On Effectiveness of Histogram of Oriented Gradient Features for Visible to Near Infrared Face Matching

Tejas Indulal Dhamecha*, Praneet Sharma*, Richa Singh, and Mayank Vatsa
IIIT-Delhi, India
Email: {tejasd, praneet10061, rsingh, mayank}@iiitd.ac.in

***Abstract*—The advent of near infrared imagery and it's applications in face recognition has instigated research in cross spectral (visible to near infrared) matching. Existing research has focused on extracting textural features including variants of histogram of oriented gradients. This paper focuses on studying the effectiveness of these features for cross spectral face recognition. On NIR-VIS-2.0 cross spectral face database, three HOG variants are analyzed along with dimensionality reduction approaches and linear discriminant analysis. The results demonstrate that DSIFT with subspace LDA outperforms a commercial matcher and other HOG variants by at least 15%. We also observe that histogram of oriented gradient features are able to encode similar facial features across spectrums.**

Keywords*—*HOG Descriptor, Cross Spectral Face Recognition

I. Introduction

Traditional face recognition utilizes visible spectrum images captured in controlled environment with cooperative users. Several algorithms have been able to achieve very high accuracy in these conditions [1], [2]. On the other hand, uncontrolled environment with variations due to pose, illumination, expression and occlusion is still a research challenge and researchers are attempting to enhance the state-of-art. To mitigate the effect of illumination variation, Li *et al.* [3] proposed to utilize near infrared (NIR) spectrum. The efficacy of NIR spectrum based face recognition systems is based on the fact that faces captured in NIR spectrum do not vary much with changes in illumination. However, the use of NIR spectrums introduced few new research directions. One such direction is matching an NIR spectrum image with a visible spectrum (VIS) image. Such a requirement arises when the gallery and probe images are not captured in the same spectrum. For instance, in law enforcement applications, the gallery image might be a VIS mugshot image whereas the probe can be a NIR image captured using surveillance cameras. This research challenge is known as cross-spectral face matching and often, it is encompassed under broader research of heterogeneous face matching [4].

As shown in Fig. 1, VIS and NIR images are acquired in different bands of electro-magnetic spectrum and information captured in the two spectrums are not same. Since VIS and NIR spectrums are adjacent, human skin has similar response to incident lights of the two spectrums. VIS image captures face texture information whereas NIR image is not very rich in texture, since NIR light penetrates into skin tissues [5]. These fundamental differences in the acquired information introduces the challenge of data heterogeneity.

*Equal contributions from student authors.

Fig. 1. Top row contains VIS images and bottom row contains NIR images, captured in multiple sessions with varying illumination and facial expressions.

A. Literature Review

Data heterogeneity in face recognition can be addressed either by finding spectrum invariant features or by modeling the heterogeneity. In recent past, significant research has been carried out for NIR to VIS face matching. Almost all the existing approaches can be broadly categorized into two approaches - synthesis and direct matching. Synthesis approaches try to model the mapping between VIS and NIR face images, thus bringing images of both the spectrums to same/pseudo same spectrum. On the other hand, direct matching approaches largely rely on using spectrum invariant features.

Synthesis approaches: NIR image is first synthesized from VIS image or vice versa followed by matching the synthesized NIR and original NIR images, or synthesized VIS and original VIS images. Chen *et al.* [6] proposed a transformation approach for synthesizing VIS images from NIR images by reducing the intra-personal differences. Similarly, Wang *et al.* [5] proposed an analysis-by-synthesis approach. Lei *et al.* [7] proposed coupled spectral regression wherein graph embedding framework is used to first represent every face and then two different learned projections are used to project the NIR and VIS images on a discriminative common subspace for classification. Zhu et al. [8] introduced a transductive framework which is based on learning a subspace such that modality differences are removed but intra-domain features are preserved.

