrix between the learned features (Fig. 2(b)) [8]. These techniques do not explicitly minimize the inter-domain variations, and require learning multiple transforms, resulting in a large number of learnable parameters, which is often challenging with the limited training data available. This research addresses the above limitations by proposing *Discriminative Shared Transform Learning (DSTL)* to learn domain-invariant features for matching data belonging to different domains, primarily sketch and digital images. The proposed algorithm enables learning of domain invariant features, which encode differentiable information, and is not data intensive, i.e. suitable for limited training data.

4.1. DSTL: components and formulation

The DSTL algorithm is built using two novel components: (i) shared transform learning, and (ii) discriminative feature learning. Details regarding each are provided in this subsection.

Shared Transform Learning (STL): It involves learning a single transform matrix ($\mathbf{T}$) for data belonging to two domains. A single transform enables the model to learn features common across domains, thereby reducing the inter-domain variations. In sketch-to-digital image matching, this would correspond to projecting the digital and sketch images using the same transform matrix. STL eliminates the need for multiple domain-specific transforms, and facilitates learning of a transform capable of representing data of both the domains. For training samples $\mathbf{X_1}$ and $\mathbf{X_2}$ of two domains, STL is formulated as:

$$\min_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1}-\mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2}-\mathbf{Z_2}\|_F^2 + \lambda\left(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\right) \quad (6)$$

where, $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are the learned representations for the corresponding input data, and $\mathbf{T}$ is the shared transform matrix.

Discriminative Feature Learning: Since STL is unsupervised in nature, it does not model the relationship between different classes. To this effect, the second component, i.e. *discriminative feature learning*, is incorporated into the proposed DSTL algorithm. Class information is utilized via discriminative losses, such that

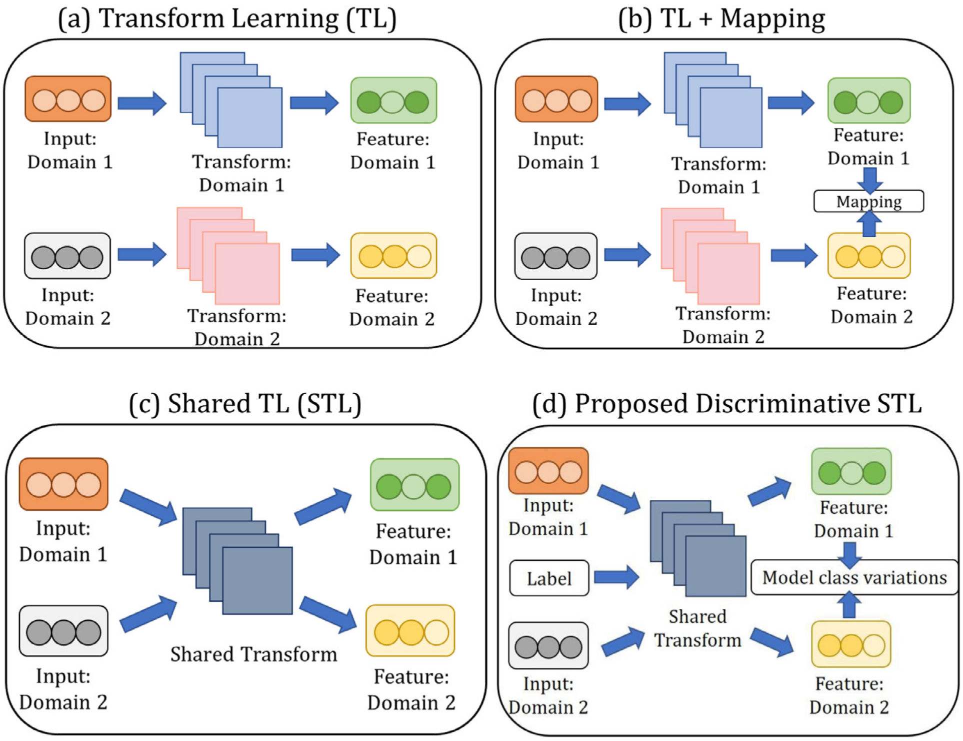


Fig. 2. Diagrammatic representation of how transform learning can be used for cross-domain matching.

the learned features are useful for classification. A distance term, $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ is introduced in the loss function, which aids in learning discriminative features by encoding class variations at the time of feature learning.

Given training data pertaining to two different input domains, namely digital image, $\mathbf{X_1}$, and sketch image, $\mathbf{X_2}$, the corresponding representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are learned using the proposed DSTL algorithm. Building on Eq. (6), DSTL is mathematically expressed as follows:

$$\min_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\left(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\right) + \mu\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y}) \tag{7}$$

where, λ, μ, and ϵ correspond to the regularization parameters which govern the weight given to different terms.

4.2. Variants of the DSTL algorithm

DSTL enables learning domain invariant features under supervised constraints. The distance term ($\mathcal{F}(.)$) in Eq. (7) can be modeled for handling the class variations across domains. In sketch to digital image matching, samples of the same class, belonging to different domains, suffer from the challenge of high intra-class variations due to the varying information content in both the input samples. In order to model the class variations across domains, the *Contractive Model (C-model)* of DSTL is proposed to learn features for sketch to digital image matching.

Contractive Model (C-model): The C-model learns a shared transform matrix while reducing the intra-class variations. Here, the input $\mathbf{X_1}$ and $\mathbf{X_2}$ are created such that they contain pair-wise data. That is, each pair contains samples of the same class, belonging to the two domains. Specifically, in case of sketch-to-digital image matching, the i^{th} sample of $\mathbf{X_1}$ and $\mathbf{X_2}$ would belong to class-A, where one would be a sketch image and the other is a digital image. Therefore, C-model does not utilize the class labels (Y) directly into the formulation, however, it uses the class information to create pairs for training. Given same-class (or genuine) pairs only, $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ is modeled as the Euclidean distance between the representations. This corresponds to a similarity-preserving loss which minimizes the distance between the representations of the same class pairs. C-model is thus formulated as:

$$\min_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\left(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\right) + \mu\|\mathbf{Z_2} - \mathbf{Z_1}\|_F^2 \tag{8}$$

where, the first two terms learn the shared transform and features, followed by the regularizers to prevent a degenerate solution, and a term to reduce the intra-class variations. The final term promotes features which minimize the difference between the representations of the same class, cross-domain pairs. While C-model handles only the intra-class similarity, cross-domain matching also suffers from the intertwined problem of low inter-class variations as well. In order to incorporate both intra-class and inter-class variations, building on the C-model, the *Divergent Model (D-model)* model is proposed.

Divergent Model (D-model): As shown in Fig. 3, D-model utilizes the class labels at the time of feature learning, such that the samples of same class have similar representations, and representations of samples of different classes are far apart. With reference to Eq. (7), $\mathbf{X_1}$ and $\mathbf{X_2}$ refer to cross-domain image pairs, and $\mathbf{y}$ contains the label corresponding to each pair specifying whether they belong to the same class (0) or different classes (1). In case of sketch to digital face image matching, the i^{th} sample of $\mathbf{X_1}$ and $\mathbf{X_2}$ may correspond to the sketch and digital image of class-A with the i^{th} element of $\mathbf{y}$ as 0. On the other hand, another j^{th} sample of $\mathbf{X_1}$ and $\mathbf{X_2}$ may correspond to the sketch and digital image of class-A and class-B, respectively, with the j^{th} element of $\mathbf{y}$ as 1. This results in cross-domain pairs belonging to the same or different classes. For a set of training pairs $\mathbf{X_1}$ and $\mathbf{X_2}$ belonging to two domains, with the label $\mathbf{y}$ (0 for same or 1 for different), $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ repre-

