December 2016 Technical Report: KUL/ESAT/PSI/1604

Hyperspectral CNN for image classification & band selection, with application to face recognition

Vivek Sharma, Ali Diba, Tinne Tuytelaars, and Luc Van Gool

PSI

Kasteelpark Arenberg 10 box 2441
3001 Leuven, Belgium

KU LEUVEN

Telephone: +32-(0)16-32.17.13 Fax: +32-(0)16-32.17.23 http://www.esat.kuleuven.be/psi

Hyperspectral CNN for Image Classification & Band Selection, with Application to Face Recognition

Vivek Sharma[◇], Ali Diba[◇], Tinne Tuytelaars[◇], and Luc Van Gool[◇∓]
[◇]KU Leuven, ESAT-PSI, iMinds [∓] BIWI, CVL, ETH Zürich
{firstname.lastname}@esat.kuleuven.be

Abstract

With hyperspectral sensor technology evolving and becoming more cost-effective, it is likely we will see hyperspectral cameras replace standard RGB cameras in a multitude of applications beyond these traditional niches of medical and aerial image segmentation in the near future. Rather than generating an image that is optimal for the human eye, responses of these new cameras will be tuned towards specific computer vision algorithms. This calls for new methods for hyperspectral band selection, optimized for those tasks. In this work, we present a novel pipeline for discriminative band selection in hyperspectral images for the task of image-level classification. It is based on a convolutional neural network to learn appropriate representations combined with AdaBoostSVM for band selection. We test our method on two standard hyperspectral face datasets in the context of face recognition. Our exhaustive experiments show that the proposed method outperforms the existing state-of-the-art methods.

1. Introduction

For each pixel in an image, a hyperspectral camera acquires the light intensity for a large number of contiguous spectral bands. Every pixel in the image contains a continuous spectrum, which can be used to characterize the objects in the scene with great precision and detail. Following the recent advances in sensor development and computational power, hyperspectral imaging has moved from rather slow and unreliable experimental prototypes to reliable and accurate analytical instruments. The rich spectral information contained in the hyperspectral image or hyperspectral cube (HSI) makes them well suited not just for pixel-level classification in medical or aerial images, but also for accurate classification of objects in computer vision tasks like scene recognition [3], pedestrian detection [16, 27], medical imaging [32, 50], and more. However, the large volume of hyperspectral image is considered a strong drawback. In this paper, we propose a framework for HSI classification and band selection to find potentially uncorrelated, and discriminative wavelengths in combination with a powerful representation of their content that characterize the object and improve the classification performance using end-to-end learning. We investigate in the visible ($380 - 700nm$) to near-infrared ($750 - 1100nm$) spectral range (V-NIR).

Silicon sensors are naturally sensitive up until 900nm, but even able to capture wavelengths up to 1200nm. Commercial digital cameras use a infrared blocking filter to prevent an unwanted NIR response. Nowadays, cameras in cars exploit V-NIR range for pedestrian detection in night vision [27]. Likewise, in mammography, single-shot spectral imaging is used for breast tumour and developing cancer detection [32]. We foresee that in the near future, *task-specific* consumer cameras will take over RGB cameras. As task-specific wavelengths pertinent to objects are found, we can use a color filter array for capturing those wavelengths only. This can lead to more accurate classification than what can be obtained with visible RGB image. The photocells are only sensitive to the spectrum of our interest, limited to a specific range. This makes the imaging system compact, computationally efficient, cost-effective, and a perfect fit for real-time applications.

In a HSI, the adjacent neighboring spectral bands are highly correlated, and it has been observed that high redundancy leads to poor generalization capabilities of the classifier [15]. To the best of our knowledge, only pixel-level band selection and classification [4, 5, 38] have been addressed so far for hyperspectral imaging, where principal component analysis [18] and vector quantization [13] are the most commonly used techniques for dimensionality reduction in hyperspectral data. The high dimensionality of hyperspectral images makes it difficult to separate the discriminative bands using statistical methods [21, 35] due to the high computational burden at the pixel-level. However, in our method the band selection and classification is done *at the image-level*, where each band in a hyperspectral cube can be considered as a separate image. We target the band selection problem at the image-level because we believe that

this allows to exploit high-level information from shapes and abstract concepts from images in comparison to pixelwise selected bands. To this end, we use state-of-the-art methods for image classification, i.e. convolutional neural networks [6, 20, 22, 43] (CNN). We propose a new framework to learn a spectral CNN for obtaining a new feature space using V-NIR information in images. The proposed scheme for training this CNN helps to handle the requirement of many images in training process. Additionally, we use AdaBoostSVM [23, 34] for band selection based on image-level classification. We choose AdaBoostSVM because of several reasons: (i) high performance in remote sensing literature [23, 34]; (ii) excellent generalization in imbalanced classification problems [46]; and (iii) ability to distinguish highly uncorrelated and discriminative features [23, 34, 46]. This makes AdaBoostSVM very promising for classification of hyperspectral data.

We have evaluated our proposed methods using two standard hyperspectral face datasets [9, 11] for face recognition. In summary, we have the following contributions: (i) We propose a scheme for training a CNN for hyperspectral image classification; (ii) We propose a new effective approach to exploit CNN features for discriminative band selection using AdaBoostSVM; (iii) Our proposed method outperforms state-of-the-art methods and traditional hand-crafted features on both datasets.

The remainder of the paper is structured as follows. In Section 2, we discuss the related work. Section 3 describes our proposed method. Results and experimental evaluation are given in Section 4. Finally, in Section 5 we conclude the paper .

2. Related Work

This section first discusses band selection techniques. It then continues with a short description of face recognition in the hyperspectral domain. Finally, we show the impact of multispectral information in the computer vision domain.

Band selection techniques: Discriminative spectral band selection in a hyperspectral cube is a fundamental problem. Maximally discriminative wavelengths increase the recognition accuracy. Therefore, it is advantageous to drop the least discriminative bands from the hyperspectral cube. This reduces the data redundancy, computation complexity, and the acquisition time of the hyperspectral cube, which is very good for real-time applications. In the last two decades, many band selection techniques have been proposed in a remote sensing context, but in this community the band selection and classification problems are done at the pixel-level. The different band selection techniques are: exhaustive search [12, 17], branch-and-bound search [29], best individual features [19], sequential forward/backward selection [17, 19], sequential forward/backward floating search [33], and more. All of these band selection methods use a criterion function that is usually linked to some performance metric, either based on a distance metric (e.g. Jeffries-Matusita distance, Bhattacharyya distance [37]) or on information measure (mutual-information or entropy [4]). In our work, we show how AdaBoostSVM can be applied to this setting for band selection based on image-level classification. We follow [19] as baselines for band selection.