Direct matching approaches: Yi *et al.* [9] used canonical correlation analysis based learning in linear discriminant analysis (LDA) subspace for matching. Liao *et al.* [10] utilized multi-scale block local binary patterns as the feature set with regularized LDA classifier. Klare *et al.* [11] extracted local binary pattern (LBP) and histogram of oriented gradients (HOG) features from VIS and NIR images. Random subspaces based

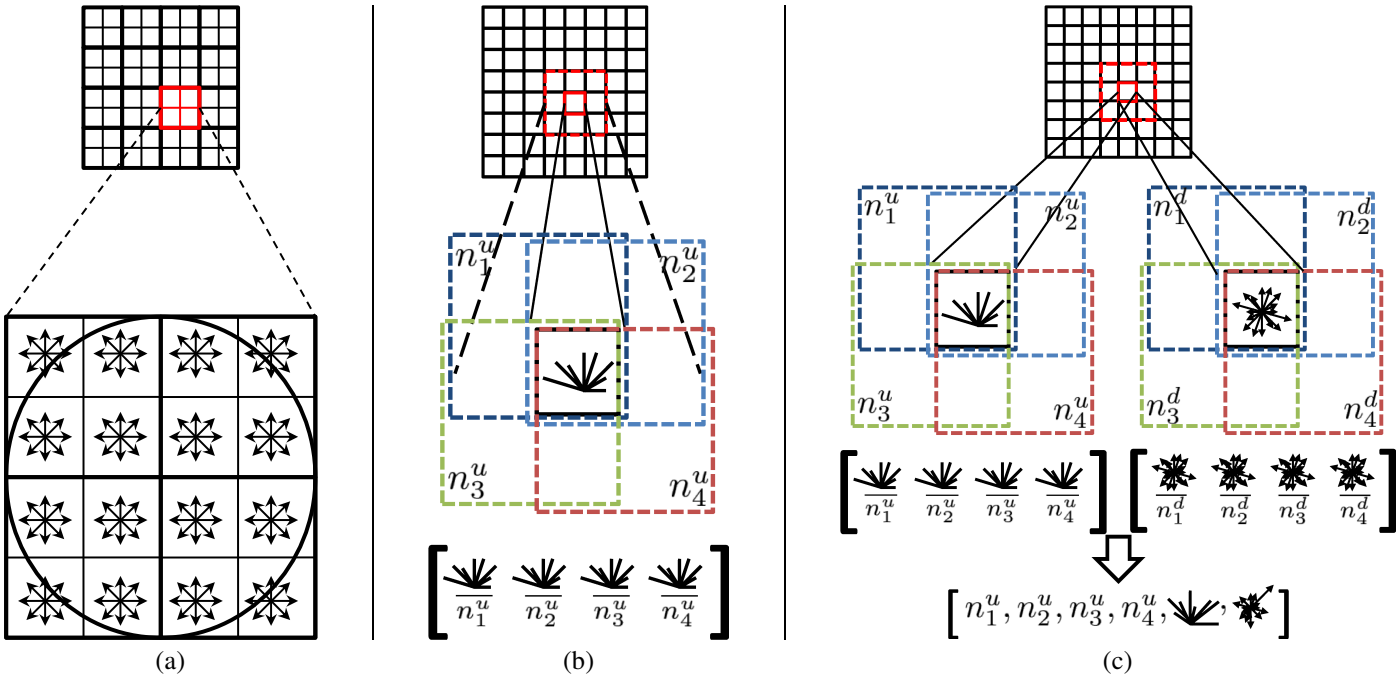


Fig. 2. Illustrating the computation of (a) DSIFT, (b) Dalal-Triggs variant of HOG, and (c) UoCTTI variant of HOG.

ensemble of classifier is utilized along with nearest neighbor (NN) and sparse representation based matching. Similarly, Maeng *et al.* [12] utilized HOG features for cross-spectrum and cross-distance face matching. Most of these algorithms are evaluated on small scale datasets, such as heterogeneous face biometrics (HFB) dataset [13] and CARL[14] which comprise limited number of subjects and/or undefined experimental protocols. Therefore, claims regarding generalizability of performance may not be made confidently and benchmarking can be challenging.