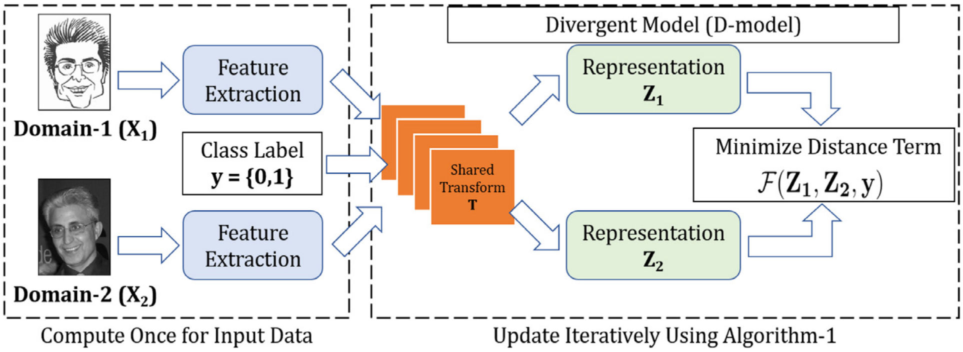


Fig. 3. The proposed D-model uses cross-domain image pairs along with their labels. D-model learns a shared transform **T** while modeling the inter-class and intra-class variations. It is feature-agnostic and can be learned on raw input or extracted features such as HOG or VGG-Face.

sents a distance function which introduces discriminability during the feature learning process by handling the inter-class and intra-class variations. In this research, $\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$ has been conceptualized as the contrastive loss [45], expressed as:

$$\begin{aligned}\mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y}) = (1-\mathbf{y})\tfrac{1}{2}\ (\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T) \\ + (\mathbf{y})\tfrac{1}{2}\ max\Big(0, (m - \mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T)\Big)\end{aligned} \tag{9}$$

where, $\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})$ is a distance function applied on the learned representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$ to model the relationship between the two, and m is the margin value. $\mathbf{y}$ is a $1 \times n$ vector containing the class labels, and the transpose operator ensures compatibility between the vectors. The margin ensures that only those inter-class pairs whose distance is less than the margin contribute to the loss function. The distance function $\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})$ is modeled as the Euclidean distance between the representations $\mathbf{Z_1}$ and $\mathbf{Z_2}$. The above loss minimizes the Euclidean distance between cross-domain intra-class pairs, and focuses on those inter-class pairs which have a distance less than the margin m. After incorporating the contrastive loss, the proposed D-model is formulated as:

$$\begin{aligned}\min\nolimits_{\mathbf{T},\mathbf{Z_1},\mathbf{Z_2}} \|\mathbf{TX_1}-\mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2}-\mathbf{Z_2}\|_F^2 + \lambda\Big(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\Big) \\ + \mu\bigg((1-\mathbf{y})\tfrac{1}{2}\ (\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T) + (\mathbf{y})\tfrac{1}{2}\ max\Big(0, (m-\mathcal{D}(\mathbf{Z_1}, \mathbf{Z_2})^T)\Big)\bigg)\end{aligned} \tag{10}$$

where, similar to C-model, the first two terms learn the shared transform $\mathbf{T}$, and the representations $\mathbf{Z_1}, \mathbf{Z_2}$ for the two domains. The next two terms are the regularizers to prevent a degenerate solution, and the final two terms correspond to the distance loss applied on the representation. The proposed D-model thus learns a shared transform for cross-domain data. Representations are learned such that the intra-class variations across domains are minimized, and the inter-class separability is maximized. One of the major highlights of the proposed algorithm is that it is feature-agnostic, i.e. it can be applied to raw input or other features extracted from the images.

4.3. Optimization of C-model and D-model

Both C-model and D-model are optimized using the alternating minimization technique, which iteratively optimizes over all the variables. Each variable ($\mathbf{T}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$) is optimized in an alternate manner, while keeping the other variables constant. The following three steps are performed in an iterative manner for learning the C-model:

Update for T:

$$\begin{aligned}\min\nolimits_{\mathbf{T}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2}-\mathbf{Z_2}\|_F^2 + \lambda\Big(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\Big) \\ \equiv \min\nolimits_{\mathbf{T}} \left\|\mathbf{T}\binom{\mathbf{X_1}}{\mathbf{X_2}} - \binom{\mathbf{Z_1}}{\mathbf{Z_2}}\right\|_F^2 + \lambda\Big(\epsilon\|\mathbf{T}\|_F^2 - log\det\mathbf{T}\Big)\end{aligned} \tag{11}$$

The above equation is in the form of a standard transform learning model (Eq. (3)) and thus can be solved using the three step approach described earlier (Eqs. (4)–(5)).

Update for $\mathbf{Z_1}$:

$$\min\nolimits_{\mathbf{Z_1}} \|\mathbf{TX_1}-\mathbf{Z_1}\|_F^2 + \mu\|\mathbf{Z_2}-\mathbf{Z_1}\|_F^2 \equiv \min\nolimits_{\mathbf{Z_1}} \left\|\binom{\mathbf{TX_1}}{\sqrt{\mu}\mathbf{Z_2}} - \binom{\mathbf{I}}{\sqrt{\mu}\ \mathbf{I}}\mathbf{Z_1}\right\|_F^2 \tag{12}$$

Update for $\mathbf{Z_2}$:

$$\min\nolimits_{\mathbf{Z_2}} \|\mathbf{TX_2}-\mathbf{Z_2}\|_F^2 + \mu\|\mathbf{Z_2}-\mathbf{Z_1}\|_F^2 \equiv \min\nolimits_{\mathbf{Z_2}} \left\|\binom{\mathbf{TX_2}}{\sqrt{\mu}\mathbf{Z_1}} - \binom{\mathbf{I}}{\sqrt{\mu}\ \mathbf{I}}\mathbf{Z_2}\right\|_F^2 \tag{13}$$

Eqs. (12) and (13) are least square problems having closed form solutions. The above three updates are repeated iteratively until convergence or maximum iterations. Similarly, Algorithm 1 presents the step-wise approach for optimizing D-model.

Algorithm 1: Step wise optimization technique for training the D-model.

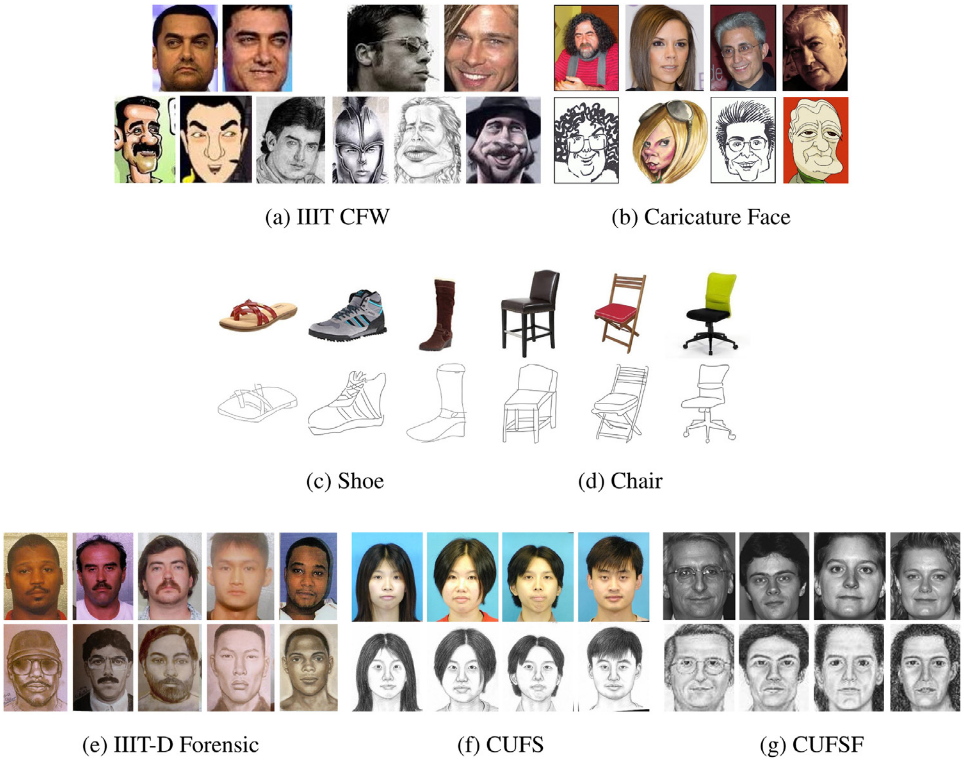
```
Input  : $\mathbf{X_1}, \mathbf{X_2}, \mathbf{y}$
Output: $\mathbf{T}, \mathbf{Z_1}, \mathbf{Z_2}$
while MaxIter do
    $\mathbf{T} \leftarrow \arg\min_{\mathbf{T}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \lambda\big(\epsilon\|\mathbf{T}\|_F^2 - \log\det\mathbf{T}\big)$
    $\mathbf{Z_1} \leftarrow \arg\min_{\mathbf{Z_1}} \|\mathbf{TX_1} - \mathbf{Z_1}\|_F^2 + \mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$
    $\mathbf{Z_2} \leftarrow \arg\min_{\mathbf{Z_2}} \|\mathbf{TX_2} - \mathbf{Z_2}\|_F^2 + \mathcal{F}(\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{y})$
end
```

