GENERALIZED BILINEAR DEEP CONVOLUTIONAL NEURAL NETWORKS FOR MULTIMODAL BIOMETRIC IDENTIFICATION

Sobhan Soleymani, Amirsina Torfi, Jeremy Dawson, and Nasser M. Nasrabadi, Fellow, IEEE

West Virginia University

ABSTRACT

In this paper, we propose to employ a bank of modality-dedicated Convolutional Neural Networks (CNNs), fuse, train, and optimize them together for person classification tasks. A modality-dedicated CNN is used for each modality to extract modality-specific features. We demonstrate that, rather than spatial fusion at the convolutional layers, the fusion can be performed on the outputs of the fully-connected layers of the modality-specific CNNs without any loss of performance and with significant reduction in the number of parameters. We show that, using multiple CNNs with multimodal fusion at the feature-level, we significantly outperform systems that use unimodal representation. We study weighted feature, bilinear, and compact bilinear feature-level fusion algorithms for multimodal biometric person identification. Finally, We propose generalized compact bilinear fusion algorithm to deploy both the weighted feature fusion and compact bilinear schemes. We provide the results for the proposed algorithms on three challenging databases: CMU Multi-PIE, BioCop, and BIOMDATA.

***Index Terms*—** Biometrics, multimodal fusion, tensor sketch, compact bilinear pooling.

1. INTRODUCTION

The permanence and uniqueness of human physical characteristics such as face, iris, fingerprint, and voice is widely utilized in biometric systems deploying the corresponding feature representation of these characteristics [1]. Multimodal biometric models have demonstrated more robustness to noisy data, non-universality and category-based variations [2, 3]. The multimodal networks can improve recognition task in cases where one or more of the biometric traits are distorted. A recognition algorithm using a multimodal architecture, requires selecting the discriminative and informative features from each modality as well as exploring the dependencies between different modalities. This architecture should also discard the single modality features that are not useful in joint recognition.

However, employing a fusion algorithm is the most prominent challenge in multimodal biometric systems [4]. The fusion algorithm can be performed at signal, feature, score, rank or decision levels [5] using different schemes such as feature concatenation [6, 7, 8] and bilinear feature multiplication [9, 10]. Although score-, rank- and decision-level fusion are studied in the literature extensively, since these levels are easier to access in the biometric systems, feature-level fusion results in a better discriminative classifier [11] due to the preservation of raw information [1]. Feature level fusion integrates different features extracted from different modalities to a more abstract feature representation, which can further be used for classification, verification, or identification [12].

To integrate the features from different modalities, several fusion methods have been considered [6]. The prevalent fusion method in the literature is feature concatenation, which is very inefficient exploiting the dependency between the modalities as the feature space dimensionality increases [4, 7]. To overcome this shortcoming, bilinear multiplication of the individual modalities is proposed [9, 10]. Using bilinear multiplication, the higher-level dependencies between the modalities are exploited and enforced through the backpropagation algorithm. The bilinear multiplication is effective since all of the elements of the single modalities interact through multiplication. The main issue in bilinear operation is the high dimensionality of its output regarding the cardinality of the inputs. Recently, to handle this shortcoming, compact bilinear pooling is proposed [13, 14, 15].

Convolutional neural networks are recently utilized for classification of multimodal biometric data. Although, CNNs are mainly used as classifiers, they are also efficient tools to extract and represent discriminative features from the raw data. Compared to hand-crafted features, employing CNN as domain feature extractors has demonstrated to be more promising when facing different biometric modalities such as face [16, 17], iris [18] and fingerprint [19].

In this paper, we make the following contributions: (i) instead of spatial fusion at the convolutional layers, modality-dedicated networks are designed to extract modality-specific features for the fusion; (ii) a fully data-driven architecture using fused CNNs and end-to-end joint optimization of the overall network, is proposed for joint domain-specific feature extraction and representation with the application of person classification; finally (iii) weighted feature fusion and generalized compact bilinear feature fusion are considered at the fully-connected level.

