DOI: 10.21917/ijivp.2025.0512

JOINT FUSION CROSS SPECTRAL ASSOCIATIVE DEEP LEARNING MODEL FOR FACE RECOGNITION

Anita Sigamani[1] and Prema Selvaraj[2]
[1]*Department of Computer Science, B.M.S College for Women, India*
[2]*Department Computer Applications, Arulmigu Arthanareeswarar Arts and Science College, India*

Abstract

Several models have been previously developed for learning correlated representations between source and target modalities. In this paper, we present a novel Joint Fusion model for learning cross spectral image representation for heterogenous face recognition. The coupled receptive face recognition model is built using ResNet architecture as backbone, a fully connected neural network and triple auto encoders for learning perceptible feature points invariant to changes in Spectrum. The performance of this model is tested using CelebA and LFW datasets. Moreover, the empirical results show that the learnt Common Latent Embeddings by the integrated networks produce competitive cross-spectrum face recognition results. These results are obtained by training the model using Adam Optimizer and Mean Squared Error (MSE) loss function. The proposed model has shown a performance improvement of 20% in AUC (Area Under the Curve) measure than the State-of-the-art with Polarization State Information, and 23% improvement in AUC over the State-of-the art models in traditional Thermal-to-Visible synthesis process. As well 12% improvement in EER (Equal Error Rate) measure polar measure and 9% improvement in EER (Conventional) are observed while comparing with Sate-of-the-art Models in Traditional thermal case.

Keywords:

Flexible Filter, Bi-HCRV, FCNN, Face Detection, Recognition

1. INTRODUCTION

Face detection has evolved significantly over the years, transitioning from early feature-based techniques to sophisticated deep learning models. Initially, methods like edge detection, the Viola-Jones algorithm [1], and Histogram of Oriented Gradients (HOG) [2] [3] were employed to detect facial features based on geometric and structural patterns. However, the advent of deep learning revolutionized this field, allowing models to automatically learn complex feature hierarchies from vast datasets, leading to a substantial increase in accuracy and robustness. The performance of face detection models is often influenced by various factors, including image and video quality, noise, and environmental conditions. In forensic analysis, in real-time scenarios such as low lighting or motion blur present critical challenges that can severely affect the accuracy of face detection. To address these challenges, image processing filters play a crucial role. These filters, which can range from simple noise reduction techniques to more advanced image restoration methods, are crucial in enhancing image quality and improving the performance of face detection algorithms. Filter operations in image processing are widely used across multiple applications, including denoising, edge detection, and feature enhancement. A comparative analysis of these filters, such as Gaussian, Median, and Bilateral filters, reveals their impact on the accuracy of face detection models, particularly in noisy environments. For instance, Dlib's HOG-based model has shown varying levels of success depending on the filter applied to the input data. However, even the most advanced face detection models are prone to failure due to certain factors, including occlusion, pose variation, or poor lighting conditions. Understanding these common reasons for failure is critical for developing more robust systems. One approach to enhance accuracy is through the automated UI for detecting human faces from noisy inputs, ensuring better real-time performance in challenging conditions. In anatomical studies, the extraction of facial landmarks plays an important role in aligning skull images with facial photographs. Techniques like Procrustes Analysis, particularly the Generalized Procrustes Analysis, are employed to align anatomical points between different images, reducing discrepancies. This alignment, often quantified through metrics such as Root Mean Square Deviation (RMSD) [4], ensures accurate positioning of anatomical points, which is vital in forensic applications. This paper explores the evolution of face detection techniques, analyzes the performance of various filters in noisy conditions, and discusses the challenges faced in real-time and forensic scenarios. Additionally, we delve into methods like Procrustes Analysis for anatomical landmark alignment, offering insights into improving face detection accuracy across diverse applications.

Face detection has evolved significantly over the years, transitioning from early feature-based techniques to sophisticated deep learning models. Initially, methods like edge detection, the Viola-Jones algorithm [21], and Histogram of Oriented Gradients (HOG) [22] [23] were employed to detect facial features based on geometric and structural patterns. However, the advent of deep learning revolutionized this field, allowing models to automatically learn complex feature hierarchies from vast datasets, leading to a substantial increase in accuracy and robustness. The performance of face detection models is often influenced by various factors, including image and video quality, noise, and environmental conditions. In forensic analysis, images with low lighting, motion blur and occlusion raise severe challenges in face detection. To address these challenges, image processing filters play a crucial role.

Significance of Infrared modality-based face recognition has increased enormously since 2010 [22]-[24]. Thermal images are ideal for working with images captured at low light situation and at night times. To overcome Occlusions issues in face detection and recognition instead of working with visible images, in this paper we choose to work with a collection of thermal (HotCold) images to capture the face structure, shape and size. Short et al. [5] used Polarimetric imaging to improve the performance of Cross Modal Face Recognition.

In the study of traditional heterogeneous face recognition system, images from different modalities are considered, like Visible, Infrared, Short-wave Infrared, Mid-wave Infrared, Long-wave Infrared, Or Forensic Sketches [6]-[9]. Image synthesis

requires precise image representation and a well framed fusion rule to influence the accuracy and performance of the system. Image fusion algorithms [10]-[15] have emerged in the recent years are developed. But these algorithms are not suitable to increase the spatial resolution in the context of image fusion process. To generate better results in image fusion process, better quality images have to be passed as input to the model. Yin et al. [16] used parallel image fusion and super resolution using sparse representation. J. Zhong et al. [17] performed Image Fusion and Super Resolution using Convolutional Neural Network. Li et al. [18] used Fractional Differential and Variational method for Image fusion and Super-resolution. Though these methods are used for parallel fusion and super resolution, still there are more chances to improve the image resolution. To improve Fusion results at the same time to preserve image details and clarity we propose a joint fusion model with contrastive learning to extract features from multispectral inputs using triple auto encoder and decoder framework. Many state-of-the-Art methods have shown significant performance in face recognition systems in the past decades [19]-[21], seems to be too complex for integration with real time systems. A solution to this, in this paper we propose an Idiosyncratic Neural network model to learn a common Latent Embeddings from Visible, Thermal and Synthesised images for training a classifier to correctly identify matching faces in the dataset.