5. Experimental setup

To demonstrate the applicability of the proposed approach across different sketch to digital image matching applications, experiments are performed with seven datasets pertaining to three case studies: (i) caricature to face image matching, (ii) sketch-based object image retrieval, and (iii) face sketch to digital image matching. Most of the experiments are performed under the real world scenario of limited data availability. Details regarding the datasets and protocols are given below.

5.1. Datasets and protocol

The proposed C-model and D-model have been evaluated on seven datasets (as shown in Fig. 4) for three different cross-domain



(a) IIIT CFW (b) Caricature Face

(c) Shoe (d) Chair

(e) IIIT-D Forensic (f) CUFS (g) CUFSF

Fig. 4. Sample images from the datasets used in experiments. The first row of each dataset contains the digital images and the second row contains images of the other domain.

Table 1
Dataset details for the three case studies of sketch to digital image matching. 'Extnd. Gallery' refers to the number of images in the extended gallery for that dataset. Domain-A refers to digital photographs, while Domain-B refers to the sketch based images.

Case-Study	Dataset	Number of Images in	
		Domain-A	Domain-B
Caricature Face Recognition	Caricature DB [34]	207	207
	IIIT CFW DB [46]	1,000	8,298
Sketch Based Image Retrieval	Chair DB [17]	297	297
	Shoe DB [17]	419	419
Sketch Face Recognition	IIIT-D Forensic Sketch DB [26]	190 + 7,125 Extnd. Gallery	190
	CUFS DB [47]	188	188
	CUFSF DB[48]	1,194	1,194

matching applications. Details regarding each dataset are given in Table 1. Unless explicitly mentioned, pre-defined protocols provided in the respective publication of each dataset have been followed for generating the training and testing splits. Each protocol results in disjoint training and testing partitions.

Case Study-1: Caricature Face Recognition: With the advent of social media and availability of multiple communication platforms, the use of caricatures has increased tremendously.[2] This creates a need for performing caricature to digital face image matching[3], applicable in several scenarios including image retrieval and social media profile linking. Results have been demonstrated on two publicly available datasets:

- IIIT Cartoon Faces in the Wild (IIIT-CFW) dataset [46] contains 8,298 caricatures and 1,000 face images of 100 public figures. Pre-defined training and testing splits have been used for performing the recognition experiment, where data corresponding to 42 identities is used for training, while the remaining 58 subjects form the test set.
- Caricature dataset [34] contains paired caricature and digital face images of 207 subjects, such that each subject has a single face image and a caricature image. The dataset contains both hand-drawn and digital caricature images collected from the Internet, thereby making the problem further challenging. On this dataset, images pertaining to 138 subjects are used for training, while the remaining 69 subjects form the test set.

Case Study-2: Sketch Based Image Retrieval (SBIR): SBIR refers to the task of retrieving digital images corresponding to a probe sketch image. SBIR has received dedicated attention for its applicability in scenarios of online shopping, similar object retrieval, and image based search. Experiments have been performed on two datasets:

- Chair dataset [17] contains 297 paired sketch and photo images. Experiments are performed using the pre-defined protocol, where 200 pairs are used for training, while the remaining form the test set.
- Shoe dataset [17] contains 419 pairs of sketch and photo images. Experiments are performed using the pre-defined protocol, where 304 pairs form the training partition, and the remaining form the test set.

Case Study-3: Sketch Face Recognition: Sketch face recognition has been a long standing problem which involves matching a sketch face image with its corresponding digital face image. Due to its applicability in forensic scenarios, it has also received substantial attention; however, matching forensic sketch images with digital face images remains a challenging and unsolved task. The

[2] https://tinyurl.com/y8z2kje6

[3] Caricature face recognition refers to matching a caricature face image with a gallery of digital face images.

reasons include larger intra-class variations and availability of limited training data. Results have been demonstrated on three publicly available datasets - two for viewed sketch face recognition and one for forensic sketch recognition. Details of each database are given as follows:

- IIIT-D Forensic Sketch dataset [26] has been used for evaluating the proposed model for forensic sketch face recognition. It contains 190 forensic sketches with corresponding digital face images. The sketches have been created based on the description provided by an eye-witness to the artist in real-world crime scene scenarios. Consistent with existing literature [8], and simulating the real-world scenarios, matching has been performed against an extended gallery of 7,265 digital face images, including 190 digital face images from the IIIT-D Forensic Sketch dataset. Due to the lack of a training set for the IIIT-D Forensic Sketch dataset, cross-dataset experiment is performed, where IIIT-D Semi-Forensic dataset [26] has been used for training. The IIIT-D Semi-Forensic dataset contains 140 sketch-digital face image pairs which simulate semi-forensic scenarios. It is ensured that the training and testing sets are subject disjoint.
- CUHK Face Sketch (CUFS) dataset [47] contains viewed sketches hand-drawn by an artist for a frontal image captured under normal lighting conditions with a neutral expression. Consistent with the existing protocol in the literature [49], experiments have been performed on the CUHK student dataset, containing 188 sketch-photo pairs. 100 image pairs are used for training, while the test set contains the remaining 88 image pairs.
- CUHK Face Sketch FERET (CUFSF) dataset [48] contains 1,194 images from the FERET dataset [50] and their corresponding shape exaggerated hand-drawn sketches, created while viewing the image with lighting variations. Consistent with the literature [49], 297 pairs are used for testing and the rest form the train set.

5.2. Experimental details

Since a key highlight of the proposed approach is its independence to the input feature, experiments are performed with four different inputs: raw pixel values, Local Binary Patterns (LBP) [51], Histogram of Oriented Gradients (HOG) [52], and a pre-trained deep learning based feature extractor. For face recognition, the VGG-Face model [53] is the pre-trained deep learning model, while ResNet-152 [54] has been used for experiments on object images. VGG-Face is a pre-trained CNN model on the large-scale VGG-Face dataset, and ResNet-152 is pre-trained on the ImageNet dataset [55]. In the literature, both the models have demonstrated state-of-the-art classification performance for face and object recognition, respectively. Comparison has also been performed with state-of-the-art algorithms for each case-study.

5.3. Implementation details

For training the C-model and D-model, same-class pairs are generated by combining each sample of domain A with all the samples of the same class in domain B. Same number of different-class pairs are generated by combining samples of domain A with randomly chosen samples of domain B from a different class. The proposed models are trained on the input data for 50 epochs. The margin parameter (m) for D-model has been set based on grid search. Classification is performed on the learned features using Euclidean distance. Data augmentation is performed by flipping across the y-axis and performing illumination variations. The dimensions of the existing models used for comparison like DeepTransformer [33] are the same as proposed in the paper. The neural network used with DeepTransformer is of dimension $[k/2, k/4]$. The proposed models have been implemented in the Matlab R2018a environment. VLFeat and MatConvNet have been used for the extraction of LBP, HOG, and deep learning features. The experiments were performed on a workstation with 64GB RAM and one Nvidia K40 GPU. We observed that the time taken to perform matching is comparable with classical object and face recognition matching algorithms. The proposed approach takes less than 1 ms to match a pair of images.