B. Research Contributions

Although HOG features have been explored for VIS to NIR matching, to the best of our knowledge, a study analyzing the effectiveness of different HOG variants, along with different classifiers and distance metrics is not yet performed. Since the overall performance of a recognition system is a function of the combination of feature and classifier, such a study may unravel the best suited framework. It would also help understand the nature of these features in context of VIS to NIR face matching. NIR-VIS 2.0 dataset [15], a recently created dataset, addresses the issues of fixed protocol and dataset size. It is our assertion that this dataset may be more helpful in understanding and/or solving the problem of VIS to NIR face matching. Therefore, in this research, all the experiments are performed on the NIR-VIS 2.0 dataset. The key contributions of this work are:

- Evaluation and performance analysis of three HOG variants, namely
  - Dense Scale Invariant Feature Transform (DSIFT) [16],
  - Dalal-Triggs HOG (HOG-DT) [17], and
  - HOG-UoCTTI [18],

  along with classification by LDA and direct feature matching.
- Studying the role of dimensionality reduction and distance metrics.
- Understanding the difficulty of the problem by establishing baseline performance using a commercial-off-the-shelf system.

II. Variants of Histogram of Oriented Gradient Features

HOG features have been widely used for object detection [16]. The basic intuition that an object can be effectively represented using edge orientations is the base of HOG feature based object detection approaches. There are many variants of HOG descriptors such as DSIFT [16], HOG-DT [17], and HOG-UoCTTI [18]. All the variants involved computing image gradient orientations. The image is tessellated into cells and for each cell gradient orientations are computed and binned into discrete orientation bins. These binned orientations are weighted and/or normalized to obtain the final feature descriptor. The details of the three aforementioned variants are as follows:

1) DSIFT or Classical HOG [16]: As the name suggests, DSIFT computes SIFT features [16] on a dense grid of fixed key points. As shown in Fig. 2(a), 4×4 grid of cells is used keeping the key point in the center of the grid. Histogram of directed gradient orientation are computed for each cell, which are weighted using a Gaussian function centered at the key point. The final descriptor is obtained by stacking the weighted histograms of each cell. If the number of orientations is set as n_o, the gradient directions are uniformly sampled in $[0, 2\pi)$, and the final descriptor of every key point is of length $16n_o$.

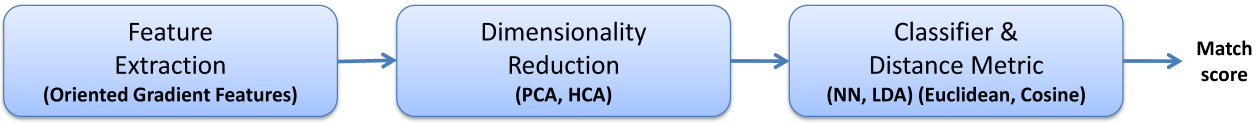


Fig. 3. Illustrating the stages involved in the proposed evaluation framework.

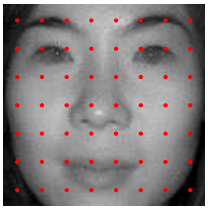


Fig. 4. Illustrating the 8×8 grid of key points on a face image.

2) Dalal-Triggs HOG [17]: As shown in Fig. 2(b), HOG-DT employs undirected gradient orientations and for every cell, binning is performed in $2n_o$ directions uniformly sampled in $[0, 2\pi)$, which is folded into two to get n_o dimensional histogram. A grid of 2×2 cells defines a block, where each block has two overlapping cells with adjacent horizontal and vertical blocks. Thus, each cell is shared by four blocks (represented with different colored squares in Fig. 2(b)). l_2-norm of each block is used as the normalization factor. Thus, for every cell four normalization factors are obtained from the four blocks which share the specified cell. The final descriptor of each cell is defined as stacking of four copies of the cell histograms, each normalized with one normalization factor. Thus, the final descriptor of every cell is of length $4n_o$.