2. RELATED WORKS

Over the decades, the results of super resolution has shown improved performance on Image Fusion [24], Pattern Recognition and Classi cation [25]. Traditional Image Fusion methods based on Super Resolution treats the input Image as a an individual component and uses only one dictionary to represent the image [26]. Representing multiple features of an image in a single dictionary becomes difficult and hard to characterize the complex structure of the image [27]. As a solution Image fusion based on Multicomponent analysis has emerged.

Jiang and Wang [28] in their work to represent complex structures used curvelet and discrete cosine transform (dct) dictionaries. Liu et. al. [29] has proposed a fusion algorithm using curvelet and local DCT. The dictionaries used in [29] are just analytical dictionaries which have weak adaptability nature and are not suitable to learn complex patterns and structures of images. As a solution, Li et al. [30] using Low-rank and Sparse decomposition developed a new image fusion method.

We design a novel model for image synthesis using a Fully Connected Neural Network (FCNN) MODEL, by using Joint Fusion technique with Adam Optimizer to create Joint Latent Embedding and Mean Squared Error (MSE) loss function to minimize the loss among Latent Embeddings.

3. CONTRIBUTIONS AND OUTLINE

In this work we make three major contributions:

- We propose a Flexible filter to increase face detection accuracy.
- We develop a novel framework for cross spectrum face recognition using receptive face recognition ResNet model.
- We design a Fusion rule for Multi-spectral Image Synthesis.
- We introduce a Joint fusion with contrastive learning.
- We present a novel network for coupling features extracted into a Common Latent Embeddings using FCNN and Triple Auto Encoder-Decoder Framework.
- We evaluate the system using CelebA [34] and LFW [35] for cross spectrum Face Recognition.

4. PROPOSED IMAGE FILTER

The proposed flexibility filter gives us a finer control over the filtering process by adjusting how aggressively noise is reduced and how well edges are preserved. By modifying the parameter cp, we can control the level of noise reduction. Smaller values of ‘cp’ (closer to 1) make the filter behave like a median filter, effectively preserving edges while removing noise. Larger values of ‘cp’ result in more smoothing, reducing more noise but potentially blurring the edges. This allows the filter to maintain sharp edges by averaging fewer central pixels and reduce the computation time taken. The results obtained are shown in Fig.1. The flexible_filter (FF) is introduced for preprocessing images efficiently by reducing noise and preserving edges.

4.1 NOVEL CROSS SPECTRUM FACE RECOGNITION FRAMEWORK

A novel framework for overcoming image occlusion by using Cross Spectrum input is proposed. Instead of working with single image based on visible spectrum, we choose to work with Bi-HCRV (2- hot, 2-cold and 1-visible) images of same identity from different modalities.

4.2 MULTI-SPECTRAL IMAGE SYNTHESIS

A new image fusing rule is proposed for the fusing the images using image pattern and size. The fusion rule is designed based on spatial resolution of the initial input Bi-HCRV images, and Weighted averaging scheme is used to fuse the pixel values based on size and pattern. The adaptive weights of each image are calculated dynamically using Eq.(5) and Eq.(7). The weights are calculated using two methods, such as Intensity based weighting and pattern-based weighting.

To synthesize Bi-HCRV images into a single composite image, as a pre-processing step the input images are converted to gray scale image using Eq.(2) based on the Coefficients ITU-R BT.601 standard based Coefficients are used to define the intensity of each Channel.

$$\begin{aligned} &C_i^n\ (i=1,2,\ldots,n) \equiv G_i^n\ (i=1,2,\ldots,n),\\ &h_i^n\ (i=1,2,\ldots,n) \equiv G_i^n\ (i=1,2,\ldots,n),\\ &V_{(i=1)} \equiv G_{(i=1)} \end{aligned} \quad (1)$$

$$G_i^n(x,y) = 0.2989 \cdot R(x,y) + 0.5870 \cdot G(x,y) + 0.1140 \cdot B(x,y) \quad (2)$$

The images are normalized between the range (0, 1) as a prior step to adaptive weight calculation using Eq.(3). We do this in order to prevent Normalize pixel values to avoid domination of a certain image type on other types. Finally, each Normalized image is multiplied with calculated average weighting values and we

combine the normalized images with custom weighting using Eq.(10).

$$\text{Norm}_i(x,y) = \frac{\text{Org}_i(x,y) - \text{Min}_i(x,y)}{\text{Max}_i(x,y) - \text{Min}_i(x,y)} \quad (3)$$

The size-based weights are calculated using Eq.(4) and Eq.(5) based on the total intensity of each image.

$$S_i = \sum_{x=1}^{M_i}\sum_{y=1}^{N_i} I_i(x,y) \quad (4)$$

$$\text{SW}_i = \frac{S_i}{\sum_{j=1}^{N} S_j} \quad (5)$$

The pattern-based weights are calculated using Eq.(6) and Eq.(7) by detecting the edges using Canny edge detection algorithm and by computing the sum of those pixels found in the detected edges.

$$E_i = \sum_{x=1}^{M_i}\sum_{y=1}^{N_i} C_i(x,y) \quad (6)$$

$$\text{PW}_i = \frac{E_i}{\sum_{j=1}^{N} E_j} \quad (7)$$

The computed Weights are normalized as in Eq.(8), before computing the total adaptive weight value using Eq.(10) or Eq.(11).

$$\sum_{i=1}^{N} W_i = 1 \quad (8)$$

$$\text{TW}_i = \alpha_s \cdot \text{SW}_i + \alpha_p \cdot \text{PW}_i \quad (9)$$

where α_s and α_p are the two Scaling Factors used in the final image synthesis equation.

$$\text{Synth}_i(x,y) = \sum_{i=1}^{n} \text{TW}_i \cdot I_i(x,y) \quad (10)$$

or

$$I_{\text{Synth}}(x,y) = \sum_{i=1}^{n} \left(W_{\text{size},i} + W_{\text{pattern},i}\right) \cdot I_i(x,y) \quad (11)$$

4.3 FEATURE EXTRACTION

Down the pipeline of Cross spectrum face recognition we extract face features, using a pre-trained Network ResNet18 model. The synthesised image, a hot image and a real-visible image are passed as sequential inputs to ResNet18 as shown in Fig.1. Features are extracted from three different modalities thermal image, visible image and synthesised image by passing through Resnet18 model by removing the final classification layer and using it as a feature extractor. The features are first extracted from the global region of the detected face using Max Margin Object Detection using Fully connected convolutional neural network and local fiduciary regions of face (eyes, nose and mouth) using ResNet18 architecture given in Fig.2.