For compact representation, a panchromatic image, which is a well-known single channel grayscale representation of a hyperspectral image is often used. In literature good results have been demonstrated using panchromatic images for classification [7]. For comparison, we also obtain the panchromatic and RGB images as baselines.

Hyperspectral image classification using CNN: While most previous work in remote sensing using CNNs [14, 49, 42] consider pixel-level classification. We propose *image-level* classification that allows us to exploit higher-level information like shapes and abstract concepts from images: making it more suitable for high-level visual recognition tasks, similar is not possible at pixel-level. Also, classification at the pixel-level (think of raw pixel values) comes at high computational burden, and it turns out that, it is difficult to disambiguate objects with large classes dataset.

Hyperspectral image classification for face recognition: Several hyperspectral feature extraction methods have been proposed lately for face recognition [25, 40, 44, 45]. They transform the high-dimensional space to a low-dimensional space, and exploit this low-dimensional space for classification. Shen et al. [40] utilize 3D Gabor wavelets to extract orientation, scale, and wavelength-dependent features from the hyperspectral images. Liang et al. [25] focus on 3D texture pattern descriptors based on local derivative patterns as features. Uzair et al. [44] use the encoded low-frequency components of the 3D discrete cosine transformation as features for face recognition. Further in [45], Uzair et al. extract the spatiospectral covariance features using 3D cubelets. Bianco [2] performs 1D projections along the spectral dimension using an unbounded linear combination (ULC), and obtains the optimal projection using Particle Optimal Optimization (PSO) as features. Also recently the traditional hand-crafted feature descriptors (HOG, LBP, SIFT) are explored by Sharma et al. [39]. We compare against these methods in our experimental section.

Vision: A lot of recent applications have been demonstrated in computer vision that exploits multispectral information in V-NIR and SWIR ($1400 - 3000nm$) range. They report high detection and recognition rates for different tasks in these ranges. Such as, scene recognition [3], predicting clinical outcomes [50], 3D reconstruction [52], salient object detection [24], pedestrian tracking [16, 27], eye tracking [51], material classification [36, 41], cultural heritage [1] and so on. All of them have reported a promi-

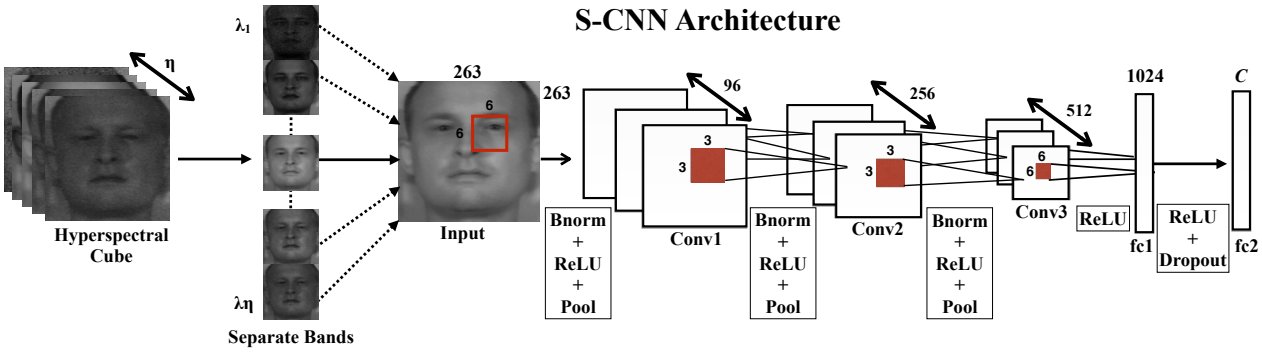


Figure 1: S-CNN architecture for hyperspectral image classification. At a time, one spectral image (or band) is fed to the network for training. S-CNN consists of three convolutional and two fully-connected layers, with a final C-way softmax. Note how a single network is learned that is applied to all the different bands.

nent improvement in classification. In our work we not only improve the classification performance, but also find the most discriminative bands that play a role behind it.

3. Method

In this section, we go through our scheme for training a CNN for hyperspectral image classification and also a new approach to exploit CNN features for discriminative and complementary band selection using AdaBoostSVM. In the first part of this section we describe our proposed hyperspectral CNN for image classification. Second part of the section introduces our proposed discriminative band selection technique using CNN and AdaBoostSVM. For the full pipeline of the proposed method of hyperspectral image classification and band selection, refer to the supplimentary material.

3.1. Hyperspectral CNN for Image Classification

In the last decade, many works in the computer vision community have focused on feature designing and descriptor engineering. Hand-crafted descriptors (like SIFT, HOG, ...) are popular in many computer vision domains, but none of them are trainable for new problems and different tasks. Recently, features derived from learning-based representation have been shown to outperform these engineered descriptors, because they have the power of discovering and optimizing visual description for the specific task to be solved. In this context, convolutional neural networks are leader in the field of learning-based feature extraction methods [20, 22, 43]. With variations in architecture of these network, researchers can obtain new models which are well fitted for their desired problem. In this work, we propose a scheme for training a CNN for hyperspectral image classification. This network can capture discriminative visual information in useful bands for our task.

We will now describe our CNN architecture for hyperspectral image classification. For our convenience, we refer to our CNN as S-CNN (see Fig. 1). In our scheme we make this hyperspectral cube flat by treating each band as a separate image. Each of these images has the same class as the hyperspectral cube it was extracted from. Having the CNN work on a single band as a generic CNN helps to improve the classification performance: treating the data in this flat way handles the requirement of many images in the CNN training process. We design our network to have 3 convolution layers (conv 1-3), followed by 2 fully-connected layers (fc 1-2) with a final C-way softmax. The softmax output layer produces a distribution over the C output class labels using softmax loss function, where C is the number of classes. All the convolution layers are followed by a batch normalization layer (bnorm), a rectification linear unit layer (ReLU), and a max pooling layer. The ReLU and a dropout layer are applied to the output of fc1.

Notation: We start from a hyperspectral image (or hyperspectral cube) with a size of $h \times w \times \eta$ where η is the number of wavelengths (or bands), h and w are the height and width of the frame, respectively. In our work, each frame $(h \times w)_{\lambda_i}$, $i \in [1...\eta]$ of the hyperspectral cube is treated as a separate image $x_{(h \times w)_{\lambda_i}}$ or x_{λ_i} (e.g. grayscale image) for image classification and recognition tasks.