3) HOG-UoCTTI [18]: The HOG-HoCTTI variant performs similar operations as HOG-DT with and without folding the cell histogram into two. Thus, it employs both undirected and directed gradients. As shown in Fig. 2(c), four histograms of undirected gradients are averaged to obtain an n_o dimensional histogram. Similar operation is performed for directed gradients to obtain an $2n_o$ dimensional histogram. The final descriptor is obtained by stacking the averaged directed histogram, averaged undirected histogram, and four normalization factors pertaining to undirected histograms. This leads to the final descriptor of size $4 + 3 \times n_o$.

III. Evaluation Framework

As shown in Fig. 3, the evaluation framework for matching VIS and NIR images involves feature extraction, followed by an optional stage of dimensionality reduction and classification. Evaluation is performed with the three variants of HOG and raw pixel intensity features. The reduced dimensional representation of features is achieved either by principal component analysis (PCA) [19] or hetero-component analysis (HCA) [15]. The dimensionality reduced features are classified using nearest neighbor (NN) and LDA with Euclidean and Cosine distance metrics. This section provides details about each component of the matching pipeline.

A. Feature Extraction

As illustrated in Fig. 4, fixed key points based approach is used to extract oriented gradient features. Let the image size be $H \times W$, a grid of $m \times m$ key points is defined on an image by specifying the step-size to cell-size ratio of $\frac{W}{m}$ along the width and $\frac{H}{m}$ along the height of the image. For the selected key points, features are extracted which contain magnitude and gradient. For every key point, these are represented in a single vector of length d. For classical HOG $d = 128$, for HOG-DT $d = 36$, and for HOG-UoCTTI $d = 31$ (n_o is set as 8, 9, and 9, respectively and $m = 8$ leads to a total of 64 key points located on a uniform grid). For every key point, HOG is then stacked to generate the final feature descriptor for a given face image and it is denoted by f. Size of the extracted feature descriptor is $(m \times m) \times d$. If the shape of the object is not affected in the images captured in two different spectrums, the edge orientation related features are spectrum invariant. As shown in Fig. 5, if the shape of the input face is same, the edge orientation information encoded using HOG is similar for VIS and NIR face images for the same subject. Therefore, HOG features make a suitable choice for matching VIS and NIR face images.

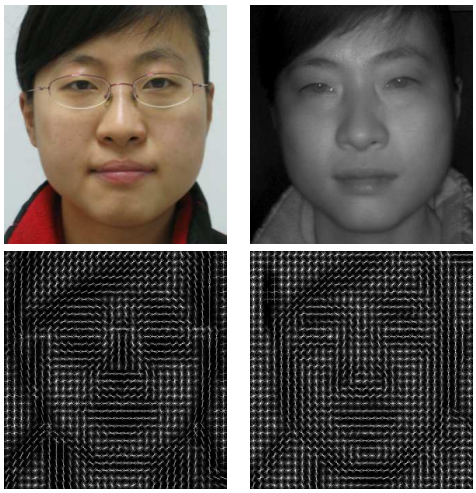


Fig. 5. A VIS (top left) and NIR (top right) images of an individual, and visualization of their corresponding HOG features are shown in bottom row.

B. Dimensionality Reduction

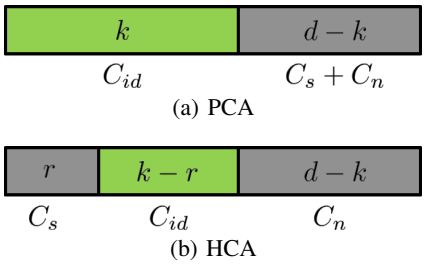
Dimensionality of the feature descriptor f is $(m \times m) \times d$. Generally, the size of this descriptor is 2,000-10,000 which is a significantly large feature vector size with the training set of approximately 2000 samples. Therefore, it may be advisable to perform dimensionality reduction prior to matching. PCA [19] and HCA [15] are chosen as two candidate dimensionality reduction algorithms.