4.4 JOINT FUSION WITH CONTRASTIVE LEARNING

A Naïve Joint Fusion Model, given in Fig.4 is proposed to extract features from Multispectral inputs (Real-visible, Synthesised, Thermal) and combine them into single latent space using Triple auto encoder and decoder framework given in Fig.3. For extracting high dimensional feature vectors, pre-trained ResNet18 model is used by removing the classification layer to retain the feature maps from second last layer of the model. The obtained feature maps are passed through a FCNN model, which uses Joint Fusion technique and contrastive learning to create Joint Latent Embedding. The training pipeline uses Adam Optimizer and Mean Squared Error (MSE) loss function to minimize the loss among Latent Embeddings and a Zero tensor. The training loop is iterated for 10 Epochs, and the loss values are monitored for each epoch.

4.4.1 Feature Extraction:

For each spectrum $s \in \{V_i, S_i, H_i\}$, the feature extractor used maps the input image of X_i to Feature vector F_i: $F_i = F_{\text{EXT}}(X_i)$. F_i is the input image of spectrum s as in Eq.(12). $F_i \in \mathbb{R}^{512}$ is the extracted feature vector. Resnet18, pretrained on ImageNet dataset is used to extract feature vectors from cross spectrum image input by removing the last classification layer.

$$F_i = f_{\text{ResNet18}}(X_i),\ s \in \{V_i, S_i, H_i\} \quad (12)$$

4.4.2 Feature Fusion:

The proposed Joint Fusion Model is used to couple the feature-vectors of three spectrums into a single latent vector as in Eq.(13) and Eq.(14):

$$F_{\text{Combined}} = \sum\left(F_{V_i}, F_{S_i}, F_{H_i}\right) \quad (13)$$

$$F_{\text{Combined}} \in \mathbb{R}^{3\times 512} = \mathbb{R}^{1536} \quad (14)$$

The Joint Fusion vector obtained is passed through three Fully Connected Layers of the model for processing using Eq.(15) - Eq.(17).

$$\text{The First layer, } F_{C_1} = \sigma(W_1 F_{\text{Combined}} + b_1) \quad (15)$$

$W_1 \in \mathbb{R}^{512\times 1536}$ (weights), $b_1 \in \mathbb{R}^{512}$ (biases), $\sigma(x) = \max(0, x)$ refers to ReLU activation, $F_{C_1} \in \mathbb{R}^{512}$.

$$\text{The Second layer, } F_{C_2} = \sigma(W_2 F_{C_1} + b_2) \quad (16)$$

$W_2 \in \mathbb{R}^{256\times 512}$ (weights), $b_2 \in \mathbb{R}^{256}$ (biases), $F_{C_2} \in \mathbb{R}^{256}$.

$$\text{The Third layer, } F_{C_3} = W_3 F_{C_2} + b_3 \quad (17)$$

$W_3 \in \mathbb{R}^{128\times 256}$ (weights), $b_3 \in \mathbb{R}^{128}$ (biases), $F_{C_3} \in \mathbb{R}^{128}$, is the Joint Latent Embedding $\left(J_{\text{Embedding}}\right)$.

4.4.3 Loss Function:

MSE loss function is used to compute the loss between Joint Latent Embedding $\left(J_{\text{Embedding}}\right)$ and a zero tensor (0) of equal size using Eq.(18) or Eq.(19).

$$\text{MSE}_{\text{loss}} = \frac{1}{128}\sum_{j=1}^{128}\left(J_{\text{Embedding},j} - 0\right)^2 \quad (18)$$

or

$$\text{MSE}_{\text{loss}} = \frac{1}{128}\sum_{j=1}^{128} J^2_{\text{Embedding},j} \quad (19)$$

where F_{Ca_j} is the j^{th} component of Joint Latent Embedding F_{Ca}.

The Contrastive loss is calculated using Eq.(20) to verify images with same identity have relatively similar embeddings and images with different identities have dissimilar identities. An Anchor image (from any modality), a Positive image (another image same as Anchor image from different modality) and a negative image (completely different image) are considered during training. The Euclidean distance is calculated between Anchor image and Positive image pairs, and Anchor image and Negative image pairs. The Contrastive loss function is defined as:

$$Loss=\text{Max}(0,D_{POS}-D_{POS},Margin) \quad (20)$$

The main goal of contrastive loss is to group images with similar feature vector pairs in a first row and dissimilar feature vector pairs in second row resembling a matrix structure in the feature space. The detailed loss function for contrastive loss is defined as:

$$L_{\text{contrastive}} = \frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0, D(a_i, p_i) - D(a_i, n_i) + \text{Margin}\right)\right] \quad (21)$$

The positive distance measure is used to check the relative closeness measure of Anchor and Positive images using Eq.(22). The negative distance measure is used to check the relative distance (far) measure of Anchor and Positive images. The margin parameter defines the minimum essential difference between Positive and Negative pairs. When the distance between Positive pair > Negative pair, then the Loss is counted to be Negative. The distance between the Anchor and Positive and Anchor and Negative is defined by,

$$D(I_1, I_2) = \| I_1 - I_2 \|_2 = \sqrt{\sum_{j=1}^{d}\left(I_{1,j} - I_{2,j}\right)^2} \quad (22)$$

where d represents the feature vector dimensions (128), $I_{1,j}$ and $I_{2,j}$ are the components of the vectors I_1 and I_2. Thus, the final contrastive Loss yields the following objective function of the Joint Fusion Model with contrastive learning as shown in Eq.(23).

$$L_{\text{contrastive}} = \frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0, \| a_i - p_i \|_2 - \| a_i - n_i \|_2 + \text{Margin}\right)\right] \quad (23)$$

We are interested in minimizing the Coupling error over training the images (X, Y and Z) from cross-spectrum. The coupling error is defined as the variance between the Real visible F_{V_i} and thermal latent features F_{H_i} in Eq.(24) and Real visible and Synthesis latent features F_{S_i} in Eq.(25).

$$\sigma(F_{V_i}, F_{H_i}) = \sum_i \| F_{V_i} - F_{H_i} \|^2 \quad (24)$$

$$\sigma(F_{V_i}, F_{S_i}) = \sum_i \| F_{V_i} - F_{S_i} \|^2 \quad (25)$$

Minimizing the coupling error enables sufficient learning of latent vectors from V_i, S_i, T_i and forces the features to be much similar. To have a smooth coupling process and to prevent the model from over fitting, [31] [47] in the Joint Fusion model, we include Kullback-Leibler (KL) divergence to align the latent vector Embeddings with the Input requirements of classification and prediction Model. For the downstream Classification task, we use a Face Predictor model based on Ensemble of Regression Trees. To align the latent embeddings with Face Predictor model (target distribution), Multivariate Gaussian Distribution is used with a covariance matrix $\sum$ to capture the correlations between face features.