Architecture Study: We train our hyperspectral CNN architecture (S-CNN) for face recognition. The key reason behind training our own network rather than using a pretrained one is to enable our S-CNN to learn representations of discriminative texture patterns in the NIR range, that is unavailable in the visible range. From the literature [6, 20, 43], we exploited a lot of insights about good architectures. For small datasets, it is highly recommended to train small networks. It turns out that deep networks have many parameters to learn and an insufficient number of training samples leads to over-fitting.

Figure 2 shows the accuracy of different architectures on the CMU-HSFD dataset [9]. The 5-layer network of

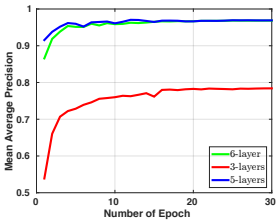


Figure 2: S-CNN architecture study. Face recognition performance on CMU validation dataset for different network architectures. The 5-layer network performs the best among all other networks.

Fig. 1 described before performs the best when compared with other networks. Note that the 3-layer network performs the worst, and the 6-layer one shows similar performance as our 5-layer network. In the 3-layer network we have 2 convolutional and 1 fully-connected layers, while the 6-layer network consists of 3 convolutional and 3 fully-connected layers. Using the 5-layer network is computationally more time efficient over the 6-layer network. So we use the 5-layer network as S-CNN network. The number of parameters for 5-layer network is 20 million. We also trained Alexnet [20] (60 million parameters) and VGG [6] (99-144 million parameters) architectures from scratch, but we found that these networks were very deep and are not suitable for small datasets: the huge number of parameters quickly led to over-fitting (see Sec. 4.3).

We also tried to train S-CNN with more complex architectures, such as by passing the whole cube as an input, but we found that the model was underperforming for the given limited number of sample set.

3.2. Discriminative Band Selection Using S-CNN

Next, we describe the discriminative band selection at image-level in the hyperspectral image through S-CNN. For this purpose, we use AdaBoostSVM [23, 34, 46] for image-level band selection. We believe that AdaBoostSVM is most suitable for our task. Each band is an independent diverse feature set. AdaBoost depends on diversity, and demonstrates excellent generalization performance for combining the complementary relationships between different bands. Thus using it will help us to select and combine the best bands. We choose SVM as a learner because of their lower sensitivity to imbalanced datasets, which is very promising for classification of hyperspectral data.

Using our trained S-CNN, for a given wavelength λ_i, $i \in [1...\eta]$, we extract S-CNN features: activations of first fully-connected layer fc1 for each image, followed by L2 normalization. Once we have extracted the normalized features for all bands, we assign the S-CNN features of the i-th band to a SVM learner in AdaBoost (see Algorithm 1). We use multi-class weighted SVM (with the RBF kernel) as a learner in AdaBoost.

Data: Input training data: $\{(x_{([\lambda_1,...,\lambda_\eta],i)}, y_i)\}$, $i = 1...M$ and $y_i \in \{1, ..., C\}$. Number of required complimentary bands, K
Result: Output: weight of K learned models, $\mathbf{W}$
Initialization:
Weights: $w_i := 1/M$. Bands: $B := 1 : \eta$
Buffer for Accuracy: $\mathbf{T} \leftarrow \emptyset$;
for *k = 1 to* K **do**
  for $b = B$ **do**
    $D_b := \{(x_{(\lambda_b,1)}, y_1)...(x_{(\lambda_b,M)}, y_M)\}$;
    learn model h_b from D_b with w_i, using SVM;
    $Acc_b :=$ Classification accuracy of model h_b;
    $\mathbf{T} := \mathbf{T} \cup \{Acc_b\}$;
  end
  Calculate $\underset{b' \in B}{\operatorname{argmax}}\ \mathbf{T}_{b'}$. Band $b' \in B$ for which maximum accuracy is obtained;
  $\epsilon_{b'} := \sum_{h_{b'}(x_{(\lambda_{b'},i)}) \neq y_i} w_i$;
  $\beta_{b'} := \epsilon_{b'}/(1 - \epsilon_{b'})$;
  for i = 1 to M : **if** $h_{b'}(x_{(\lambda_{b'},i)}) = y_i$ **then** $w_i := w_i * \beta_{b'}$;
  $s := \sum_i w_i$;
  for all w_i: $w_i := w_i/s$;
  $W_k := log(1/\beta_{b'})$;
  $B := B \backslash \{b'\}$;
end
return: $\mathbf{W}$= $[W_1, W_2, ..., W_K]$

Algorithm 1: Discriminative band selection in the hyperspectral image using AdaBoostSVM with S-CNN features.

Given that we are interested in finding K complementary bands out of η bands in the hyperspectral image. We initialize each training sample x with a weight of $1/M$. M is the number of training samples. The train/test set samples are the same for all bands. In the first iteration of AdaBoost, the best performing band is chosen. The weight of each sample in the training set is updated depending on whether the predicted label is correctly classified or misclassified. If correctly classified x gets lower weights, otherwise higher weights. In the second iteration, the second best performing band is chosen that is complementary to the first one. This time AdaBoost focuses on the training samples which were misclassified previously (i.e. training samples with higher weights). After the second band is chosen, the weights of x are again re-weighted, and the same process continues for K iterations. Finally, AdaBoost combines all the K-learned models together as an ensemble. The predictions of this ensemble are combined through a weighted majority vote among the prediction of the different models. These models

correspond to the most K discriminative and complementary spectral wavelengths in the hyperspectral cube which can be used for accurate classification. Figure. 3 shows the accuracy of S-CNN+AdaBoostSVM for K complimentary bands on CMU-HSFD and PolyU-HSFD datasets.

4. Experiments

We evaluate our proposed method on two datasets for face recognition. Our experiments consist of four parts (i) Comparison of our proposed method with state-of-the-art methods (Sec. 4.2); (ii) Comparison of our proposed image-level band selection method with other methods for band selection (Sec. 4.2); (iii) Comparison of S-CNN feature with the traditional hand-crafted features (Sec. 4.3); (iv) Comparison of S-CNN best bands with RGB and panchromatic image representations (Sec. 4.3); (v) Comparison of S-CNN architecture with other state-of-the-art architectures (Sec. 4.3). In Section 4.1, we first explain the experimental details with dataset, implementation details, training protocol, and baselines.

4.1. Experimental details

All experiments were performed using the publicly available vlfeat library [47] and matconvnet framework [48].