2. GENERALIZED COMPACT BILINEAR FUSION

Consider a fusion operation $f : (X_1, X_2, ..., X_n) \rightarrow Y$ that fuses n modalities; $X_i \in R^{H_i \times W_i \times D_i}$, $i = 1, 2, .., n$. The fusion operation results in $Y \in R^{\widehat{H} \times \widehat{W} \times \widehat{D}}$, where W, H and D correspond to width, height and depth of the feature maps. Fusion can be performed using the feature maps of the CNNs when the corresponding feature maps from different modalities are compatible. However, in multimodal biometric networks, the feature maps can vary in the spatial dimension due to the different spatial dimensionality of the inputs. To handle this issue, instead of utilizing convolutional layers feature maps for fusion, fully-connected layers are considered in our architecture for ultimate modality-dedicated feature representation. Therefore, in our proposed architecture, $H_i = W_i = \widehat{H} = \widehat{W} = 1$, and there is no condition on D. We show that the fully-connected representation provides promising results in the case of recognition applications.

In the proposed fusion algorithm, prior to the fusion, each modality is represented by the output of a fully-connected layer which we call the modality-dedicated embedding layer. In **weighted feature fusion** algorithm, the fusion function concatenates the modality-dedicated embedding layers of the multiple modalities, in which $Y \in R^{1 \times \widehat{D}}$, where $\widehat{D} = \sum_i D_i$. In **bilinear fusion** algorithm, $Y = X_1^T X_2$. If $H_i \neq 1$, the outer product is applied on two feature maps at the pixel level, followed by global average pooling over the spatial dimensions [9, 10]. However, the bilinear fusion over fully-connected layers computes the outer product of the modality-dedicated embedding layers, where $Y \in R^{1 \times \tilde{D}}$ and $\widehat{D} = \prod_i D_i$. The resulting feature-level representation $Y \in R^{1 \times \widehat{D}}$, projects all possible feature-level interactions between the n modalities. In the case that the n is larger than two, in each step the outer product is vectorized and then multiplied by the next modality.

Generalized compact bilinear feature-level fusion algorithm: Compact bilinear fusion projects the outer product of two vectors into a low-dimensional sub-space with very little loss in performance compared to bilinear fusion [13]. Random Maclaurin projection and Tensor Sketch projection [13] are the most prominent algorithms proposed for compact bilinear pooling. Here, we deploy the tensor sketch projection. This algorithm uses the count sketch projection introduced in [20] to estimate the outer product of two vectors without computing the outer product explicitly. The count sketch of the outer product of two vectors can be expressed as the convolution of count sketches of the vectors [15]. However, this convolution can be computed as the inverse Fourier transform of the element-wise product of the count sketches in the frequency domain. Therefore, the bilinear outer product of multiple modalities can be computed through element-wise multiplication of Fourier domain count sketches. Let $\mathbf{x}_1 \in \mathbb{R}^{c_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{c_2}$ be the modality-dedicated embedding layers:

$$\mathbf{y} = \text{FFT}^{-1}(\text{FFT}(\Psi(\mathbf{x}_1, h_1, s_1)) \circ \text{FFT}(\Psi(\mathbf{x}_2, h_2, s_2))), \quad (1)$$

where hash functions $h_1 \in \mathbb{N}^{c_1}$ and $h_2 \in \mathbb{N}^{c_2}$ are random, but fixed vectors uniformly drawn from $\{1, 2, ..., d\}$, $s_1 \in \{-1, +1\}^{c_1}$, and $s_2 \in \{-1, +1\}^{c_2}$. The count sketch function is defined as:

$$\Psi(\mathbf{x}_1, h_1, s_1)) = \{(Qx_1)_1, (Qx_1)_2, ..., (Qx_1)_d\}, \quad (2)$$

where $(Qx_1)_j = \sum_{n:h_1[n]=j} s_1[n]\mathbf{x}_1[n]$. This algorithm can be expanded to fuse multiple modalities as well.

In the proposed generalized compact bilinear fusion algorithm, single modalities and all possible 2-,3-,..., n-compact bilinear products are concatenated to form vector $\mathbf{y}$. For instance, when $n = 3$, three modality-dedicated embedding layer, three two-modality tensor sketch projection, and one three-modality tensor sketch projection are concatenated.