4.4.4 Definition:

Let us consider the Latent Embeddings as samples from $\mathcal{N}(\mu, \Sigma)$ to confine the embeddings to probable structure to reduce overfitting and to enable smoother Geometric relations. μ refers to the Mean value of face landmark and the Covariance Matrix $\sum$ found using the face landmarks. landmarks. To regularize the KL Divergence $\mathcal{N}(\mu, \Sigma)$, we align the Embedding Distribution, $q(z)$ with $\mathcal{N}(\mu, \Sigma)$ using Eq.(26).

$$d_{\text{KL}}(q(z) \| \mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\begin{pmatrix} \text{Tr}(\Sigma^{-1}\Sigma_q) \\ +(\mu_q - \mu)^T \Sigma^{-1}(\mu_q - \mu) \\ -k \\ +\log\frac{|\Sigma|}{|\Sigma_q|} \end{pmatrix} \quad (26)$$

4.5 SHAPE PREDICTOR MODEL

In order to compare the query image with the collection of images found in dataset, as a prior step for face recognition we pass the extracted features to the Shape predictor model. The model's task is defined as a two-step equation process involving the computation of facial landmarks and the visualization of landmarks.

The function of shape predictor model to compute 128-embedding vectors e_j based on the face landmarks of all detected faces $e_j \in F$ and visualize landmarks by overlaying face landmarks $\{(P_K)\}$ over images using Eq.(27) and (28).

$$\varepsilon(I) = \{e_j : e_j = F_{\text{Rec}}(I, F_{\text{Shape}}(I, F_{\text{Detect}}(I))) \quad \forall j \in F\} \quad (27)$$

$$L(I) = D\left(I, \{P_K : P_K = F_{\text{Shape}}(I, F_{\text{Detect}}(I)) \ \forall K \in [1,68]\}\right) \quad (28)$$