Hyperspectral datasets: For training a CNN, we need to have a hyperspectral dataset with enough training samples per object class. To the best of our knowledge, there are only two publicly available hyperspectral datasets, which contain enough samples per subject. Both are face datasets [9, 11]. Therefore, our experiments and evaluations are conducted on these two datasets. They are Carnegie Mellon University [9] (CMU-HSFD) and Hong Kong Polytechnic University [11] (PolyU-HSFD) hyperspectral face datasets.

PolyU-HSFD (see Fig. 4) is acquired with the CRI's VariSpec Liquid-Crystal-Tuneable-Filter (LCTF) with a halogen light. Each hyperspectral cube contains 33 bands covering the visible range of 400-720nm with a step size of 10 nm. For each individual, frontal, left and right views

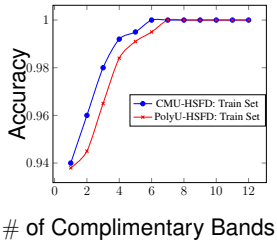


Figure 3: S-CNN+AdaBoostSVM: Accuracy vs # of complimentary bands (see Table 2). Figure shows that using K complimentary bands, the misclassified samples are correctly classified when trained and tested on train set.

with neutral-expression for multiple sessions were acquired. Each subject has 1-7 cubes-per-view. The database contains 48 subjects with 3-21 cubes-per-subject over all sessions. Following the experimental protocol of [25, 44, 45], we use 25 subjects for our experiments. The captured images are quite noisy, and have a low signal-to-noise ratio for the first 6 and the last 3 bands of the spectral range. Following the same protocol, we have discarded them. The relative proportion of signal-to-noise ratio is low in this dataset. This implies the shot noise is high. Therefore, we apply a median filter to remove shot noise as a preprocessing step.

CMU-HSFD (see Fig. 4) is acquired with the CMU-developed Acousto-Optic-Tunable-Filter (AOTF) with three 600W halogen bulbs. Each hyperspectral cube contains 65 bands covering the V-NIR range of 450-1090nm with a step size of 10 nm. For each individual, frontal, left and right views with neutral-expression for multiple sessions were acquired. Each subject has 1-5 cubes-per-view. The database contains 54 subjects with 4-20 cubes-per-subject over all sessions. Following the experimental protocol of [25, 44, 45], we use 48 subjects for our experiments. We also apply a median filter in this dataset to remove shot noise.

Training Protocol for S-CNN: Following the experimental protocol of [25, 44, 45], we use only the frontal views, so we can compare our technique with state-of-the-art methods for hyperspectral face recognition. For the evaluation using the defined protocol, the gallery set (training set) is constructed by randomly selecting one cube-per-subject, while the probe set (test set) is constructed by randomly selecting two cubes-per-subject. The remaining cubes are used for training the S-CNN for both datasets. We use this setup in Sec. 4.2.

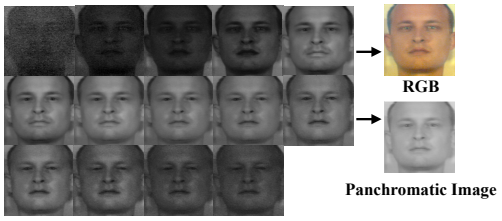
In our second setup, we use frontal, left and right views for comparison of our technique with the traditional hand-crafted features. We construct the gallery, probe, and S-CNN training set in the same way as in the first setup, but this time we use also left and right views. We use this setup in Sec. 4.3.

Once we have trained our S-CNN, we extract the S-CNN features (fc1) for the gallery and the probe set images. Then, a multi-class linear SVM is trained on S-CNN gallery set and tested on the S-CNN probe set, and then we report the recognition accuracy using majority voting.

Data preprocessing: We use flip augmentation, mirroring images about the x-axis and y-axis, thus generating horizontal and vertical reflections for each image. This allows us to train deeper network and avoid over-fitting. Augmentation consistently improves performance [6]. All images were cropped and resized to 263×263 for S-CNN training.

S-CNN implementation details: An input image of size 263×263 pixels is fed to conv1 with 96 kernels of size $6 \times 6 \times 96$ with a stride of 2 pixels. The conv2 takes as

(a) PolyU-HSFD: Visible Range (440-690nm). Bands with a step-size of 40nm are shown.



(b) CMU-HSFD: V-NIR Range (450-1090nm). Bands with a step-size of 50nm are shown.

Figure 4: Example hyperspectral image for a subject with its corresponding RGB and panchromatic images. For the detailed explanation of the conversion of HSI to RGB and panchromatic images, we encourage the reader to visit the supplementary material.

input the (bnorm, ReLU, pooled) output feature maps of conv1 and filters it with 256 kernels of size $3 \times 3 \times 256$ with a stride of 2 pixels. This is followed by conv3 which takes as input the (bnorm, ReLU, pooled) output feature maps of conv2 and filters it with 512 kernels of size $3 \times 3 \times 512$ with a stride of 1 pixel. The fc1 has 1024 outputs, where dropout was applied after fc1 with a rate of 0.5. Each output of the fc is connected to all the inputs. The resulting output of fc2 is fed to a C-way softmax to compute the class posterior probabilities.

S-CNN details for training: We train our S-CNN from scratch using stochastic gradient descent with momentum set to 0.9, weight decay of 0.0005, and with a batch size of 60. We initialize an equal learning rate for all trainable layers to 0.05, which is manually decreased by a factor of 10 when the validation error stopped decreasing. Prior to the termination the learning rate was reduced two times at 15th and 25th epoch. We initialize the weights of the network with normally distributed random numbers. We trained the network for 30 epochs which took 6 hours on a single NVIDIA GTX 980 4GB GPUs.

Baseline features: We compare S-CNN feature with a few baselines: the traditional hand-crafted features, namely SIFT [26], HOG [8], and LBP [30]. For HOG and LBP, we use a cell-size of 8×8, and number of oriented histogram bins is 9 for HOG. For SIFT, we use a bin-size of 8 and step-size of 4, then we Fisher encode the extracted SIFT features and return the L2-normalized feature vector. To compute Fisher encoding, we need to obtain a visual word dictionary, for that we use a GMM with 100 clusters. We denote dense SIFT Fisher vectors by DSIFT-FVs. For all computation and training of methods based on hand-crafted features, we keep these parameters fixed.