Let F_i be the vectorized version of feature descriptor of the i^{th} training image. If feature descriptor of the i^{th} image has dimensions $(m \times m) \times d$ then F_i, a column vector, has

dimensions $(m \times m \times d) \times 1$. Let the size of the training set be N, which contains a mix of NIR and VIS images. The feature matrix of the complete training set is of size $N \times (m \times m \times d)$ and every row corresponds to the feature vector of a training image. In order to apply PCA, the covariance matrix is computed as follows,

$$C = \frac{1}{N} \sum_{i=1}^{N} (F_i - \mu)^T (F_i - \mu) \quad (1)$$

where $\mu = \frac{1}{N} \sum_{i=1}^{N} F_i$ is the mean feature vector. Eigenvalue decomposition of covariance matrix is performed to obtain its eigenvectors and eigenvalues. As illustrated in Fig. 6, PCA subspace is the vector space spanned by k eigenvectors corresponding to the largest eigenvalues. Out of the top-k vectors, HCA [15] discards top-r eigenvectors as they may not represent information pertaining to identity. Thus, the HCA subspace is spanned by $k - r$ vectors. Let the total variance consist of three components, C_{id}, C_s, and C_n. Here, C_{id} represents the variation due to the presence of multiple subjects, C_s represents the variation due to two spectrums, and C_n represents the variation due to noisy or un-useful components. The rationale for selecting k components in PCA is the assumption that top components represent C_{id}, whereas HCA assumes that $k - r$ components obtained after rejecting the top r components represent the same information. k is chosen such that top-k eigenvectors represent 99% of the total eigenenergy. Usually, r is chosen empirically from the development set. To obtain the reduced dimensional representation of a feature descriptor, it is mean normalized and projected into the respective subspace. The outcome of this step is a dimensionality reduced feature descriptor for each image which is a very close approximation of the original feature vector.



(a) PCA

(b) HCA

Fig. 6. Illustrating PCA and HCA in terms of the components used. d is the total number of components. k is the number of components used in PCA and r is the number of components rejected in HCA.

C. Classifier and Distance Metrics

Matching in feature space and in learned discriminative space is performed using nearest neighbor and LDA. The projections of the gallery and probe features are matched using two distance metrics, namely euclidean and cosine distance. Let x and y be the two projections to be matched, the euclidean distance is defined as

$$E(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

and the cosine distances is defined as

$$CosDist(x, y) = 1 - \frac{xy^T}{\sqrt{xx^T}\sqrt{yy^T}}.$$

IV. Experiments and Analysis

Following the evaluation framework shown in Fig. 3, NIR-VIS 2.0 dataset is used for performance evaluation. Three sets of experiments are performed: (1) establish the baseline performance using FaceVacs which is a commercial-off-the-shelf (COTS) system, (2) evaluate the effectiveness of HOG variants, classifiers, dimensionality reduction techniques, and distance metrics, and (3) to compare the effectiveness of HCA with PCA.

A. Database and Protocol

NIR-VIS 2.0 dataset [15] consists of 17,850 NIR and VIS images (combined) pertaining to 725 subjects of varying age groups. The images were acquired in four different sessions with a resolution of 128×128. The protocols defined for performance evaluations consist of two views. View 1 is for development of solution/algorithm and selection of optimum parameters. It consists of train development and test development sets. View 2 is for reporting performance, which may also require optimizing parameters. View 2 consists of 10 splits of train and test sets for random subsampling cross validation. In both the views, there are equal number of subjects in train and test sets. The performance is reported in terms of rank-1 identification accuracy and cumulative match characteristic (CMC) curves are shown. Mean and standard deviation over the 10 folds are also reported. Face images used for evaluation are already cropped and registered. Hence, no further preprocessing is performed on the face images.