Table 2
Rank-10 matching accuracy (%) for caricature face recognition and sketch-based image retrieval. Accuracies have been reported after matching with the original features, transform learning (TL only) features, shared transform learning (STL) features, contractive model (C-model), and divergent model (D-model). The proposed models demonstrate improved performance across different input features.

Input	Original	TL Only	STL	C-model	D-model
Caricature Face Recognition					
Caricature Face Dataset					
Pixels	19.56	23.19	22.46	27.53	**30.43**
LBP	18.84	18.84	20.29	26.81	**26.81**
HOG	34.05	28.26	31.16	39.13	**39.85**
VGG-Face	44.93	61.59	68.11	69.57	**78.98**
IIIT CFW Dataset					
Pixels	22.44	23.47	23.49	24.53	**28.16**
LBP	26.71	25.27	25.89	33.74	**35.20**
HOG	32.70	29.89	29.19	34.17	**36.47**
VGG-Face	59.26	79.16	77.29	79.98	**86.05**
Sketch-based Object Image Retrieval					
Shoe Dataset					
Pixels	8.70	9.57	9.57	13.91	**18.26**
LBP	37.39	39.13	30.43	40.00	**52.17**
HOG	80.00	80.00	79.13	80.00	**86.96**
ResNet-152	39.13	46.09	26.96	50.43	**55.65**
Chair Dataset					
Pixels	14.43	20.62	14.43	25.77	**45.36**
LBP	61.86	63.92	55.67	77.32	**83.51**
HOG	94.85	93.81	92.78	95.88	**96.91**
ResNet-152	85.57	86.60	72.16	**94.85**	93.81

Table 3
Rank-10 matching accuracy (%) comparing the proposed D-model with other algorithms. Accuracies are reported on the Caricature Face Dataset (CF), and the IIIT CFW dataset. The proposed D-model and DeepTransformer models are trained with VGG-Face features.

Algorithm	CF	IIIT CFW
COTS [56]	0	8.06
MMSS [57]	39.13	38.13
GSMFL [58]	47.10	41.01
Pixel + Neural Network	23.19	24.43
VGG-Face + DeepTransformer [8] + NNet	31.15	31.86
VGG-Face + Proposed D-model	**78.98**	**86.05**

Table 4
Rank-10 accuracy (%) of the proposed D-model and other algorithms on the Shoe and Chair datasets.

Algorithm	Shoe DB	Chair DB
BoW-HOG [17]	67.83	67.01
Dense-HOG [17]	65.22	93.81
ISN Deep [17]	62.61	82.47
Triplet model [17]	**87.83**	**97.94**
Song et al. [60]	**91.30**	**98.97**
Song et al. [61]	**94.78**	95.88
HOG + DeepTransformer [8] + NNet	53.91	91.75
HOG + Proposed D-model	**86.96**	**96.91**

6. Results and analysis

Experiments have been performed by using (i) the features as it is (referred to as Original), (ii) vanilla transform learning [37] (TL only), where transforms are learned for each domain indepen-

Table 5
Rank-10 matching accuracy (%) of the proposed D-model and other algorithms on the IIIT-D Forensic Sketch dataset with a large-scale gallery.

Algorithm	Forensic DB
COTS [56]	2.63
MCWLD [26]	14.71
GSMFL [58]	13.46
L-CSSE + DeepTransformer [8] + NNet	19.44
HOG + Proposed D-model	**35.26**

dently, (iii) Shared Transform Learning (STL), and the proposed (iv) C-model and (v) D-model (Table 2). As mentioned previously, results are demonstrated with different input features, i.e., (i) raw pixels, (ii) LBP, (iii) HOG, and (iv) deep learning based feature extractor. Comparison has also been performed with the other cross-domain matching techniques and state-of-the-art results reported on each dataset (Tables 3–6). Consistent with the existing protocols, for all the datasets, results are reported in the form of accuracy at top-k, where $k = 10$, i.e. percentage of test images whose corresponding true match was retrieved in the top 10 ranks. For some sketch face recognition experiments, $k = 1$, i.e. rank-1 accuracy is reported. The following subsections present results pertaining to each case study, and analysis of the proposed models.

6.1. Case study specific analysis

Tables 2 to 6 present the results obtained for the different sketch to digital image experiments. Analysis corresponding to each case-study is provided below:

Case Study-1: Caricature Face Recognition: Table 2 presents the rank-10 or top-10 accuracy of the proposed models with different input features. The proposed D-model+VGG-Face features performs the best by achieving 78.98% and 86.05% on the Caricature Face dataset and the IIIT-CFW dataset, respectively (Table 2). This can primarily be attributed to the fact that VGG-Face is a learning based model which has specifically been trained for faces, and thus demonstrates the best performance. Table 3 compares the performance of the proposed model with other recently proposed cross-domain algorithms, a Commercial off-the-shelf system (COTS) [56], and a neural network of dimension $[\frac{k}{2}\, \frac{k}{4}]$ trained on raw pixels. It can be observed that VGG-Face+D-model outperforms other cross-domain matching techniques by at least 30% on both datasets.

In literature, the only comparative accuracy on the Caricature Face Dataset is by Klare et al. [34], where the authors achieve a rank-10 accuracy of 74.8%. However, they performed experiments with 197 subjects, whereas the released dataset contains 207 subjects, which has been used in this study.[4] The proposed D-model achieves an improvement of almost 5% on this dataset. There do not exist any reported results on the IIIT-CFW dataset. Further, the proposed model outperforms the existing transform learning based cross-domain matching model, DeepTransformer [8], which demonstrates the benefit of encoding the class variations across domains during feature learning, and learning a shared transform as opposed to multiple transforms.

Fig. 5(a) presents sample learned weights of the transform matrix on raw pixels. Outline of faces and caricatures, along with some salient features can be observed. For instance, some of the weights are characterized by exaggerated jaw lines and nose boundaries, while some contain balanced facial features. Upon analyzing the mis-classified samples for caricature face recognition (Fig. 5(b)), we observed that most of the images were of varying pose. Exaggeration of facial features, extreme pose variations, and different artistic techniques renders the problem further challenging.

We have also performed additional experiments on the IIIT-CFW dataset for two recent few shot protocols provided by Zheng et al. [59]. For the task of Few-Shot Photo to Caricature Recognition, we obtain an accuracy of 86.73% which is 1.2% higher than what is reported by Zheng et al. [59]. Additionally, on the second protocol of Few-Shot Photo to Caricature Recognition, we obtain 97.64%, as opposed to 93.7% reported by authors. The results demonstrate the effectiveness of the proposed approach on benchmark datasets.