End-to-end training of the architecture: Generalized compact bilinear fusion algorithm consists of random, but fixed functions $\{s_i\}$ and $\{h_i\}$, Fourier and inverse Fourier transforms. Since these transforms are differentiable, the error can be back-propagated through the fusion layer, the end-to-end training of the proposed generalized compact bilinear fusion algorithm is possible, and the multimodal architecture can be jointly optimized. For two-modality tensor sketch fusion algorithm, the error is back-propagated through the fusion layer using the equation below. Let L represent the loss function at the fusion layer [13]:

$$\frac{\partial L}{\partial \mathbf{x}_1} = \sum_d \frac{\partial L}{\partial \mathbf{y}[d]} T_d^2(\mathbf{x}_2) \circ s_1, \quad (3)$$

where ${T_d}^2(\mathbf{x}) \in \mathbb{R}^{c_1}$, ${T_d}^2(\mathbf{x}_2)[j] = \Psi(\mathbf{x}_2, h_2, s_2)[d - h_1[j]]$. Similarly, $\frac{\partial L}{\partial \mathbf{x}_2}$ can be calculated.

3. JOINT OPTIMIZATION OF ARCHITECTURE

The multimodal CNN architecture consists of modality-dedicated CNN networks, a joint representation layer, and a softmax classification layer that are jointly trained and optimized. The modality-dedicated networks are trained to extract the modality specific features and the joint representation is trained to explore and enforce dependency between different modalities. The joint optimization of the networks, discards the unuseful features.

Modality-dedicated networks: Each modality-dedicated CNN, consists of the first 16 layers of a conventional VGG19 network [21] and a fully-connected modality-dedicated embedding layer (FC6) of size 1024. The fully-connected layers of the conventional VGG19 network are not practical for our application, since the joint optimization of the modality-dedicated networks and the joint representation layer is practically impossible due to the massive number of parameters that need to be trained and the limited number of training samples. The details for each modality-dedicated network can be found in Table 1.

Joint representation layer: The output of the modality-dedicated networks are fused using one of the discussed fusion algorithm, then fed to a fully connected layer of size 1024 and finally, fed to the softmax classification layer.

network	CNN-Face	CNN-Iris	CNN-Fingerprint
input	$224 \times 224 \times 3$	$512 \times 64 \times 3$	$224 \times 224 \times 3$
layer	kernel	kernel	kernel
conv1 (1-2)	$3 \times 3 \times 64$	$3 \times 3 \times 64$	$3 \times 3 \times 64$
maxpool1	2×2	2×2	2×2
conv2 (1-2)	$3 \times 3 \times 128$	$3 \times 3 \times 128$	$3 \times 3 \times 128$
maxpool2	2×2	2×2	2×2
conv3 (1-4)	$3 \times 3 \times 256$	$3 \times 3 \times 256$	$3 \times 3 \times 256$
maxpool3	2×2	2×2	2×2
conv4 (1-4)	$3 \times 3 \times 512$	$3 \times 3 \times 512$	$3 \times 3 \times 512$
maxpool4	2×2	2×2	2×2
conv5 (1-4)	$3 \times 3 \times 512$	$3 \times 3 \times 512$	$3 \times 3 \times 512$
FC6	$7 \times 7 \times 1024$	$2 \times 16 \times 1024$	$7 \times 7 \times 1024$

Table 1: The modality-dedicated CNN architectures.

		Train set	Test set	KNN	SVM	CNN
BioCop	Face	6833	6960	89.68	88.76	98.14
	Iris	36636	39725	70.52	79.26	99.05
	Fingerprint	1822	991	91.22	90.61	97.28
BIOMDATA	Left iris	874	584	66.61	71.92	99.35
	Right iris	871	581	64.89	71.08	98.95
	Left thumb	875	644	61.23	63.96	80.15
	Left index	872	632	82.91	84.70	93.43
	Right thumb	871	647	62.11	63.52	82.63
	Right Index	870	624	82.05	84.46	93.12
Multi-Pie	Left view	10320	30940	45.52	47.30	87.50
	Frontal view	15480	38700	40.87	41.15	90.29
	Right view	10320	30960	45.13	47.30	85.49

Table 2: The number of samples in training and test sets and rank-one recognition rate for single modalities.