B. Analysis

The baseline algorithm by Li *et al.* [15] utilizes PCA with cosine metric as the similarity measure. The face symmetry is leveraged along with application of HCA to further enhance the performance. As shown in Table I, this leads to rank-1 accuracy of 23.70% and standard deviation of 1.89%. On this dataset, this has been the state-of-the-art accuracy till now. To understand the performance of existing commercial matchers, identification performance is computed with FaceVacs. Though the performance of FaceVacs is significantly better than existing baseline, there is a significant scope of further improvement.

TABLE I. Baseline Rank-1 identification accuracies (%).

Classifier	Accuracy
PCA + Symmetry [15]	09.26±0.66
PCA + Symmetry + HCA [15]	23.70±1.89
FaceVacs	**58.56±1.19**

Table II and Fig. 7 report the rank-1 identification accuracies and CMC curves of various combinations of features, dimensionality reduction techniques, and classifiers in the proposed evaluation framework. The key observations are as follows.

- The best performance is achieved by DSIFT with subspace LDA i.e. PCA+LDA and cosine distance

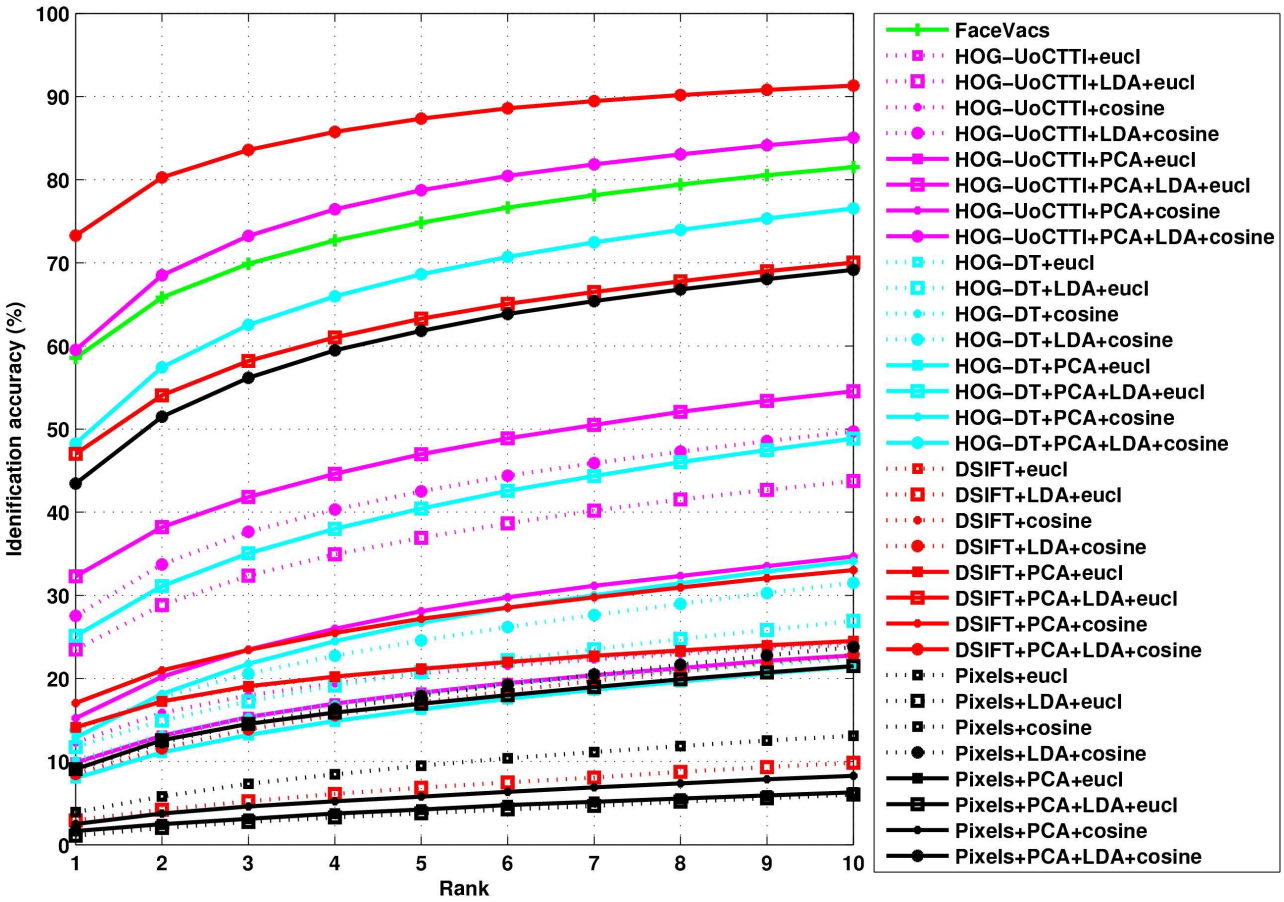


Fig. 7. CMC plots illustrating the performance of different combinations of features, dimensionality reduction techniques, and classifiers in the proposed framework.

metric. This even outperforms the FaceVacs matcher with a significant margin of ~15%.

- Consistently across all the features, performing dimensionality reduction with PCA either results in significant improvement or at least maintains the recognition accuracy (with exception in only one case).
- The performances of DSIFT, HOG-UoCTTI, and HOG-DT reveal that the later lags in the comparison of formers. Note that, DSIFT and HOG-UoCTTI utilize directed gradient whereas HOG-DT utilizes undirected gradient. The performance difference may be attributed to this property and for VIS to NIR face matching directed gradients are more effective than undirected gradients.