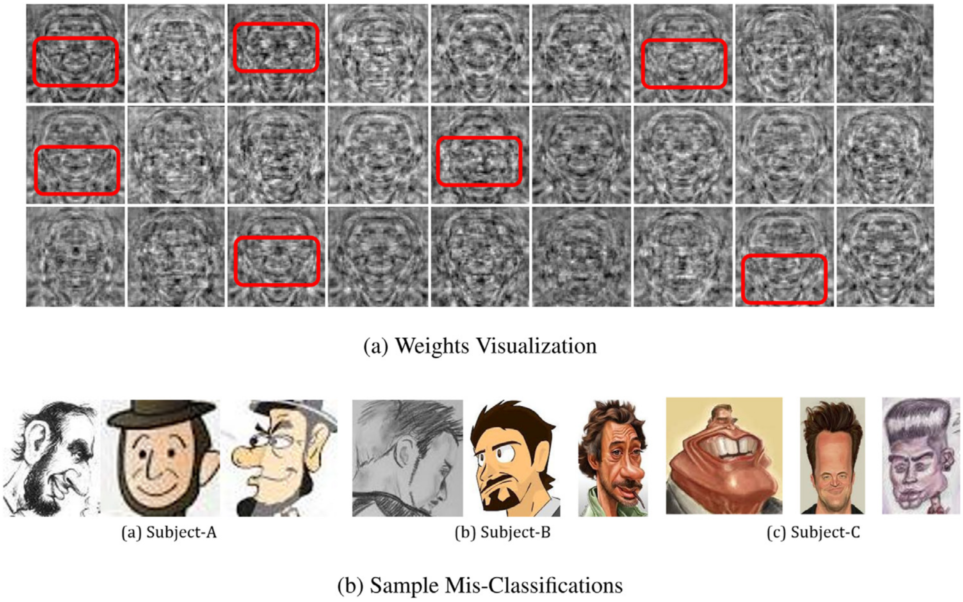
Case Study-2: Sketch based Image Retrieval (SBIR): As observed in Table 2, the proposed D-model with HOG features performs best by achieving 86.96% and 96.91% on the Shoe and Chair datasets, respectively. One of the key reasons for this behavior is that HOG features encode the gradient information, which provide important distinct information about sketches. Table 4 compares the matching accuracy of the proposed D-model with other techniques for sketch based object image retrieval. Owing to the fixed protocol of the dataset, results have directly been taken from the respective publications. It can be observed that the proposed D-model with HOG features is among the top performing models for the given datasets. It is interesting to note that the proposed model mis-classifies only three samples of the chair database, whereas the best performing model mis-classifies a single sample [60], and most of the recent models have been pre-trained on the ImageNet photo-edgemap pairs [61].

For the application of sketch-based object image retrieval, we observe that the proposed technique is able to model shape and visual features of the images quite well. Fig. 6(a) presents the rank-list of two samples which are *not* classified by the proposed HOG+D-model. In the first case, all the retrieved samples have the same shape as the query image, and a similar pattern of legs with wheels. Similarly, in the second case, the retrieved samples have the same shape, which demonstrates the utility of the proposed model for retrieving similar images as well. Fig. 6(b) presents the rank-list obtained by HOG features and HOG+D-model features for a sample query of the Shoe dataset. HOG+D-model is able to retrieve the correct match at rank-2, which is not the case with HOG features based matching. We believe that the D-model further enforces discriminability in the HOG features, thus enhancing the classification performance.

Case Study-3: Sketch Face Recognition: On the IIIT-D Forensic Sketch dataset, the proposed D-model (with HOG features as input) provides the best matching performance. Table 5 compares its performance with the state-of-the-art results. Since we follow the same benchmark protocol, results have directly been taken from Nagpal et al. [8]. As compared to state-of-the-art DeepTransformer with L-CSSE (deep learned feature), the proposed D-model with HOG demonstrates an improvement of around 15%, and an improvement of over 32% is observed from COTS at rank-10. Fig. 7 presents the Cumulative Match Characteristics (CMC) curves obtained on the forensic sketch dataset. At rank-50, an improvement of over 25% is observed with the proposed HOG+D-model, as compared to the state-of-the-art. Fig. 8 presents visualizations of sample weights learned by the transform learning model on the sketch images for pixel input. The model appears to encode the facial features, especially at the global level. For example, different weights appear to model varying face shape and hairstyles. Since face sketches tend to have more holistic information as compared to minute local features, modeling the global information enables better feature extraction, thereby facilitating improved classification performance.

Table 6 presents the results on the two viewed sketch face datasets: CUFS [47] and CUFSF [48]. The proposed HOG+D-model

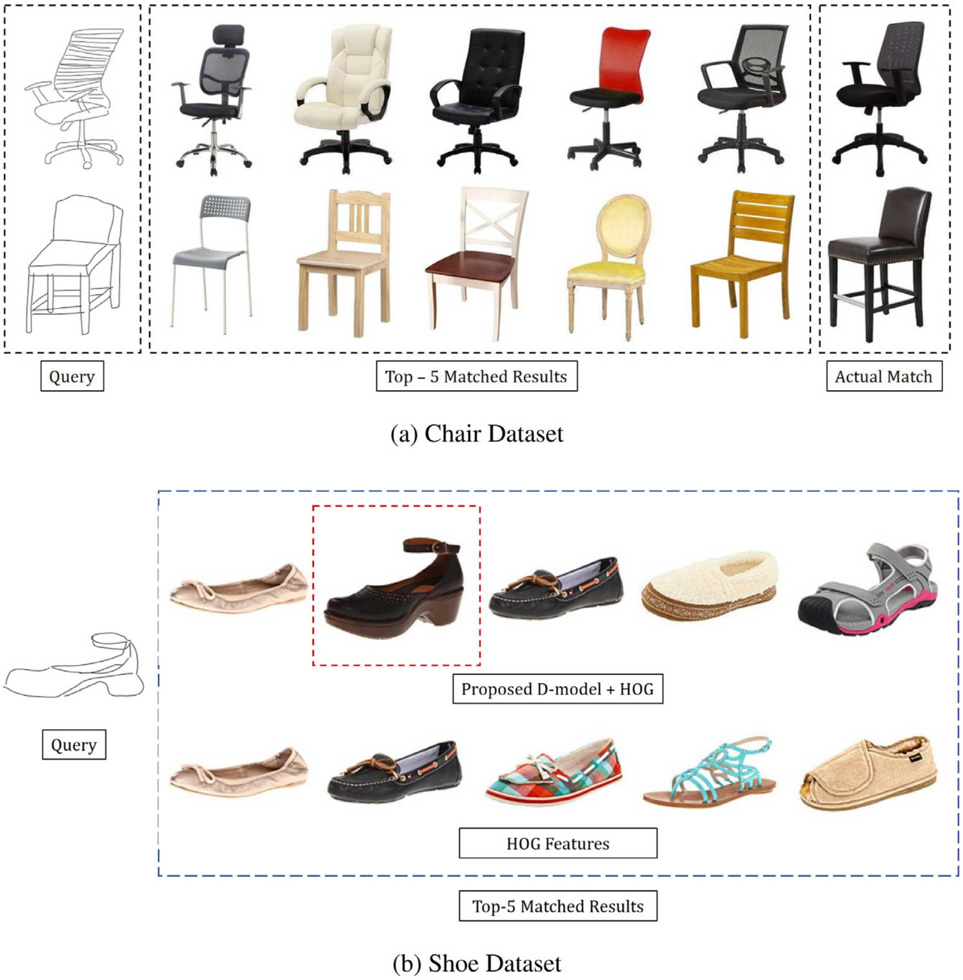
[4] For consistency with Klare et al. [34], one-third and two-third subjects are used for testing and training, respectively.



(a) Weights Visualization

(b) Sample Mis-Classifications

Fig. 5. (a) Visualizations of sample weights learned by the proposed D-model on the Caricature Face dataset, for pixel input. It can be observed that the model learns different components of faces and caricatures - both exaggerated and balanced. (b) Sample images mis-classified by the D-model.



(a) Chair Dataset

(b) Shoe Dataset

Fig. 6. Sample retrieval results for the Chair and Shoe dataset with the proposed D-model with HOG features as input.

Table 6
Rank-1 accuracy (%) of the proposed D-model and other comparative algorithms on the CUFS and CUFSF datasets.