4. EXPERIMENTS AND DISCUSSIONS

CMU Multi-PIE database: This database [22] consists of face images under different illuminations, viewpoints, and expressions which are recorded in four sessions. Following the setup in [23], we consider the multi-view face images for 129 subjects that are present in all sessions. The available views are divided into three modalities of $\{-90°, -75°, -60°, -45°\}$, $\{0°, \pm 15°, \pm 30°\}$ and $\{45°, 60°, 75°, 90°\}$. Images from session 1 at views $\{0°, \pm 30°, \pm 60°, \pm 90°\}$ are used as training samples. Test images are obtained from all available view angles from session 2.

BioCop multimodal database: This database [24] is one of the few databases that allows disjoint training and testing of multimodal fusion at feature level. The BioCop database is collected under four disjoint years; 2008, 2009, 2012, and 2013. To make the training-test splits mutually exclusive, the 294 subject that are common in years 2012 and 2013 are considered. The proposed algorithm is trained on 294 mutual subjects in year 2013 dataset, and is tested on the same subjects in year 2012 dataset. It is worth mentioning that although the databases are labeled as 2012 and 2013, the date of data acquisition for common subjects in the datasets can vary between one to three years, which has also the advantage of investigating the effect of age-progression. We also consider the left and right irises as a single class, which results in heterogeneous classes for the iris modality.

BIOMDATA multimodal database: This database [25] is a challenging database, since many of the samples are damaged with blur, occlusion, sensor noise and shadows [12]. Following the setup in [12], six biometric modalities are considered: left and right irises, and thumb and index fingerprints from both hands. The experiments are conducted on 219 subjects that have samples in all six modalities. For each modality, four randomly chosen samples are used for the training and the remaining samples are used for the test set. For any modality in which the number of the samples is less than five, one sample is used for the test set and the remaining samples are used for the training. A summary of the databases is presented in Table 2.

Training and test phases: For each databases, the number of samples per individual and per modality varies. Therefore, for the training phase, for each individual 250 sets of modalities are randomly chosen from the training set. Similarly 250 sets are chosen from test set for the test phase. For Multi-Pie and BioCop databases, each triplet includes one sample from each modality. Similarly, for BIOMDATA database each set includes normalized left and right irises, and enhanced left index, right index, left thumb and right thumb fingerprint images. For Multi-Pie database the number of triplets in training and test phases is the same and equal to $32,250$. The number of triplets in BioCop database and sets of six images in BIOMDATA database for training and test phase are equal to $73,500$ and $54,750$, respectively.

Data representation: The face images are cropped, aligned to a template [26, 27], and resized to 224×224 images. Iris images are segmented, normalized using OSIRIS [28], and transformed into 64×512 strips. Each fingerprint image is enhanced using the method described in [29], The core point is detected from the enhanced image [30], and finally a 224×224 region centered by the core point is cropped.

Implementation: Initially, each modality-dedicated CNNs is trained independently, and each CNN is optimized on a single modality. For each modality, the conventional VGG19 network is pre-trained on Imagenet [31]. Pre-training helps with additional training data when the number of domain specific training data is limited. For the CNN-Face networks, the network is fine-tuned on CASIA-Webface [32] and the corresponding database (BioCop 2013 or CMU Multi-Pie databases). The preprocessing algorithm includes the channel-wise mean subtraction on RGB values, where the channel means are calculated on the whole training set. CNN-Iris networks are fine-tuned on CASIA-Iris-Thousand [33], Notre Dame-IRIS 04-05 [34], and finally the corresponding

Modality	{1,2}	{1,3}	{2,3}	{1,2,3}
SVM-Major	53.18	54.47	57.61	62.95
SVM-Sum	51.15	53.84	55.43	69.30
SMDL	71.65	74.14	70.27	81.30
JSRC	68.16	66.42	64.53	73.30
CNN-Major	92.18	93.75	89.74	95.87
CNN-Sum	91.58	93.28	89.13	94.51
Weighted feature fusion	94.12	94.96	91.53	96.59
Generalized compact bilinear	94.67	95.53	92.18	97.27

Table 3: Accuracy evaluation for different fusion settings for Multi-PIE database. 1, 2 and 3 represent frontal, right, and left views, respectively.

database (BioCop-Iris 2013 or BIOMDATA database). For the BioCop database, the CNN-Fingerprint network is fine-tuned on the BioCop 2013 right index fingerprint database. For the BIOMDATA database, the networks are fine-tuned on the corresponding fingerprint databases.