4.2. Frontal View Experiments

Evaluation of our S-CNN features: In order to evaluate the performance of S-CNN, we compare it with state-of-the-art methods on both datasets. For a fair comparison, we use the same baseline as in [25, 44, 45] and use only the frontal view for training and testing over all the bands. In this experiment, we do not perform any band selection yet, but use all available bands. To evaluate the quality of the feature we use a simple linear SVM classifier. We extract the S-CNN features and feed band-by-band to a multi-class linear SVM. The decisions of the different bands are merged using majority voting. In Table 1, we report the recognition accuracy of S-CNN and compare it with the best results reported in the literature. Our S-CNN classification results outperform all the state-of-the-art methods, and traditional hand-crafted features for hyperspectral face recognition on both datasets. Our method achieved the highest recognition accuracy, exceeding DSIFT-FVs [39], Band fusion+PLS [45], and 1D ULC+PSO [2].

Comparison of band selection methods: We compare our proposed discriminative band selection in Section 3.2 with the best individual bands [19], randomly selected band subsets, and gradual band removal band selection methods. In the best individual bands method, we obtain the classification accuracy for each individual band, and a subset of best bands with maximum performance are chosen. In randomly selected bands method, we randomly select a number of disjoint band subsets from the whole spectral range. Gradual band removal method involves successively removing one band at a time, whose removal increases the accu-

Methods	Accuracy	
	PolyU-HSFD	CMU-HSFD
3D DCT [44]	84.0	88.6
3D Gabor Wavelets [40]	90.1	91.6
3D LDP [25]	95.3	94.8
Band fusion+PLS [45]	95.2	99.1
1D ULC+PSO [2]	99.1	–
LBP [39]	85.6	86.1
HOG [39]	92.3	91.5
DSIFT-FVs [39]	96.1	96.9
S-CNN: Majority-Voting (ours)	97.2	98.8
S-CNN+SVM: Majority-Voting (ours)	**99.3**	**99.2**

Table 1: Comparison of S-CNN with state-of-the-art methods using all bands.

Band Selection Methods	PolyU-HSFD	CMU-HSFD
Best Individual Bands [19]	[560, 640, 490, 500, 630]	[650, 660, 720, 910]
Gradual Band Removal	[490, 530, 570, 630, 690]	[850, 930, 990, 1050]
S-CNN+AdaBoostSVM (ours)	[640, 470, 590, 570, 520]	[1010, 900, 970, 730]

Table 2: Selected bands for S-CNN and other methods for band selection.

racy of the remaining bands most. We do it repeatedly, until we are left with the bands with maximum accuracy.

For a fair comparison, we use the same baseline as in [45] and select the best 5 bands for PolyU-HSFD and best 4 bands for CMU-HSFD. Table 2 shows the selected best subset of bands chosen by the different band selection methods. In Table 3, we compare the recognition accuracy of selected bands by different methods on both datasets. This shows that classification using our selected bands outperforms significantly the other methods. It appears clear that our proposed method, S-CNN+AdaBoostSVM outperforms significantly the other methods with 99.6% on PolyU-HSFD and 99.4% on CMU-HSFD datasets. Note that NIR range (700-1090nm) is present only in CMU-HSFD database. The blue wavelength bands of electromagnetic spectrum are consistently discarded on both datasets showing that they are less informative and discriminative in comparison to green, red and NIR wavelength bands for face recognition. Also we should note that our method finds maximally discriminative bands in the hyperspectral cube, and it very well takes into account the complementary relationships between different bands for image classification task.

Table 4 presents the accuracy results of our S-CNN+AdaBoostSVM compared with Bandfusion+PLS [45]. For both datasets, our method outperforms Bandfusion+PLS and achieves state-of-the-art accuracy, which is currently the best published result, by 4.7% on PolyU-HSFD and 0.6% on CMU-HSFD datasets. The selected subset of bands chosen by Uzair et al. in [45] were {530, 540, 550, 630, 670}nm for PolyU-HSFD and {570, 640, 720, 1000}nm for CMU-HSFD.

Band Selection Methods	PolyU-HSFD	CMU-HSFD
Best Individual Bands [19]	88.9	97.6
Randomly Selected Bands	72.7	78.3
Gradual Band Removal	88.8	97.2
S-CNN+AdaBoostSVM (ours)	**99.6**	**99.4**

Table 3: Comparison of our method with other band selection methods using their corresponding best bands combined with our S-CNN features and linear SVM shown in Table 2 for both datasets.

Methods	PolyU-HSFD	CMU-HSFD
Bandfusion+PLS [45]	94.9	98.8
S-CNN+AdaBoostSVM (ours)	**99.6**	**99.4**

Table 4: Comparison of our method with state-of-the-art [45] using the same number of selected bands shown in Table 2 for both datasets.

	HOG	LBP	DSIFT-FVs	S-CNN (ours)
PolyU-HSFD	[460, 510, 570]	[560, 650, 510]	[460, 640, 520]	[590, 570, 520]
CMU-HSFD	[1030, 1010, 1050]	[720, 730, 650]	[630, 680, 690]	[900, 970, 730]

Table 5: Selected best 3 bands for S-CNN+AdaBoostSVM and hand-crafted features+AdaBoostSVM.

4.3. All Views Experiments

Comparison of S-CNN with baseline features: For this evaluation, we extract the HOG, LBP and DSIFT-FVs features for each individual band in the hyperspectral cube. Then, we apply the hyperspectral AdaBoostSVM algorithm in the same way as we applied to S-CNN features, discussed earlier in Section 3.2 for finding the best bands for each hand-crafted feature. For this evaluation, we select only 3 bands (Best-3-Bands). Table 5 shows the best 3 bands selected by each feature extraction method by AdaBoostSVM on both datasets. For evaluation, we extract the features for the chosen 3 bands and feed them band-by-band to a multi-class linear SVM and the decisions are merged using AdaBoost weighted majority voting.

In Table 6, we compare the recognition accuracy of Best-3-Bands selected for S-CNN with the Best-3-Bands selected for baseline features. We observe that the accuracy of S-CNN outperforms all the baseline features listed in Table 6 and achieves state-of-the-art accuracy with 99.4% on PolyU-HSFD and 99.2% on CMU-HSFD datasets for all views. We can say that, S-CNN learns the high-level abstract/semantic representations, in comparison to hand-crafted features that are designed to work well on visible range.