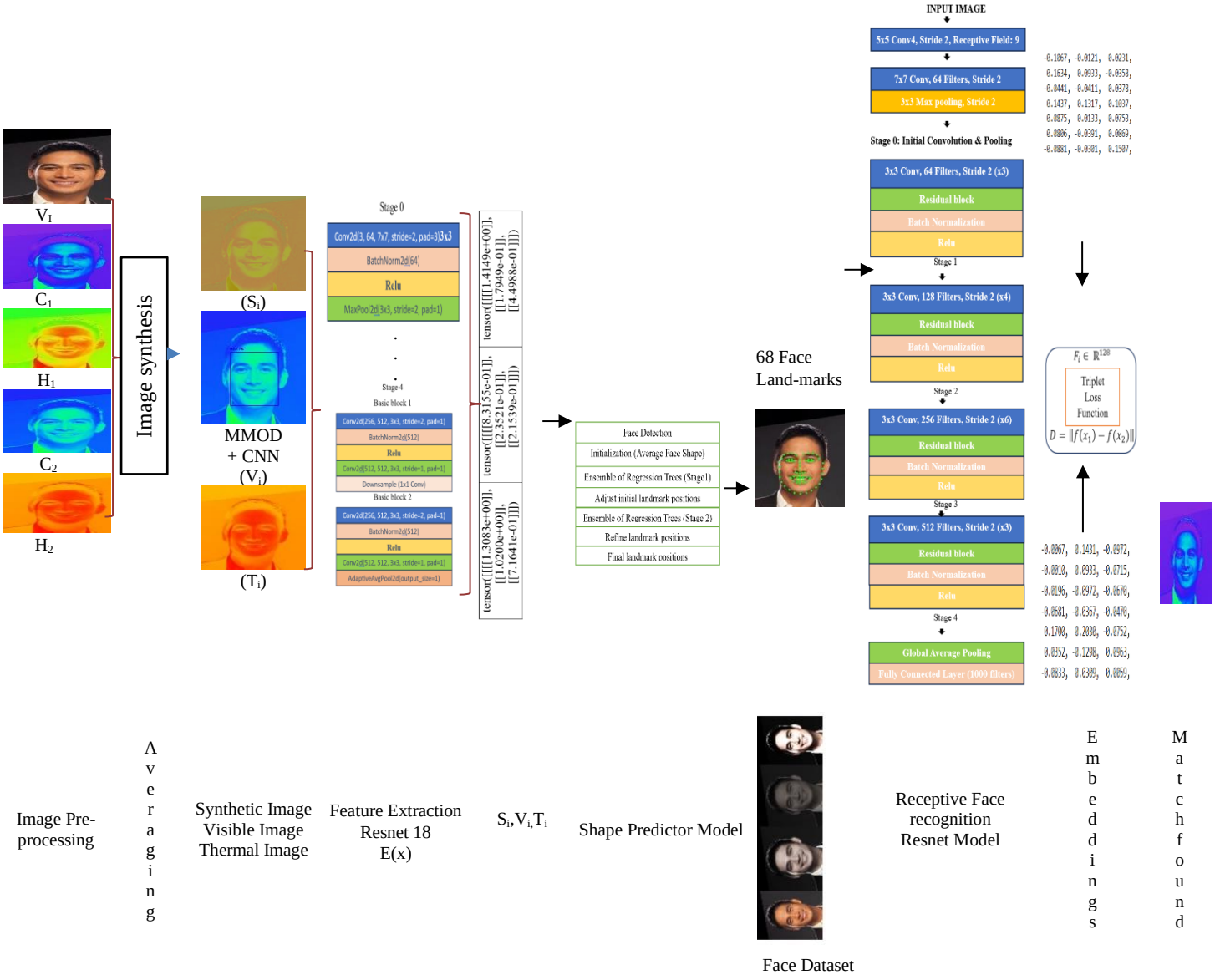

Fig.1. Cross Spectrum Face Recognition

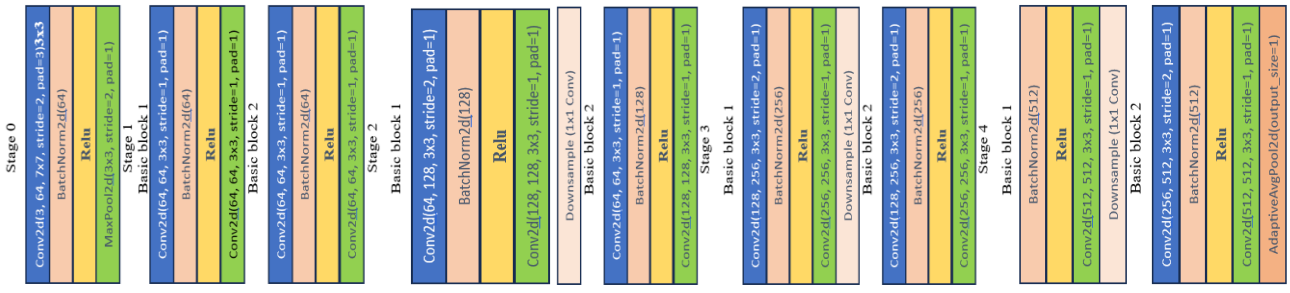

Fig.2. ResNet18 Architecture

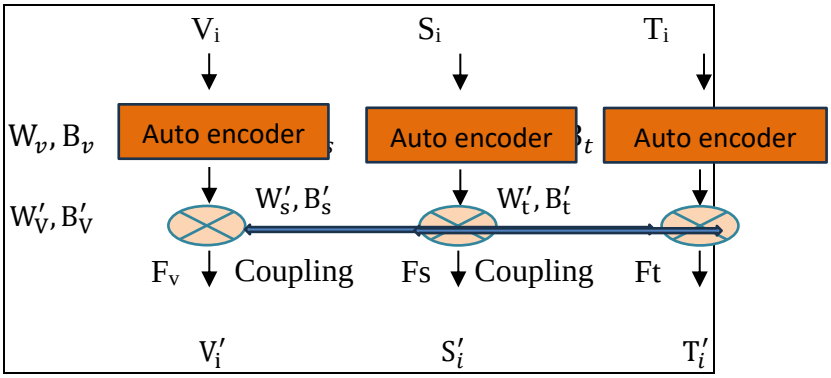

Fig.3. Coupling Embeddings into common Latent space using triple auto encoder and decoder framework

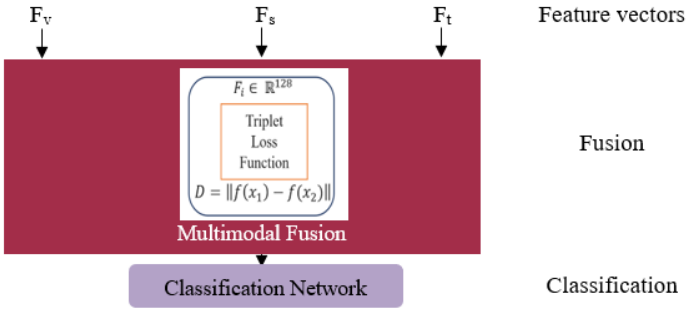

Fig.4. Multimodal Fusion

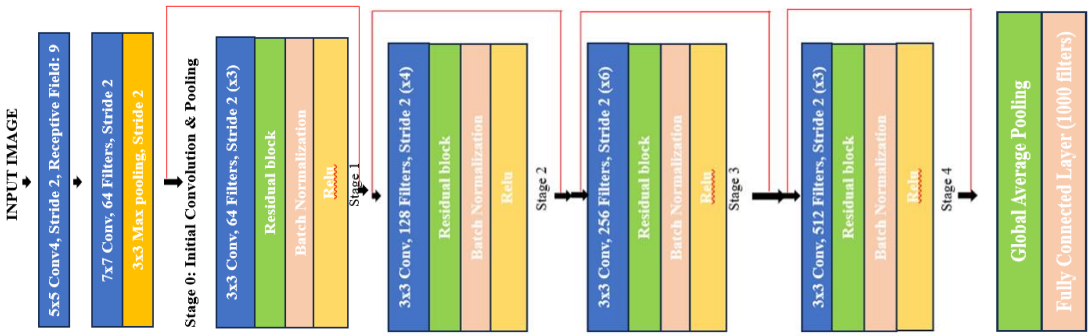

Fig.5. Receptive Face recognition Resnet Model

4.6 TRIPLE AUTO ENCODER AND DECODER FRAMEWORK

A Combined autoencoder is used for cross spectrum face recognition as shown in Fig.3. The primary objective of this work is to extract common feature vectors from three domains (Real-Visible, Synthesised image and thermal image) and to train the classifier using one modality to generalize with other modalities to increase face recognition. A naïve fusion model is proposed using Joint Fusion with Contrastive Learning. To combine features from different modalities (e.g., visible and thermal images) into a single latent space for face recognition, we use the framework of Triple auto encoder and decoder from Fig.3. The latent features of V_i, S_i, and T_i are computed from the Bi-HCRV images. The encodings of Visible image W_V',B_V' is computed from feature vectors of F_v and the encoder parameters W_v,B_v. The encodings of Synthesised image W_s, B_s, is computed from feature vectors of Fs and the encoder parameters W_s',B_s'. The encodings of Thermal image W_t, B_t, is computed from feature vectors of Ft and the encoder parameters W_t',B_t'. The computed latent vectors are coupled using the proposed naïve Joint Fusion Model given in Fig.4.

The Fusion Model in Fig.4, is a FCNN consisting of three fully connected layers. The first Fully Connected Linear Layer takes features from three different spectrum inputs and outputs 512-dimensional features. The second Fully Connected Layer takes 512-dimensional feature from previous layer and outputs a 256-dimensional feature. The third Fully Connected Layer takes 256-dimensional feature from previous layer and outputs a latent space dimension of 128-dimensional feature. The method takes three inputs from Real-Visible features, Thermal features and Synthesised features. The features are coupled using Joint fusion technique with Contrastive learning resulting as a single feature vector consisting information from all three spectrums. The

combined feature vectors are passed through First Fully Connected layer (FC1) and towards a ReLU Activation Function. The resulting output of First layer is passed through Second Fully Connected layer (FC2) and towards a ReLU Activation Function. The output from the second layer is passed through Third Fully Connected layer (FC3). The output generated from FCNN results as a 128-latent space dimensional feature vector with combined information from the three input modalities.