Algorithm	CUFS	CUFSF
Pix2Pix [62]	100.0	37.0
CycleGAN [64]	99.0	25.0
DualGAN [63]	100.0	35.0
Multi-Adversarial Networks ($PS^2 - GAN$) [49]	100.0	47.0
Semi-coupled Dictionary Learning [66]	95.2	-
Multi-Paced Dictionary [67]	98.4	-
Generalized Coupled Dictionary Learning [65]	98.0	-
Locality-Constrained Joint Dictionary + Residual Learning [68]	98.2	-
Multi-view Domain Translation [69]	98.1	-
HOG + Proposed D-model	**100.0**	**67.3**

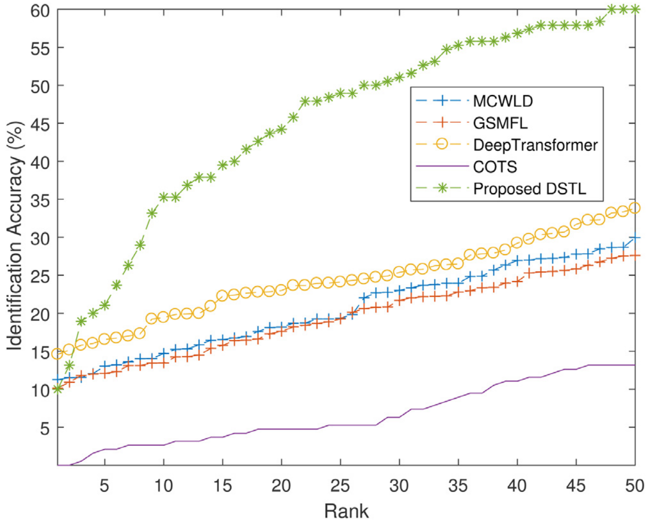


Fig. 7. Cumulative Match Characteristics (CMC) curves on the IIIT-D Forensic Sketch database.

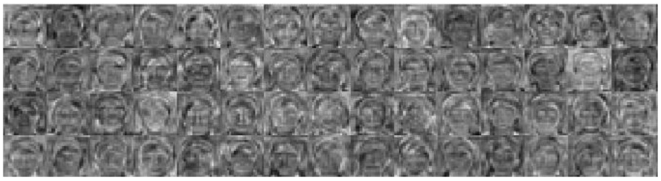


Fig. 8. Visualization of sample weights learned by the proposed D-model algorithm, for sketch face recognition on raw pixels.

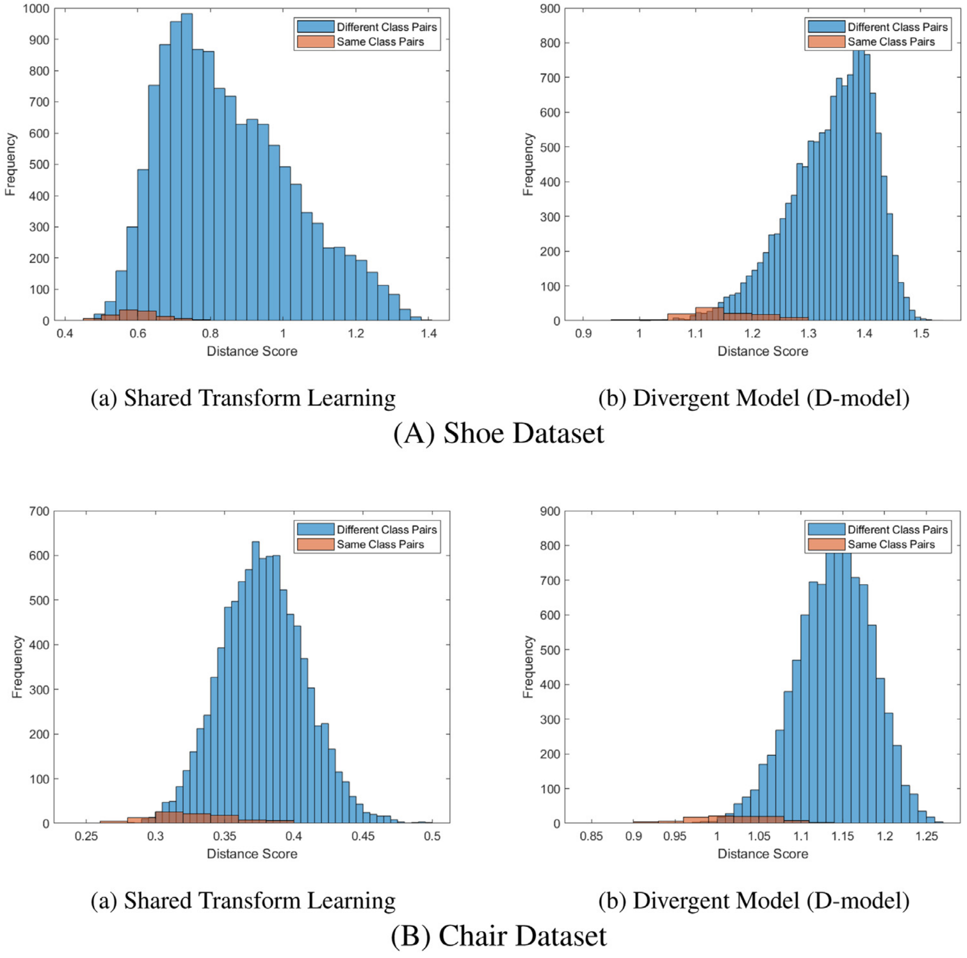
achieves 100% and 67.3% on the CUFS and CUFSF dataset, respectively. Comparison has been drawn with existing Generative Adversarial Network (GAN) based techniques such as Pix2Pix [62], DualGAN [63], and CycleGAN [64] (owing to the same protocol, results have directly been taken from Wang et al. [49]). The proposed D-model with HOG features outperforms GAN based image synthesis techniques on both the datasets. Comparison has also been performed with existing dictionary learning based techniques (owing to the same protocol, results have directly been taken from their respective publications, and Mandal and Biswas [65]), where the proposed model demonstrates an improvement of atleast 1.5% on the CUFS dataset. The results showcase the efficacy of the D-model for sketch to photo face recognition.

6.2. Overall analysis of the proposed models

The proposed C-model and D-model have also been analyzed in terms of the effect of input feature, ablation study, and number of trainable parameters:

Effect of Input Features: The proposed models are evaluated with different input features, both hand-crafted and deep learning based. From Table 2, it is observed that both hand-crafted and deep learning based features perform better than raw pixels, across all case-studies. This can be attributed to the challenging nature of the problem, which requires informative features for matching across domains. Moreover, features which outperform others without any transform, attain the best results when given as input to the D-model as well. For example, VGG-Face features (original) yields the highest accuracy as compared to other features on the Caricature Face dataset and the IIIT CFW dataset. With the proposed D-model, it achieves state-of-the-art performance with an improvement of around 26-33%. Further, for shoe and chair databases, handcrafted features yield better results as compared to deep learning features.

Ablation Study: Table 2 also presents the ablation study performed on the proposed D-model. The table can be read from right to left to analyze the effect of each component. D-model embeds both intra-class and inter-class variations, while C-model focuses on reducing the intra-class variations only. The Shared Transform Learning (STL) model does not utilize the class labels, and only projects the input onto a common space. Further, Transform Learning only (TL only) eliminates the use of shared transform and uses two transforms for feature learning, without any class information. The first column, Original, uses Euclidean distance directly on the input for matching. Across different case-studies and input features, the proposed D-model achieves highest classification performance, as compared to other techniques. C-model, which focuses on reducing the intra-class variations only, demonstrates improved performance as compared to the original features, and unsupervised Transform Learning models. This motivates the usage of reducing intra-class variations while learning representations. Upon adding the term for modeling the inter-class separability as well (D-model), the improvement in accuracy is further pronounced. In most cases, across features, D-model outperforms C-model, which motivates the utility of the proposed model for different cross-domain matching tasks. The classification accuracy of TL only and STL models can be compared with C-model and D-model in order to further promote the inclusion of class information in the proposed models. An improvement of 4-31% is observed from STL to D-model, which can be attributed to the discriminative supervision constraints. It is interesting to note that while TL only improves the classification performance as compared to the Original features in almost all scenarios, similar improvement is not observed for Shared Transform Learning. This implies that projecting cross-domain data onto a common space, without additional supervision constraints may not enhance classification. The benefit of incorporating discriminative supervision constraints can also be observed from Fig. 9. Modeling the inter-class and intra-class variations provides a separation between the same-class and different-class pairs, thus enhancing the classification performance.



(a) Shared Transform Learning

(b) Divergent Model (D-model)

(A) Shoe Dataset

(a) Shared Transform Learning

(b) Divergent Model (D-model)

(B) Chair Dataset

Fig. 9. Score distributions obtained after (a) Shared Transform Learning and (b) Divergent Model (D-model) on two datasets for sketch based image retrieval. The separation of same and different class scores promote inclusion of discriminative class-specific terms in the proposed model (best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Further, D-model pushes different-class pairs to have a higher distance score as compared to the same-class pairs.