Comparison with RGB and panchromatic image representation: We generate the RGB and panchromatic images for HSI. Then, we fine-tune the S-CNN trained network on the target images. We extract S-CNN: fc1 features for each image, and then train a multi-class linear SVM and we report the recognition accuracy. For a fair comparison

	HOG	LBP	DSIFT-FVs	S-CNN+AdaBoostSVM
PolyU-HSFD$_{Best-3-Bands}$	80.0	79.4	88.3	**99.4**
CMU-HSFD$_{Best-3-Bands}$	91.5	99.6	98.4	**99.2**

Table 6: Comparison of S-CNN+AdaBoostSVM with baseline features using their corresponding Best-3-Bands shown in Table 5.

	RGB	Best-3-Bands (ours)	Panchromatic	Best-1-Band (ours)
PolyU-HSFD	86.0	**99.4**	88.0	91.3
CMU-HSFD	83.9	**99.2**	89.2	93.0

Table 7: Comparison of S-CNN+AdaBoostSVM: Best-3-Bands (shown in Table 5) with RGB and panchromatic images.

with 3-channel RGB image, we select only three bands for evaluation.

In Table 7, we compare the recognition accuracy of best-3 and best-1 bands selected for S-CNN (Table 5), with the RGB image and a single-channel panchromatic image representations. We observe that the accuracy of S-CNN outperforms the RGB image by a significant margin. It has been reported by Pan et al. [31], using NIR range for face recognition gives higher accuracy in comparison to visible range. The reason being, in NIR range the subsurface tissue features are more discriminative due to larger penetration depth in the human skin. Also, NIR images are to a great degree invariant to illumination, and have discriminative texture patterns that characterize the object [41] and scene recognition [3] better than the computed features in the R, G, and B color channels directly. S-CNN (Best-3-Bands) also performs better than single-channel panchromatic image, by 11.4% on PolyU-HSFD, and 10% on CMU-HSFD.

To qualitatively evaluate the performance of our best 3 bands (Table 5), we compare it with RGB image features. We visualize the learned S-CNN feature embedding on both CMU-HSFD and PolyU-HSFD datasets. We extracted the 1024-dimensional S-CNN: fc1 features for RGB wevelengths and best 3 bands (i.e. {900, 970, 730}nm for CMU-HSFD and {590, 570, 520}nm for PolyU-HSFD) and then projected to 2-dimensional space using t-SNE [28]. Fig. 5 shows the learned feature embedding of the S-CNN features for RGB and our best 3 bands on both datasets. We can observer that our best 3 bands are qualitatively better than the RGB wavelengths.

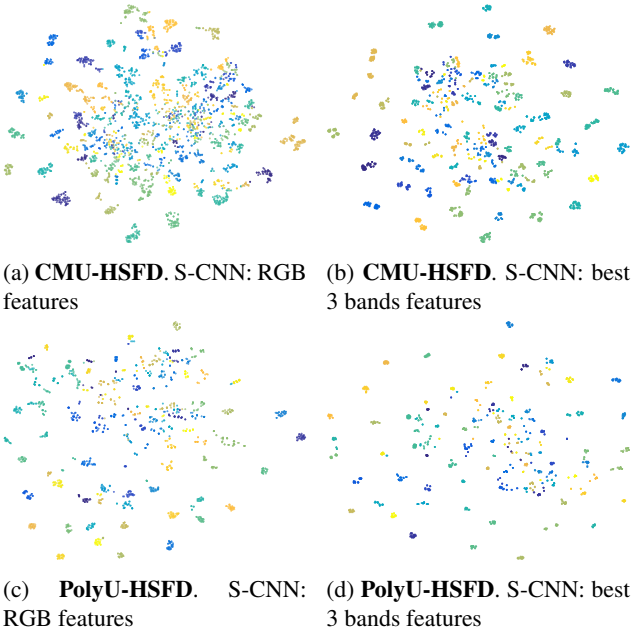
Comparison with other Architectures: In this experiment, we show AlexNet [20] and VGG-M [6] architectures trained on ImageNet [10] and fine-tuned on CMU-HSFD and PolyU-HSFD underperform S-CNN network by a significant margin. Here, we use all views for training and testing over all the bands. For this evaluation, we extract the features and feed band-by-band to a multi-class linear SVM. The decisions of the different bands are merged using majority voting. In Table 8, we compare the recognition accuracy of S-CNN with state-of-the-art architectures.

5. Conclusion

In this paper, we presented a scheme for training a CNN for HSI, along with a new effective approach for discriminative and complimentary band selection in HSI for the task of image- level classification. Our proposed methods are evaluated on hyperspectral face datasets for the task of face recognition. The proposed methods significantly outperform state-of-the-art methods and methods based on hand-crafted features. Even though in this paper we have focussed on the face recognition task. Our method has the potential to generalize to other object classes, given that the HSI literature has shown that spectral information captured by hyperspectral cameras produces better results than RGB cameras in a multitude of applications. Furthermore with the increasing deployment of the HSI devices e.g. car-mounted cameras, cameras for medical applications, precision farming, new datasets will be available. Using our approach, wavelengths pertinent to objects of different categories can be found in the HSI, thus leading to more accurate classification.

Methods	Accuracy		Parameters (million)
	PolyU-HSFD	CMU-HSFD	
VGG-M [6]+SVM: Majority-Voting	63.7	66.4	103
Alexnet [20]+SVM: Majority-Voting	74.3	80.9	60
S-CNN+SVM: Majority-Voting (ours)	**99.3**	**99.2**	**20**

Table 8: Comparison of S-CNN with current state-of-the-art architectures.



(a) **CMU-HSFD**. S-CNN: RGB features (b) **CMU-HSFD**. S-CNN: best 3 bands features

(c) **PolyU-HSFD**. S-CNN: RGB features (d) **PolyU-HSFD**. S-CNN: best 3 bands features

Figure 5: **t-SNE visualization of S-CNN features**. Feature embedding visualizations of S-CNN RGB and best 3 bands features on both datasets. We can observe that the S-CNN best 3 bands are semantically separated better compared to the S-CNN RGB image features. Each class (subject) has a separate color for the map points.

References

[1] C. Balas, V. Papadakis, N. Papadakis, A. Papadakis, E. Vazgiouraki, and G. Themelis. A novel hyper-spectral imaging apparatus for the non-destructive analysis of objects of artistic and historic value. *Journal of Cultural Heritage*, 2003.

[2] S. Bianco. Can linear data projection improve hyperspectral face recognition? In *CCIW*, 2015.

[3] M. Brown and S. Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR*, 2011.

[4] C. Cariou, K. Chehdi, and S. L. Moan. Bandclust: an unsupervised band reduction method for hyperspectral remote sensing. *GRSL*, 2011.