4.7 RECEPTIVE FACE RECOGNITION RESNET MODEL

Cross-spectrum face recognition system works on two different modalities of images such as real-visible image and hot-cold thermal images. The aim of the system is to map the feature vectors of images chosen from different modalities into common latent space to enable face recognition. In the proposed architecture given in Fig.6, we choose to work with 3x3 kernels at convolutional layers for capturing local features and 5x5 kernel to capture more spatial information. To capture fine grained details Stride 1 is used and to reduce the computation load and to increase receptive field Stride 2 is used. A convolutional neural network is defined with receptive field to capture local features such as edges and textures which are vital for detecting local features like eyes, nose and mouth.

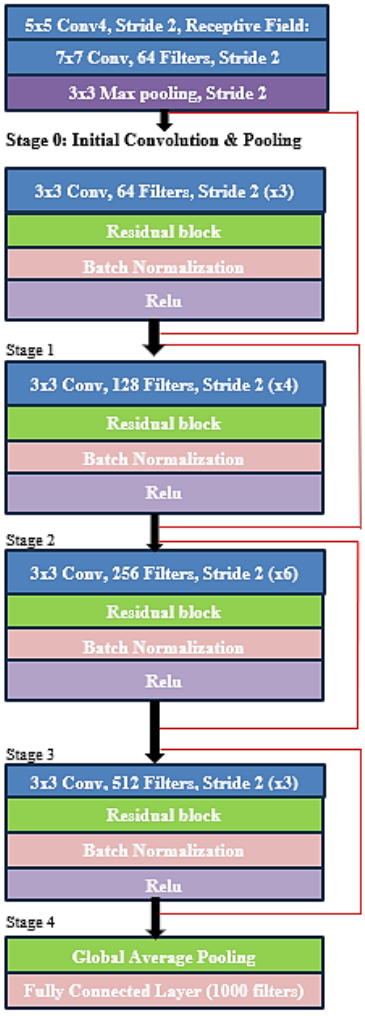

Fig.6. Receptive Face recognition Resnet Model

$$RF = \text{KernelSize} + (\text{KernelSize} - 1) \times (\text{Stride} - 1) \quad (29)$$

With the growing receptive field value proportional to the kernel size and stride of the convolution, global features like face shape and structure details are captured. Receptive field calculation is computed as in Eq.(29). We achieve down sampling by increasing the stride instead of using normal pooling operation like CNNs do.

$$\text{Match}(I_1, I_2) = \begin{cases} \text{true}, & \text{if } \| \varepsilon(I_1) - \varepsilon(I_2) \|_2 < \text{threshold} \\ \text{false}, & \text{otherwise} \end{cases} \quad (30)$$

$\text{Match}(I_1, I_2)$ is computed using Eq.(30) based on the Euclidean Distance between I_1, I_2 returns True if a match is found (distance is below the sigma (σ) value) distance between embeddings is below the threshold and returns False if a match is not found (distance above the sigma (σ) value). We set threshold of σ 0.6 for conforming on the identities.

4.8 FACE RECOGNITION DATASETS

In order to assess the performance of the proposed model, we use two different face recognition datasets: CelebA [34] and LFW [35] for cross spectrum Face Recognition. CelebA, "CelebFaces Attributes Dataset" is a Large-Scale face Attributes dataset. It has 202,599 celebrity images with forty attributes including pose variations and cluttered background. LFW, "Labeled Faces in the Wild" has more than 13,000 images of faces collected from web resource. Each face in the dataset is labelled with person name in the picture. The dataset has 1680 images with redundant photos of same person. Considering managing runtime and processing speed efficiently, a total of 530 images are selected from CelebA dataset and 500 images from LFW. We use these datasets for Face Synthesis, Face Detection, Landmark Localization and Face Recognition Tasks as shown in Fig.8.

To quantify the difference between Feature Distributions of two discrete Probability Distributions (synthesised image with each image of the CelebA dataset) we use Eq.(31) iteratively across all features for each image in the CelebA dataset. For $p = [p_1, p_2, \ldots, p_n]$ and $q = [q_1, q_2, \ldots, q_n]$

$$d_{\text{KL}}(p \,\|\, q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \quad (31)$$

To find the variance between the two probability measures of source image (Thermal/synthetic image) and visible image, the following Eq.(32) is used. Compute $\log \frac{p_i}{q_i}$ for each image (i) in the dataset. Multiply each image (i) by p_i and sum over all images.

$$D = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} - \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j} \quad (32)$$

5. RESULTS AND DISCUSSION

We evaluate the performance of cross-spectrum Receptive Face Recognition Resnet Model using CelebA [34] and LFW [35]. A total of 530 images are selected from CelebA dataset and 500 images from LFW for testing and training the model. For the selected images we define a Square shaped bounding boxes for the global region, a Circle for left and right eye region and a

Triangle for nose and mouth region as in Fig.7. We compare the results of the image fusion and face recognition accuracy with State-of-the-art Cross-Spectrum synthesized imagery using: Performance Test of the Model, Landmark detection and Face recognition accuracy.

5.1 PERFORMANCE TEST OF THE MODEL

For matching the synthesized Bi-HCRV with the visible images in the dataset we use AUC (Area Under the Curve) and EER (Equal Error Rate) measures from the ROC (Receiver Operating Characteristic) curves. To perform an impartial comparison, the matching process follows the Tightly Cropped Regions. we perform matching using the same tightly cropped regions as in [31] [32].

Table.1. Performance Test comparison between Baseline methods, State-of-the-art methods, and our proposed CSFS-RFRR Model

Method	**AUC (polar)**	**AUC (S0)**	**EER (polar)**	**EER (S0)**
RAW	50.35%	58.64%	48.96%	43.96%
Mahendran et al. [33]	58.35%	58.64%	48.96%	43.96%
Zhang et al. [32]	79.90%	79.30%	25.17%	27.34%
Riggan et.al. [31]	75.83%	68.52%	33.20%	34.36%
CSFS-RFRR Model (ours)	**95.43%**	**92.49%**	**19.46%**	**23.25%**