- Classification of raw DSIFT features with LDA yields poor results than with NN, however reduced dimensional representations surpass FaceVacs. This may be due to the small sample size (SSS) problem as the number of samples in training (number = ~6100) is less than the raw feature length of DSIFT (dim = 8192). These results are obtained when regularization is performed to address singularity of within-class scatter which may arise due to the SSS problem.
- Consistently in all the experiments cosine metric outperformed euclidean distance. This suggests that for the proposed VIS to NIR matching framework, cosine distance may be a better distance metric.
- We have also performed experiments with LBP fea-

TABLE II. RANK-1 IDENTIFICATION ACCURACY OF PIXEL INTENSITY AND HOG FEATURES IN COMBINATION WITH DIMENSIONALITY REDUCTION TECHNIQUES, CLASSIFIERS, AND DISTANCE MEASURES.

		Pixel Intensity		DSIFT		HOG-DT		HOG-UoCTTI	
		Raw	PCA	Raw	PCA	Raw	PCA	Raw	PCA
NN	Euclidean	1.64±0.31	1.64±0.24	14.11±1.09	14.12±1.08	8.00±1.22	8.00±1.21	9.78±1.66	9.86±1.62
	Cosine	3.92±0.79	2.43±0.41	14.10±1.08	17.04±1.03	10.17±1.28	12.89±1.48	12.30±1.42	15.21±1.59
LDA	Euclidean	1.09±0.44	9.10±2.21	2.93±0.58	47.01±1.99	11.73±1.46	25.14±2.33	23.49±1.88	32.28±1.23
	Cosine	9.04± 0.63	43.44±1.59	8.52±1.26	**73.28±1.10**	13.84±1.83	48.27±1.43	27.52±1.62	59.51±1.35

tures and the results with HOG features seem to be more appropriate than LBP for visible to near-infrared face matching. Moreover, in our experiments on NIR-VIS-2.0 database, LBP does not seem to add value in feature level fusion as well.

- Note that, in Li *et al.* [15], baseline algorithm is different than in this paper which may explain the difference in accuracy between both. In this research, PCA is applied to preserve 99% eigenenergy whereas in [15] PCA preserves 98% eigenenergy with an additional unit length normalization on samples.