[5] C.-I. Chang. An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *Transactions on Information Theory*, 2000.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[7] P. Chavez, S. C. Sides, and J. A. Anderson. Comparison of three different methods to merge multiresolution and multispectral data- landsat tm and spot panchromatic. *Photogrammetric Engineering and remote sensing*, 1991.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] L. J. Denes, P. Metes, and Y. Liu. *Hyperspectral face database*. Carnegie Mellon University, 2002.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 2009.

[11] W. Di, L. Zhang, D. Zhang, and Q. Pan. Studies on hyperspectral face recognition in visible spectrum with feature band selection. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 2010.

[12] K. Fukunaga. *Introduction to statistical pattern recognition*. 2013.

[13] R. M. Gray. Vector quantization. *ASSP Magazine*, 1984.

[14] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015.

[15] G. P. Hughes. On the mean accuracy of statistical pattern recognizers. *Transactions on Information Theory*, 1968.

[16] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR Workshops*, 2015.

[17] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *PAMI*, 1997.

[18] I. Jolliffe. *Principal component analysis*. 2002.

[19] J. Kittler et al. Feature set search algorithms. *Pattern recognition and signal processing*, 1978.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[21] D. A. Landgrebe. *Signal theory methods in multispectral remote sensing*. John Wiley & Sons, 2005.

[22] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *NIPS*, 1990.

[23] X. Li, L. Wang, and E. Sung. Adaboost with svm-based component classifiers. *Engineering Applications of Artificial Intelligence*, 2008.

[24] J. Liang, J. Zhou, X. Bai, and Y. Qian. Salient object detection in hyperspectral imagery. In *ICIP*, 2013.

[25] J. Liang, J. Zhou, and Y. Gao. 3d local derivative pattern for hyperspectral face recognition. In *FGR*, 2015.

[26] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[27] Y. Luo, J. Remillard, and D. Hoetzer. Pedestrian detection in near-infrared night vision system. In *Intelligent Vehicles Symposium*, 2010.

[28] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[29] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *Computers, IEEE Transactions on*, 1977.

[30] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 2002.

[31] Z. Pan, G. Healey, M. Prasad, and B. Tromberg. Face recognition in hyperspectral images. *PAMI*, 2003.

[32] Philips. Microdose mammography si. new spectral benefits, proven low dose. 2016.

[33] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 1994.

[34] P. Ramzi, F. Samadzadegan, and P. Reinartz. Classification of hyperspectral data using an adaboostsvm technique applied on band clusters. *J-STARS*, 2014.

[35] J. A. Richards and J. Richards. *Remote sensing digital image analysis*. Springer, 1999.

[36] N. Salamati, C. Fredembach, and S. Süsstrunk. Material classification using color and nir images. In *CIC*, 2009.

[37] S. B. Serpico and L. Bruzzone. A new search algorithm for feature selection in hyperspectral remote sensing images. *TGRS*, 2001.

[38] S. B. Serpico and G. Moser. Extraction of spectral channels from hyperspectral images for classification purposes. *TGRS*, 2007.

[39] V. Sharma and L. Van Gool. Image-level classification in hyperspectral images using feature descriptors, with application to face recognition. *arXiv preprint arXiv:1605.03428*, 2016.

[40] L. Shen and S. Zheng. Hyperspectral face recognition using 3d gabor wavelets. In *ICPR*, 2012.

[41] D. Slater and G. Healey. Material classification for 3d objects in aerial hyperspectral images. In *CVPR*, 1999.

[42] V. Slavkovikj, S. Verstockt, W. De Neve, S. Van Hoecke, and R. Van de Walle. Hyperspectral image classification with convolutional neural networks. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, 2015.

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[44] M. Uzair, A. Mahmood, and A. Mian. Hyperspectral face recognition using 3d-dct and partial least squares. In *BMVC*, 2013.

[45] M. Uzair, A. Mahmood, and A. Mian. Hyperspectral face recognition with spatiospectral information fusion and pls regression. *TIP*, 2015.

[46] G. Valentini and T. G. Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *JMLR*, 2004.

[47] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, 2010.

[48] A. Vedaldi and K. Lenc. Matconvnet-convolutional neural networks for matlab. *arXiv preprint arXiv:1412.4564*, 2014.

[49] J. Yue, W. Zhao, S. Mao, and H. Liu. Spectral-spatial classification of hyperspectral images using deep convolutional neural networks. *Remote Sensing Letters*, 2015.

[50] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin. Classification of histology sections via multispectral convolutional sparse coding. In *CVPR*, 2014.

[51] Z. Zhu and Q. Ji. Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. *CVIU*, 2005.

[52] A. Zia, J. Liang, J. Zhou, and Y. Gao. 3d reconstruction from hyperspectral images. In *WACV*, 2015.

Supplementary Material: Hyperspectral CNN for Image Classification & Band Selection, with Application to Face Recognition

Vivek Sharma[◇], Ali Diba[◇], Tinne Tuytelaars[◇], and Luc Van Gool[◇∓]
[◇]KU Leuven, ESAT-PSI, iMinds [∓] BIWI, CVL, ETH Zürich
{firstname.lastname}@esat.kuleuven.be

1. Supplementary Material

1.1. Pipeline

Fig. 9 shows an overview of the proposed discriminative band selection technique in the hyperspectral images, as discussed in Section 3.2.

1.2. Standard Image Representation of HSI

In this section, we explain the conversion of HSI to RGB and panchromatic images. We obtain these images since they are the standard image representation of HSI. Also, we use these images as baselines for comparison.

1.2.1 Conversion of HSI to RGB image

In order to apply existing classification methods on the RGB color space of the hyperspectral image, we transform each hyperspectral image in to a 3-channel RGB image using CIE 2006 tristimulus color matching functions [2]. Now, we describe the steps in details.