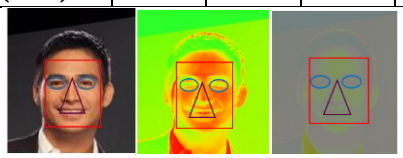

Fig.7. Global region and local regions of face image

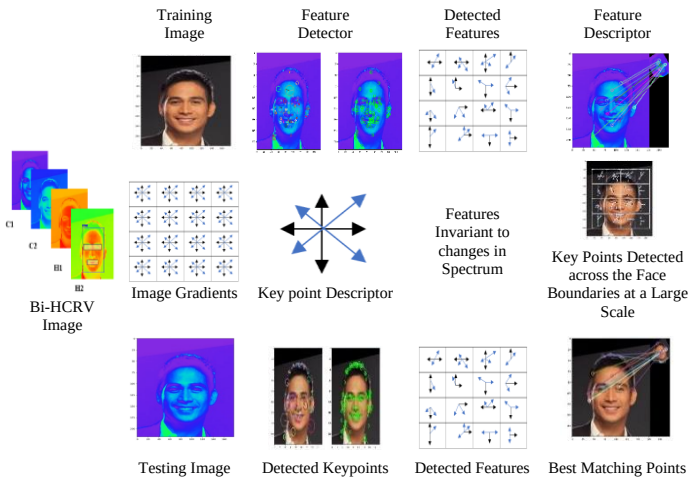

Fig.8. Perceptible Feature points invariant to changes in Spectrum

We present the results of AUC and ERR in Table.1 by comparing with State-of-the-art approaches, base line methods and our proposed model. We use the baseline methods "RAW" and Mahendran et al. [10] for prove the significance of a Joint Fusion based approach and a Cross-Modality modules. We also compare our Cross Spectrum Face recognition model with [32], to prove how Synthesized images from multiple modalities can increase the face recognition accuracy. Lastly, we compare our Cross-Spectrum face Synthesis and Receptive Face Recognition Resnet Model (CSFS-RFRR Model) the State-of-the-art performance in Zhang et al. [32].

6. CONCLUSION

A Flexible filter is introduced to reduce noise while preserving the edges simultaneously by shifting the central pixel value closer to 1 and greater to preserve edges with noise removal. The proposed filter allows to maintain sharp edges by averaging fewer central pixels thus reducing computation time. By preprocessing the input images with the proposed Flexible filter, we are able to increase the accuracy of face detection by 10%. We develop a novel framework for overcoming the issues related to occlusion by using cross spectrum image synthesis of Bi-HCRV images, by computing adaptive weight calculation using size-Based Weights and pattern-Based Weights. For the later method we detect the edges of the image using Canny edge detection algorithm and compute the sum of those pixels found in the detected edges. We extract high dimensional feature vectors by using pre-trained ResNet18 model by removing the classification layer to retain the feature maps from second last layer of the model. The obtained feature maps are passed through a FCNN MODEL, which uses Joint Fusion technique and contrastive learning to create Joint Latent Embedding. The training pipeline uses Adam Optimizer and MSE loss function to minimize the loss among Latent Embeddings and a Zero tensor. The training loop is iterated for 10 Epochs and the loss values are monitored for each epoch. We compute Contrastive loss using to verify images with same identity have relatively similar embeddings and images with different identities have dissimilar identities and to group images with similar feature vector pairs in a matrix structure. Euclidean distance measure is calculated between Anchor image and Positive image pairs, and Anchor image and Negative image pairs. The positive distance measure is used to check the relative closeness measure of Anchor and Positive images using Eq.(22). The negative distance measure is used to check the relative distance (far) measure of Anchor and Positive images. In order to make sufficient learning of latent vectors from V_i, S_i, T_i and make the features to look much similar we minimize the coupling error during the training process of images (X, Y and Z) from cross-spectrum. We include Kullback-Leibler (KL) divergence using Multivariate Gaussian Distribution to align the latent vector Embeddings with the Input requirements of classification and prediction Model, to enable a smooth coupling process and to prevent the model from over fitting. For the downstream Classification task, we use a Face Predictor model based on Ensemble of Regression Trees to compute and visualize face landmarks. The proposed Receptive Face Recognition Resnet Model enables Cross-spectrum face recognition by comparing images from two different modalities such as real-visible image and hot-cold thermal images. The model uses 3x3 kernels at convolutional layers for capturing local features and 5x5 kernel to capture more spatial information. We use Stride 1 and 2 to capture fine grained details and to reduce the computation load by

increasing the receptive field. We achieve down sampling by increasing the stride instead of using normal pooling operation like CNNs. We evaluate the model using two different datasets CelebA [34] and LFW [35] for cross spectrum Face Recognition. We compare the results of the image fusion and face recognition accuracy with State-of-the-art Cross-Spectrum synthesized imagery based on Performance Test of the Model using Landmark detection and Face recognition accuracy using AUC (Area Under the Curve) and EER (Equal Error Rate) measures from the ROC (Receiver Operating Characteristic) curves. The results derived from Table.1, clearly depicts CSFS-RFRR Model (ours) has shown a performance improvement of 20% in AUC than the State-of-the-art with Polarization State Information, and 23% improvement in AUC over the State-of-the art models in traditional Thermal-to-Visible synthesis process. As well 12% improvement in EER polar and 9% improvement in EER (S0) are observed while comparing with Sate-of-the-art Models in Traditional thermal case.

REFERENCES

[1] T. Bourlai, N. Kalka, A. Ross, B. Cukic and L. Hornak, "Cross-Spectral Face Verification in the Shortwave Infrared (SWIR) Band", *Proceedings of International Conference on Pattern Recognition*, pp. 1343-1357, 2010.

[2] F. Juefei-Xu, D.K. Pal and M. Savvides, "NIR-VIS Hetero Geneous via Cross-Spectral Joint Dictionary Learning and Reconstruction", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 141-150, 2015.

[3] B. Klare and A.K. Jain, "Heterogeneous Face Recognition: Matching NIR to Visible Light Images", *Proceedings of International Conference on Pattern Recognition*, pp. 1513-1516, 2010.

[4] F. Nicolo and N.A. Schmid, "Long Range Cross-Spectral Face Recognition: Matching SWIR against Visible Light Images", *IEEE Transactions on Information Forensics and Security*, Vol. 7, No. 6, pp. 1717-1726, 2012.

[5] N.J. Short, S. Hu, P. Gurram, K. Gurton and A. Chan, "Improving Cross-Modal Face Recognition using Polarimetric Imaging", *Optics Letters*, Vol. 40, No. 6, pp. 882-885, 2015.

[6] T. Bourlai, N. Kalka, A. Ross, B. Cukic and L. Hornak, "Cross-Spectral Face Verification in the Short Wave Infrared (SWIR) Band", *Proceedings of International Conference on Pattern Recognition*, pp. 1-9, 2010.