Number of Parameters: For a vector input of dimension $n \times 1$, and a square shared transform of $n \times n$, a representation of dimension $n \times 1$ is obtained. Here, the proposed C-model and D-model only learn $n \times n$ parameters for the single shared transform. This strengthens the usage of the proposed model, both in terms of reduced parameters and improved performance, especially for problems with limited training data.

7. Conclusion

Almost all machine learning problems suffer from the challenge of low inter-class and high intra-class variations. These challenges are further pronounced in cross-domain matching problems such as sketch to digital image matching tasks, due to the added variability in the information content across domains, and the availability of limited labeled training data pairs. This research proposes a novel *Discriminative Shared Transform Learning* algorithm, which utilizes a shared transform to perform effective feature extraction for data belonging to two domains under supervised constraints. Learning a shared transform under supervised constraints enables the model to learn domain invariant features useful for enhanced classification performance. Further, two models have been presented under the proposed algorithm: Contractive and Divergent Models (C-model and D-model). D-model focuses on learning domain invariant representations while modeling the inter-class and intra-class variations across domains. Experimental analysis demonstrates that modeling both inter-class and intra-class constraints improves the performance of existing features in most cases. A unique characteristic of the proposed algorithm is its feature agnostic behavior, i.e. given different input features (raw pixels, hand-crafted, or deep learning based), it results in improved performance for the given task. For example, an improvement of around 4-34% is observed for caricature face recognition. Such techniques provide the flexibility of utilizing task-specific features, to boost the existing performance. A thorough evaluation of the proposed models has been performed on three different case studies classified under sketch to digital image matching: (i) caricature face recognition, (ii) sketch based object retrieval, and (iii) sketch face recognition. The results demonstrate the efficacy of incorporating (i) shared (ii) and discriminative (contractive and divergent) terms in the existing transform learning model, resulting in improved state-of-the-art performance. For example, improvement of around 15% is observed on the challenging IIIT-D forensic sketch dataset. Despite the improved performance achieved by the proposed models, there still exists a vast scope for improvement, especially for real world scenarios.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgements

S. Nagpal is supported by TCS PhD Fellowship. M. Vatsa is supported by SwarnaJayanti Fellowship by the Government of India.

References

[1] B. Klare, Z. Li, A.K. Jain, Matching forensic sketches to mug shot photos, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2010) 639–646.
[2] R. Mauro, M. Kubovy, Caricature and face recognition, Mem. Cognit. 20 (4) (1992) 433–440.
[3] L.A. Pereira, R. da Silva Torres, Semi-supervised transfer subspace for domain adaptation, Pattern Recognit. 75 (2018) 235–249.
[4] B. Yang, A.J. Ma, P.C. Yuen, Learning domain-shared group-sparse representation for unsupervised domain adaptation, Pattern Recognit. 81 (2018) 615–632.
[5] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, W. Fan, Deep cascaded cross–modal correlation learning for fine-grained sketch-based image retrieval, Pattern Recognit. 100 (2020).
[6] S. Ouyang, T. Hospedales, Y. Song, X. Li, C.C. Loy, X. Wang, A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution, Image Vis. Comput. 56 (2016) 28–48.
[7] S. Wang, Z. Ding, Y. Fu, Coupled marginalized auto-encoders for cross-domain multi-view learning, in: International Joint Conferences on Artificial Intelligence, 2016, pp. 2125–2131.
[8] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, A. Majumdar, Face sketch matching via coupled deep transform learning, in: IEEE International Conference on Computer Vision, 2017, pp. 5429–5438.
[9] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, T.M. Hospedales, Deep spatial-semantic attention for fine-grained sketch-based image retrieval., in: International Conference on Computer Vision, 2017, pp. 5552–5561.
[10] A. Dutta, Z. Akata, Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5089–5098.
[11] X. Wu, L. Song, R. He, T. Tan, Coupled deep learning for heterogeneous face recognition, in: AAAI Conference on Artificial Intelligence, 2018.
[12] M. Singh, S. Nagpal, R. Singh, M. Vatsa, A. Noore, Learning a shared transform model for skull to digital face image matching, in: IEEE International Conference on Biometrics Theory, Applications and Systems, 2018.
[13] Y. Cao, C. Wang, L. Zhang, L. Zhang, Edgel index for large-scale sketch-based image search, in: International Conference on Computer Vision and Pattern Recognition, 2011, pp. 761–768.
[14] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa, Sketch-based image retrieval: benchmark and bag-of-features descriptors, IEEE Trans. Vis. Comput. Graph. 17 (11) (2011) 1624–1636.
[15] R. Hu, J. Collomosse, A performance evaluation of gradient field hog descriptor for sketch based image retrieval, Comput. Vis. Image Underst. 117 (7) (2013) 790–806.
[16] Y. Qi, Y. Song, H. Zhang, J. Liu, Sketch-based image retrieval via siamese convolutional neural network, in: IEEE International Conference on Image Processing, 2016, pp. 2460–2464.
[17] Q. Yu, F. Liu, Y. SonG, T. Xiang, T. Hospedales, C.C. Loy, Sketch me that shoe, in: International Conference on Computer Vision and Pattern Recognition, 2016, pp. 799–807.
[18] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, W. Fan, Sketch-based image retrieval with deep visual semantic descriptor, Pattern Recognit. 76 (2018) 537–548.
[19] H. Zhang, P. She, Y. Liu, J. Gan, X. Cao, H. Foroosh, Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval, IEEE Trans. Image Process. 28 (9) (2019) 4486–4499.
[20] X. Zhang, Y. Huang, Q. Zou, Y. Pei, R. Zhang, S. Wang, A hybrid convolutional neural network for sketch recognition, Pattern Recognit. Lett. 130 (2020) 73–82.
[21] Q. Liu, L. Xie, H. Wang, A.L. Yuille, Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval, in: IEEE International Conference on Computer Vision, 2019, pp. 3662–3671.
[22] S. Dey, P. Riba, A. Dutta, J. Llados, Y.-Z. Song, Doodle to search: practical zero-shot sketch-based image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2179–2188.
[23] Y. Fang, W. Deng, J. Du, J. Hu, Identity-aware cyclegan for face photo-sketch synthesis and recognition, Pattern Recognit. 102 (2020).
[24] C. Peng, X. Gao, N. Wang, J. Li, Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation, Pattern Recognit. 84 (2018) 262–272.
[25] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: International Conference on Computer Vision and Pattern Recognition, 2011, pp. 513–520.
[26] H.S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, Memetically optimized MCWLD for matching sketches with digital face images, IEEE Trans. Inf. Forensics Secur. 7 (5) (2012) 1522–1535.
[27] H. Han, B.F. Klare, K. Bonnen, A.K. Jain, Matching composite sketches to face photos: a component-based approach, IEEE Trans. Inf. Forensics Secur. 8 (1) (2012) 191–204.
[28] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967.
[29] D. Huang, Y.F. Wang, Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition, in: IEEE International Conference on Computer Vision, 2013, pp. 2496–2503.
[30] B.F. Klare, A.K. Jain, Heterogeneous face recognition using kernel prototype similarities, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2012) 1410–1422.
[31] S. Ouyang, T.M. Hospedales, Y.-Z. Song, X. Li, ForgetMeNot: memory-aware forensic facial sketch matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5571–5579.
[32] T. de Freitas Pereira, A. Anjos, S. Marcel, Heterogeneous face recognition using domain specific units, IEEE Trans. Inf. Forensics Secur. 14 (7) (2018) 1803–1816.
[33] L. Wolf, Y. Taigman, A. Polyak, Unsupervised creation of parameterized avatars, in: IEEE International Conference on Computer Vision, 2017, pp. 1539–1547.
[34] B.F. Klare, S.S. Bucak, A.K. Jain, T. Akgul, Towards automated caricature recognition, in: IAPR International Conference on Biometrics, 2012, pp. 139–146.
[35] K. Takayama, H. Johan, T. Nishita, Face detection and face recognition of cartoon characters using feature extraction, in: Image, Electronics and Visual Computing Workshop, 2012, p. 48.
[36] Y. Shi, D. Deb, A.K. Jain, WarpGAN: automatic caricature generation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10762–10771.
[37] S. Ravishankar, Y. Bresler, Learning sparsifying transforms, IEEE Trans. Signal Process. 61 (5) (2013) 1072–1086.
[38] S. Nagpal, M. Singh, A. Jain, R. Singh, M. Vatsa, A. Noore, On matching skulls to digital face images: a preliminary approach, in: IEEE International Joint Conference on Biometrics, 2017, pp. 813–819.
[39] S. Ravishankar, Y. Bresler, Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging, SIAM J. Imaging Sci. 8 (4) (2015) 2519–2557.
[40] S. Ravishankar, B. Wen, Y. Bresler, Online sparsifying transform learning 2014; Part I: algorithms, IEEE J. Sel. Top. Signal Process. 9 (4) (2015) 625–636.
[41] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee, T.J. Sejnowski, Dictionary learning algorithms for sparse representation, Neural Comput. 15 (2) (2003) 349–396.
[42] B. Kulis, M. Sustik, I. Dhillon, Learning low-rank kernel matrices, in: International Conference on Machine Learning, 2006, pp. 505–512.
[43] S. Ravishankar, Y. Bresler, Online sparsifying transform learning 2014; Part II: convergence analysis, IEEE J. Sel. Top. Signal Process. 9 (4) (2015) 637–646.
[44] Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Proceedings of Asilomar Conference on Signals, Systems and Computers, vol. 1, 1993, pp. 40–44.
[45] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2006, pp. 1735–1742.
[46] A. Mishra, S.N. Rai, A. Mishra, C.V. Jawahar, IIIT-CFW: a benchmark database of cartoon faces in the wild, in: European Conference on Computer Vision Workshops, 2016, pp. 35–47.
[47] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967.
[48] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: International Conference on Computer Vision and Pattern Recognition, 2011, pp. 513–520.
[49] L. Wang, V. Sindagi, V.M. Patel, High-quality facial photo-sketch synthesis using multi-adversarial networks, in: IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 83–90.
[50] P.J. Phillips, Hyeonjoon Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[51] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[52] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[53] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proceedings of the British Machine Vision Conference, 2015, pp. 41.1–41.12.
[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: International Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[56] Verilook Face SDK, (http://www.neurotechnology.com/verilook.html).
[57] A. Wang, J. Cai, J. Lu, T.J. Cham, MMSS: multi-modal sharable and specific feature learning for RGB-D object recognition, in: IEEE International Conference on Computer Vision, 2015, pp. 1125–1133.
[58] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1089–1102.
[59] W. Zheng, L. Yan, C. Gou, W. Zhang, F. Wang, A relation network embedded with prior features for few-shot caricature recognition, in: IEEE International Conference on Multimedia and Expo, 2019, pp. 1510–1515.
[60] J. Song, Y.-Z. Song, T. Xiang, T. Hospedales, X. Ruan, Deep multi-task attribute–driven ranking for fine-grained sketch-based image retrieval, in: Proceedings of the British Machine Vision Conference, 2016, pp. 132.1–132.11.