Since Li *et al.* [15] observe HCA to be more useful than PCA, we investigate HCA for the best performing combination of DSIFT with subspace LDA and cosine distance metric. The results are reported in Table III. It has been observed that dimensionality reduction using HCA degrades the performance compared to PCA. The key intuition behind HCA is that in case of heterogeneous data samples, the largest variation in total variance is due to spectrum differences. These variations will be even more than the actual inter-class variabilities which provide discriminative information.

Table III shows the effect of using HCA by replacing PCA in the best performing combination of DSIFT+PCA+LDA. Further, HCA with raw pixel intensity values is also evaluated. Similar to Li *et al.* [15], in nearest neighbor matching, the accuracy improves when HCA is employed instead of PCA. However, when matching is performed in the discriminant space learned by LDA, HCA leads to significant performance reduction compared to PCA. It shows that LDA benefits from the top components which are otherwise strictly rejected in HCA. It also means that the top components contain information useful for identification; however simple nearest neighbor (without LDA) is not able to leverage that information.

TABLE III. COMPARING RANK-1 IDENTIFICATION ACCURACIES (%) OBTAINED BY USING HCA AND PCA FOR DIMENSIONALITY REDUCTION.

Dimensionality Reduction		LDA		NN	
		DSIFT	Intensity	DSIFT	Intensity
PCA	Euclidean	47.01±1.99	9.10±2.21	14.12±1.09	1.64±0.24
	Cosine	73.28±1.10	43.44±1.59	17.04±1.03	2.43±0.41
HCA	Euclidean	29.82±2.08	5.77±1.91	23.64±1.58	6.39±1.99
	Cosine	58.43±1.17	34.57±1.49	38.64±1.44	21.97±1.47

V. CONCLUSION

This paper presents an evaluation framework to analyze the effectiveness of HOG features for cross spectral (visible to near infrared) face recognition. In conjunction with feature space, PCA and HCA subspaces as well as LDA, a detailed analysis is performed on the NIR-VIS-2.0 database. Experimental results show that DSIFT+PCA+LDA+Cosine distance achieves the rank-1 identification accuracy of 73.28% and outperforms state-of-the-art by a significant margin of ~50% and COTS by ~15%. It is also observed that directed orientation is more useful for recognition than undirected orientations. In future, we plan to extend the framework with distance metric learning and domain adaptation approaches to further improve the accuracy.

VI. ACKNOWLEDGEMENT

The authors would like to thank Prof. Stan Li for sharing the database used in this research. Dhamecha is partly supported through TCS research fellowship.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the evaluation of 2D still-image face recognition algorithms," *NIST Interagency Rep*, no. 7709, 2010.

[3] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627–639, 2007.

[4] B. Klare and A. Jain, "Heterogeneous face recognition using kernel prototype similarities." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, p. 1410, 2013.

[5] R. Wang, J. Yang, D. Yi, and S. Z. Li, "An analysis-by-synthesis method for heterogeneous face biometrics," in *Advances in Biometrics*, 2009, pp. 319–326.

[6] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikainen, "Learning mappings for face synthesis from near infrared to visual light images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 156–163.

[7] J.-Y. Zhu, W.-S. Zheng, and J. Lai, "Transductive VIS-NIR face matching," in *Proceedings of IEEE International Conference on Image Processing*, 2012, pp. 1437–1440.

[8] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1123–1128.

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

[10] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Advances in Biometrics*, 2009, pp. 209–218.

[11] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching nir to visible light images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1513–1516.

[12] H. Maeng, S. Liao, D. Kang, S.-W. Lee, and A. K. Jain, "Nighttime face recognition at long distance: cross-distance and cross-spectral matching," in *Proceedings of Asian Conference on Computer Vision*, 2013, pp. 708–721.

[13] S. Z. Li, Z. Lei, and M. Ao, "The HFB face database for heterogeneous face biometrics research," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2009, pp. 1–8.

[14] V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognitive Computation*, vol. 5, no. 1, pp. 119–135, 2013.

[15] S. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2013, pp. 348–353.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[19] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.