Let $I(\lambda)$ be the spectral power distribution of the CIE standard daylight illuminant (D_{65}). $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ are the color matching functions for CIE 2006, that represent the human vision [3], and $R(\lambda)$ is the spectral reflectance of the object surface (Fig. 7). The CIE tristimulus values X, Y, and Z for each spatial location (m,n) of a reflecting object can be obtained by

$$X(m,n) = k\sum_{i=1}^{\eta^{'}} I(\lambda_i)R(m,n,\lambda_i)\bar{r}(\lambda_i),$$

$$Y(m,n) = k\sum_{i=1}^{\eta^{'}} I(\lambda_i)R(m,n,\lambda_i)\bar{y}(\lambda_i),$$

$$Z(m,n) = k\sum_{i=1}^{\eta^{'}} I(\lambda_i)R(m,n,\lambda_i)\bar{z}(\lambda_i)$$

where k is a normalization factor given by $100/\sum_{i=1}^{\eta^{'}} I(\lambda_i)\bar{y}(\lambda_i)$, and $\eta^{'}$ is the number of bands in the visible range (380-700nm) in the hyperspectral cube. Similar expressions are used for the computation of $Y(m,n)$ and $Z(m,n)$, by replacing $\bar{r}(\lambda_i)$ with $\bar{g}(\lambda_i)$ and $\bar{b}(\lambda_i)$ respectively. The CIE X, Y, and Z tristimulus values are transformed into sRGB color space (Fig. 4) using

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.2404542 & -1.5371385 & -0.4985314 \\ -0.9692660 & 1.8760108 & 0.0415560 \\ 0.0556434 & -0.2040259 & 1.0572252 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

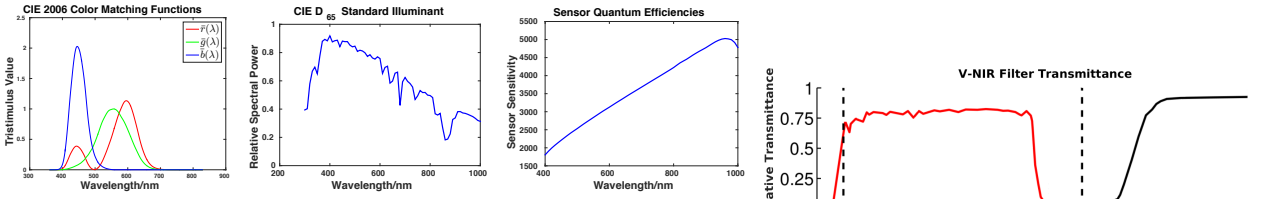
This is how we obtain a RGB image from a hyperspectral image.

1.2.2 Conversion of HSI to Panchromatic Image

In a conventional imaging system, a color filter array is placed in front of the sensor. This array limits the sensitivity of each photocell to a certain range of the spectrum. Such that, at each pixel of the sensor output image the object characteristic in only one channel is captured. Such a system is a panchromatic imaging system [1] (see Fig. 8). We simulate a panchromatic imaging system which is sensitive over a V-NIR range of approximately 440-1000nm and records the total intensity of radiance falling on each pixel.

Let $I(\lambda)$ be the spectral power distribution of the CIE standard daylight illuminant (D_{65}), $R(\lambda)$ is the spectral reflectance of the object surface, spectral transmittance of the V-NIR filter response by $T(\lambda)$, and spectral response of the silicon sensor by $S(\lambda)$ (see Fig. 7). We capture a single exposure using V-NIR filters for modulating wavelengths in 440-1000nm ($\eta^{''}$). The panchromatic camera sensor response for the V-NIR capture for each spatial location (m,n) (Fig. 4) of a reflecting object is obtained by

$$\rho(m,n) = \sum_{i=1}^{\eta^{''}} I(\lambda_i)R(m,n,\lambda_i)S(\lambda_i)T(\lambda_i)$$



(a) The CIE 2006, Color Matching Functions

(b) CIE standard daylight illuminant (D_{65}), $I(\lambda)$

(c) Silicon sensitivity of Hamamatsu camera, $S(\lambda)$

(d) Visible and Near-Infrared (V-NIR) filters, $T(\lambda)$

Figure 7: Characteristic curves of the different components of hyperspectral imaging system for the image acquisition.

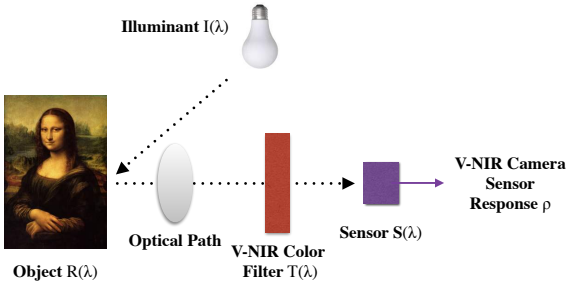


Figure 8: Components of a panchromatic image acquisition system.

References

[1] P. Chavez, S. C. Sides, and J. A. Anderson. Comparison of three different methods to merge multiresolution and multispectral data- landsat tm and spot panchromatic. *Photogrammetric Engineering and remote sensing*, 1991.

[2] CIE-CMFs. Cie 2006 color matching functions. `http://cvrl.ioo.ucl.ac.uk/cmfs.htm`. 2016.

[3] N. P. Jacobson and M. R. Gupta. Design goals and solutions for display of hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions on*, 43, 2005.

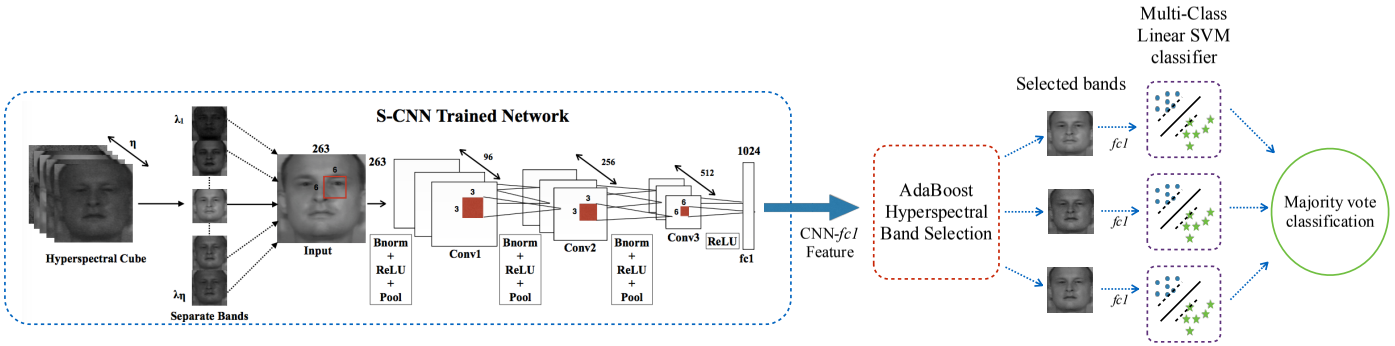


Figure 9: Pipeline of our proposed image-level band selection in the hyperspectral images using S-CNN features using AdaBoost-SVM weighted majority voting, as discussed in Section 3.2