A two-step optimization algorithm is utilized to train the joint optimization of networks, where initially the modality-dedicated networks' weights are frozen and the joint representation layer is optimized greedily upon the extracted features by modality-dedicated networks. Then, all modality-dedicated networks, fusion layer, and the classification layer are jointly optimized.

Comparison of methods: To compare the results for the proposed algorithms, with the state-of-the-art algorithms, Gabor features in five scales and eight orientations are extracted from all modalities. For each face, iris, and fingerprint image, $31,360$, $36,630$, and $31,360$ features are extracted respectively. These features are used for all the algorithms except CNN-Sum, CNN-Major, and two proposed algorithms. Table 2 presents the results for the rank-one recognition rate for the databases. The performance of the proposed fusion algorithms is compared with several state-of-the-art feature, score and decision level fusion algorithms. SVM-Sum and CNN-Sum use the probability outputs for the test sample of each modality, added together to give the final score vector. SVM-Major and CNN-Major chose the maximum number of modalities taken to be from the correct class. The feature level fusion techniques include serial feature fusion [35], parallel feature fusion [36], CCA-based feature fusion [37], JSRC [1], SMDL [23], and DCA/MDCA [12] methods. Tables 3 and 4 present the results for different fusion settings. For all the databases we have considered $d = 4096$. For BIOMDATA database, due to the vast number of possible outer products, the generalized compact bilinear method only includes single modalities and three compact bilinear multiplications (two irises, two index fingers and two thumbs). The reported values are the average values for five randomly generated training and test sets for the training and test phases.

Modality	{1,2}	{1,3}	{2,3}	{1,2,3}
SVM-Major	79.22	89.27	80.47	90.32
Serial + PCA + KNN	71.12	86.28	75.69	76.18
Serial + LDA + KNN	80.12	91.28	79.69	82.18
Parallel + PCA + KNN	74.69	88.12	77.58	-
Parallel + LDA + KNN	82.53	93.21	82.56	-
CCA + PCA + KNN	87.21	95.27	86.44	95.33
CCA + LDA + KNN	89.12	95.41	86.11	95.58
DCA/MDCA + KNN	83.02	96.36	83.44	86.49
CNN-Sum	99.10	98.85	98.92	99.14
CNN-Major	98.51	97.70	98.31	99.03
Weighted feature fusion	99.18	99.03	99.12	99.25
Generalized compact bilinear	99.27	99.12	99.16	99.30

(a) BioCop database: 1, 2, and 3 represent face, iris, and fingerprint, respectively.

Modality	2 irises	4 fingerprints	6 modalities
SVM-Major	78.12	88.34	93.31
SVM-Sum	81.23	94.13	96.85
Serial + PCA+ KNN	72.31	90.71	89.11
Serial + LDA+ KNN	79.82	92.62	92.81
Parallel + PCA+ KNN	76.45	-	-
Parallel + LDA+ KNN	83.17	-	-
CCA + PCA + KNN	88.47	94.72	94.81
CCA + LDA + KNN	90.96	94.13	95.12
JSRC	78.20	97.60	98.60
SMDL	83.77	97.56	99.10
DCA/MDCA + KNN	83.77	98.1	99.60
CNN-Sum	99.54	99.46	99.82
CNN-Major	99.31	99.42	99.48
Weighted feature fusion	99.73	99.65	99.86
Generalized compact bilinear	99.79	99.70	99.90

(b) BIOMDATA database.

Table 4: Accuracy evaluation for different fusion settings.