[7] T. Bourlai, A. Ross, C. Chen and L. Hornak, "A Study on using Mid-Wave Infrared Images for Face Recognition", *Proceedings of International Conference on Society for Optical Engineering*, Vol. 8371, pp. 1-10, 2012.

[8] N.D. Kalka, T. Bourlai, B. Cukic and L. Hornak, "Cross-Spectral Face Recognition in Heterogeneous Environments: A Case Study on Matching Visible to Short-Wave Infrared Imagery", *Proceedings of International Joint Conference on Biometrics*, pp. 1-15, 2011.

[9] X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 11, pp. 1955-1967, 2009.

[10] H. Li, H. Qiu, Z. Yu and Y. Zhang, "Infrared and Visible Image Fusion Scheme based on NSCT and Low-Level Visual Features", *Infrared Physics and Technology*, Vol. 76, pp. 174-184, 2016.

[11] K. Amolins, Y. Zhang and P. Dare, "Wavelet based Image Fusion Techniques An Introduction, Review and Comparison", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 62, No. 4, pp. 249-263, 2007.

[12] S. Li, X. Kang, L. Fang, J. Hu and H. Yin, "Pixel-Level Image Fusion: A Survey of the State of the Art", *Information Fusion*, Vol. 33, pp. 100-112, 2017.

[13] H. Li, X. Liu, Z. Yu and Y. Zhang, "Performance Improvement Scheme of Multifocus Image Fusion Derived by Difference Images", *Signal Processing*, Vol. 128, pp. 474-493, 2016.

[14] Q. Zhang, Y. Liu, R.S. Blum, J. Han and D. Tao, "Sparse Representation based Multi-Sensor Image Fusion for Multi-Focus and Multi-Modality Images: A Review", *Information Fusion*, Vol. 40, pp. 57-75, 2018.

[15] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward and X. Wang, "Deep Learning for Pixel-Level Image Fusion: Recent Advances and Future Prospects", *Information Fusion*, Vol. 42, pp. 158-173, 2018.

[16] H. Yin, S. Li and L. Fang, "Simultaneous Image Fusion and Super Resolution using Sparse Representation", *Information Fusion*, Vol. 14, No. 3, pp. 229-240, 2013.

[17] J. Zhong, B. Yang, Y. Li, F. Zhong and Z. Chen, "Image Fusion and Super Resolution with Convolutional Neural Network", *Proceedings of International Conference on Communications in Computer and Information Science*, pp. 78-88, 2016.

[18] H. Li, Z. Yu and C. Mao, "Fractional Differential and Variational Method for Image Fusion and Super-Resolution", *Neurocomputing*, Vol. 171, pp. 138-148, 2016.

[19] K. Gurton, A. Yuffa and G. Videen, "Enhanced Facial Recognition for Thermal Imagery using Polarimetric Imaging", *Optics Letters*, Vol. 39, No. 13, pp. 3857-3859, 2014.

[20] N.J. Short, S. Hu, P. Gurram and K. Gurton. "Exploiting Polarization-State Information for Cross Spectrum Face Recognition", *Proceedings of International Conference on Biometrics Theory, Applications and Systems*, pp. 1-11, 2015.

[21] N.J. Short, S. Hu, P. Gurram, K. Gurton and A. Chan, "Improving Cross-Modal Face Recognition using Polarimetric Imaging", *Optics Letters*, Vol. 40, No. 6, pp. 882-885, 2015.

[22] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", *Proceedings of International Conference on Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 1-9, 2001.

[23] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", *Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1-20, 2005.

[24] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li and D. Tao, "Discriminative Dictionary Learning-based Multiple Component Decomposition for Detail Preserving Noisy Image Fusion", *IEEE Transactions on Instrumentation and Measurement*, Vol. 69, No. 4, pp. 1082-1102, 2020.

[25] H. Li, J. Xu, J. Zhu, D. Tao and Z. Yu, "Top Distance Regularized Projection and Dictionary Learning for Person

Re-Identification", *Information Sciences*, Vol. 502, pp. 472-491, 2019.

[26] Q. Zhang, Y. Liu, R.S. Blum, J. Han and D. Tao, "Sparse Representation based Multi-Sensor Image Fusion for Multi-Focus and Multi-Modality Images: A Review", *Information Fusion*, Vol. 40, pp. 57-75, 2018.

[27] J.L. Starck, M. Elad and D.L. Donoho, "Image Decomposition via the Combination of Sparse Representations and a Variational Approach", *IEEE Transactions on Image Processing*, Vol. 14, No. 10, pp. 1570-1582, 2005.

[28] Y. Jiang and M. Wang, "Image Fusion with Morphological Component Analysis", *Information Fusion*, Vol. 18, pp. 107-118, 2014.

[29] Z. Liu, Y. Chai, H. Yin, J. Zhou and Z. Zhu, "A Novel Multi-Focus Image Fusion Approach based on Image Decomposition", *Information Fusion*, Vol. 35, pp. 102-116, 2017.

[30] H. Li, X. He, D. Tao, Y. Tang and R. Wang, "Joint Medical Image Fusion, Denoising and Enhancement via Discriminative Low-Rank Sparse Dictionaries Learning", *Pattern Recognition*, Vol. 79, pp. 130-146, 2018.

[31] B.S. Riggan, N.J. Short, S. Hu and H. Kwon, "Estimation of Visible Spectrum Faces from Polarimetric Thermal Faces", *Proceedings of International Conference on Biometrics Theory, Applications and Systems*, pp. 1-6, 2016.

[32] H. Zhang, V.M. Patel, B.S. Riggan and S. Hu, "Generative Adversarial Network-based Synthesis of Visible Faces from Polarimetric Thermal Faces", *Proceedings of International Joint Conference on Biometrics*, pp. 1-10, 2017.

[33] A. Mahendran and A. Vedaldi, "Understanding Deep Image Representations by Inverting Them", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-11, 2014.

[34] Liu Ziwei, Luo Ping, Wang Xiaogang and Tang Xiaoou, "Deep Learning Face Attributes in the Wild", *Proceedings of International Conference on Computer Vision*, pp. 1-7, 2015.

[35] Gary B. Huang, Manu Ramesh, Tamara Berg and Erik Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments", *Technical Report*, pp. 7-49, October, 2007.