[61] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, T.M. Hospedales, Deep spatial-semantic attention for fine-grained sketchbased image retrieval, in: International Conference on Computer Vision, 2017, pp. 5551–5560.

[62] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[63] Z. Yi, H.R. Zhang, P. Tan, M. Gong, DualGAN: unsupervised dual learning for image-to-image translation, in: IEEE International Conference on Computer Vision, 2017, pp. 2868–2876.

[64] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 2242–2251.

[65] D. Mandal, S. Biswas, Generalized coupled dictionary learning approach with applications to cross-modal matching, IEEE Trans. Image Process. 25 (8) (2016) 3826–3837.

[66] S. Wang, L. Zhang, Y. Liang, Q. Pan, Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2216–2223.

[67] D. Xu, J. Song, X. Alameda-Pineda, E. Ricci, N. Sebe, Multi-paced dictionary learning for cross-domain retrieval and recognition, in: International Conference on Pattern Recognition, 2016, pp. 3228–3233.

[68] J. Jiang, Y. Yu, Z. Wang, X. Liu, J. Ma, Graph-regularized locality-constrained joint dictionary and residual learning for face sketch synthesis, IEEE Trans. Image Process. 28 (2) (2019) 628–641.

[69] C. Peng, N. Wang, J. Li, X. Gao, Universal face photo-sketch style transfer via multiview domain translation, IEEE Trans. Image Process. 29 (2020) 8519–8534.

Shruti Nagpal received Bachelors of Technology in Computer Science from IIIT Delhi, India in 2015 and currently, she is pursuing her doctoral studies since 2015 at IIIT Delhi, India. Her area of research includes machine learning, deep learning, and biometrics. She was awarded the Google Women Techmakers scholarship in 2018 and she is also a recipient of the TCS PhD Fellowship. Her research received the Best Poster Awards at IEEE/IAPR International Joint Conference on Biometrics (IJCB) 2017 and 2020. She has authored over 20 publications including journals and peer-reviewed conferences. She has also served as the reviewer for Pattern Recognition, Information Fusion, IEEE Transactions on Image Processing, and IEEE Access. She is an Associate Editor of the IEEE Biometrics Council Newsletter.

Maneet Singh received Bachelors of Technology in Computer Science from IIIT Delhi, India, in 2015 and currently, she is pursuing her doctoral studies since 2015 at IIIT Delhi, India. Her research interests focus machine learning and deep learning with applications in face recognition. Her research received the Best Poster Awards in IEEE/IAPR International Joint Conference on Biometrics 2017 and 2020. She is also a recipient of the Google Anita Borg Memorial Scholarship, 2015. She has authored over 20 publications including journals and peer-reviewed conferences. She has also served as a reviewer for several journals such as Pattern Recognition, Information Fusion, Computer Vision and Image Understanding, and IEEE Transactions on Information Forensics and Security.

Richa Singh received the Ph.D. degree in computer science from West Virginia University, Morgantown, USA, in 2008. She is currently the Professor at IIT Jodhpur, India, Vice President (Publications) of IEEE Biometrics Council, and an Associate Editor-in-Chief of Pattern Recognition. She has co-edited three books including Deep Learning in Biometrics. Her areas of interest are pattern recognition, machine learning, and biometrics. She was a recipient of the Kusum and Mohandas Pai Faculty Research Fellowship at the IIIT-Delhi, the FAST Award by the Department of Science and Technology, India, and several best paper and best poster awards in international conferences. She has also served as the Program Co-Chair of IJCB2020, AFGR 2019, BTAS 2016, and IWBF 2018, and the General Co-Chair of ISBA 2017. She is currently serving as the General Co-Chair of FG2021 and Program Co-Chair of ICMI2022. She is a Fellow of IEEE and IAPR and a Senior Member of ACM.

Mayank Vatsa received the M.S. and Ph.D. degrees in computer science from West Virginia University, USA, in 2005 and 2008, respectively. He is currently a Professor at IIT Jodhpur, India, and the Project Director of the TIH on Computer Vision and Augmented and Virtual Reality. His areas of interest are biometrics, machine learning, computer vision, and information fusion. He is the recipient of the prestigious Swarnajayanti Fellowship award from Government of India, A. R. Krishnaswamy Faculty Research Fellowship at the IIIT-Delhi, the FAST Award Project by DST, India, and several Best Paper and Best Poster Awards at international conferences. He is an Area/Associate Editor of Information Fusion and Pattern Recognition Journals, General Co-Chair of IJCB 2020, and the PC Co-Chair of the ICB 2013, IJCB 2014, ISBA2017, and FG2021. He has served as the Vice President (Publications) of the Biometrics Council. He is a Senior Member of ACM and IEEE.