5. CONCLUSION

In this paper, we proposed a joint CNN architecture with feature level fusion for multimodal recognition using multiple modalities. We proposed to apply fusion at fully-connected layers instead of convolutional layers to handle the possible spatial mismatch problem. This fusion algorithm results in no loss in performance, while the number of parameters is reduced significantly. We demonstrated that the multimodal fusion at the feature level and joint optimization of multi-stream CNNs significantly improve unimodal representation accuracy by incorporating the captured multiplicative interactions of the low-dimensional modality-dedicated feature representations, by means of generalized compact bilinear pooling.

ACKNOWLEDGEMENT

This work is based upon a work supported by the Center for Identification Technology Research and the National Science Foundation under Grant #1650474.

6. REFERENCES

[1] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, 2014.

[2] H. Jaafar and D. A. Ramli, "A review of multibiometric system with fusion strategies and weighting factor," *International Journal of Computer Science Engineering (IJCSE)*, vol. 2, no. 4, pp. 158–165, 2013.

[3] C.-A. Toli and B. Preneel, "A survey on multimodal biometrics and the protection of their templates," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2014, pp. 169–184.

[4] A. Nagar, K. Nandakumar, and A. K. Jain, "Multibiometric cryptosystems based on feature-level fusion," *IEEE transactions on information forensics and security*, vol. 7, no. 1, pp. 255–268, 2012.

[5] R. Connaughton, K. W. Bowyer, and P. J. Flynn, "Fusion of face and iris biometrics," in *Handbook of Iris Recognition*. Springer, 2013, pp. 219–237.

[6] Y. Shi and R. Hu, "Rule-based feasibility decision method for big data structure fusion: Control method." *International Journal of Simulation–Systems, Science & Technology*, vol. 17, no. 31, 2016.

[7] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, "Group sparse representation based classification for multi-feature multimodal biometrics," *Information Fusion*, vol. 32, pp. 3–12, 2016.

[8] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *24th International Conference on Pattern Recognition (ICPR)*, 2018.

[9] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[10] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear cnns," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[11] A. A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Defense and Security*. International Society for Optics and Photonics, 2005, pp. 196–204.

[12] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984–1996, 2016.

[13] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.

[14] J.-B. Delbrouck and S. Dupont, "Multimodal compact bilinear pooling for multimodal neural machine translation," *arXiv preprint arXiv:1703.08084*, 2017.

[15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[16] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[17] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi, "Attribute-centered loss for soft-biometrics guided face sketch-photo recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2018.

[18] A. Gangwar and A. Joshi, "Deepirisnet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2301–2305.

[19] R. F. Nogueira, R. de Alencar Lotufo, and R. C. Machado, "Fingerprint liveness detection using convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1206–1213, 2016.

[20] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 239–247.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[23] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.

[24] "Biocop database, http://biic.wvu.edu/." [Online]. Available: http://biic.wvu.edu/

[25] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak, "A protocol for multibiometric data acquisition, storage and dissemination," *Technical Report, WVU, Lane Department of Computer Science and Electrical Engineering*, 2007.

[26] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[27] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[28] E. Krichen, A. Mellakh, S. Salicetti, and B. Dorizzi, "Osiris (open source for iris) reference system," *BioSecure Project*, 2008.

[29] S. Chikkerur, C. Wu, and V. Govindaraju, "A systematic approach for feature extraction in fingerprint images," *Biometric Authentication*, pp. 1–23, 2004.

[30] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *IEEE transactions on Image Processing*, vol. 9, no. 5, pp. 846–859, 2000.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[32] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[33] "CASIA-iris-thousand, http://biometrics.idealtest.org/." [Online]. Available: http://biometrics.idealtest.org/

[34] K. W. Bowyer and P. J. Flynn, "The ND-IRIS-0405 iris image dataset," *arXiv preprint arXiv:1606.04853*, 2016.

[35] C. Liu and H. Wechsler, "A shape-and texture-based enhanced fisher classifier for face recognition," *IEEE transactions on image processing*, vol. 10, no. 4, pp. 598–608, 2001.

[36] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern recognition*, vol. 36, no. 6, pp. 1369–1381.

[37] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.