2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Quantifying Biometric Characteristics of Hand Gestures through Feature Space Probing and Identity-Level Cross-Gesture Disentanglement

Aman Verma[1], Gaurav Jaswal[2], Seshan Srirangarajan[1,3], and Sumantra Dutta Roy[1]
[1]Department of Electrical Engineering, Indian Institute of Technology Delhi, India
[2]iHub and HCI Foundation, Indian Institute of Technology Mandi, India
[3]Bharti School of Telecommunication Technology & Management, Indian Institute of Technology Delhi, India

Abstract— **We present the delta-gesture biometrics quantification assessment (DGBQA) framework which estimates the biometric characteristics of hand gestures. The proposed framework is aimed at learning generic motion-representations of gestures instead of subject-specific details from a large number of identities. It also enables the biometric scores to be estimated for a set of gestures at a time instead of having to estimate these one at a time. In the first step, it formulates a feature space which is identity and gesture aware, and in the second step, it proceeds to compute biometric scores using inter-subject and intra-subject distance measures in the feature space. However, due to the inclusion of identity-aware objective, the identity details tend to be shared across gestures. We refer to this as identity sharing and this can lead to the score for different gestures being dependent on each other. To address this issue, we introduce an identity-level cross-gesture disentanglement loss ($\mathscr{L}_{ICGD}$) which encourages the different gestures belonging to the same identity to be orthogonal in the feature space. We demonstrate the efficacy of the proposed biometric quantification framework and the disentanglement loss function through extensive experiments on four datasets and using standard as well as proposed novel evaluation metrics. Our analysis indicates that gestures involving multiple coarse movements are better for biometrics.**

I. Introduction

Video hand gestures are an emerging biometric modality [13], [20] which has application in personalized human-computer interaction [8], [6]. Improving authentication performance has been a key objective in this domain. To achieve this, research thrust has been on: (i) developing domain-specific architectures for extracting richer biometric-features [14], [15], [16], and (ii) authentication using different gesture acquisition modalities such as depth [20] and egocentric RGB videos [18]. Although performance gains have been achieved, there is no standard protocol that is followed while building the gesture sets due to which the datasets mostly consist of disparate gestures [7], [20]. This leads us to the question: which hand gestures are best suited for biometric authentication? It has been found that some gestures tend to accentuate identity details and are characterized by lower error rates during authentication [7]. Gestures involving intricate but coarse motion patterns have been found to perform well in biometric applications [13], [20]. However, there are no comprehensive qualitative guidelines or quantitative measures for identifying gestures that are suitable for biometrics. In this work, we aim to address the need for a quantitative 'biometric goodness' measure for hand gestures. Such a quantitative measure will simplify the design decisions, such as which gestures are to be used for personalized embedded devices. In addition, these measures would help in understanding which features and motion patterns are important for biometrics.

In order to quantify biometric goodness, one approach would be to perform biometric verification experiments for each gesture across different models, and use the average error rate as a quantitative measure. However, this approach has several limitations. This process is time-consuming and has to be followed for each gesture at a time. For this evaluation to be generic, experiments should be conducted over a large number of identities. In spite of this, the measure may still be estimated from a feature space conditioned on the identities rather than only on the gesture motion descriptions. Thus, we need a measure or method that is based on gesture and identity-aware representations.

There have been attempts in the literature to quantify biometrics in other domains. In [10], a biometric score for online signature templates is presented on the basis of distinctiveness, repeatability, and complexity of the signature templates. This method relies on explicit matching and a large number of identities. In [17], biometric capacity of face representations were estimated based on inter-subject matching thresholds and dimensionality of feature embeddings. In [2], the biometric capacity estimation of faces was posed as a sphere-packing problem. This is based on the assumption that all the samples cover the representation space uniformly. However, regions with higher and lower clustering are generally evident. In [21], a personalization score for gestures is presented on the basis of intra-subject distances. However this formulation does not consider intra-gesture distances across subjects. From [4], [10], we know that biometric goodness of an individual can be characterized in terms of uniqueness and variability. We adapt these to the hand gesture setting by defining *uniqueness* to refer to the separation between a gesture performed by different subjects, and *variability* as the variation across the gesture instances performed by a subject. We will quantify these properties and fuse them to arrive at a quantitative biometric score for each gesture.

We propose the delta-gesture biometric quantification as-

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

sessment (DGBQA) framework for quantification of biometric characteristics. This framework follows a two-step procedure where in the first step, a feature space for biometric scoring is formulated. From this feature-space, uniqueness and variability parameters are estimated using distance between the feature embeddings. Finally, using these parameters, we propose the DGBQA score which estimates the biometric characteristics. The term 'delta' is used to refer to the fact that the proposed scoring formulation involves pairwise differences between the embeddings. Please note that we will use the terms biometric goodness, biometric characteristics, and biometric score interchangeably. Similarly, the terms subject and identity will be used interchangeably.

From the above discussion, we identify three desirable characteristics for the feature space as: (i) gesture-aware: the feature space should capture generic representation of the gestures, (ii) identity-aware: the feature space must preserve identity details so as to facilitate robust biometric scoring, and (iii) the feature space should allow biometric scoring of multiple gestures at a time. In order to formulate the feature space, the DGBQA framework performs joint learning of the gesture classes and subject identities. Specifically, we formulate a multi-task optimization with the objective of hand gesture and identity recognition. However, it is possible that the identity details are shared across gestures. We refer to this as *identity sharing* and it can be attributed to the identity recognition objective which allows the model to learn identities irrespective of the gesture. A consequence of identity sharing would be that the biometric score for different gestures will be dependent on each other. To address this, we introduce an identity-level cross-gesture disentanglement loss ($\mathscr{L}_{ICGD}$) as part of the optimization objective. The $\mathscr{L}_{ICGD}$ term encourages the different gesture representations belonging to a given identity to be disentangled or separated in the feature space. Once the DGBQA scores are computed, they must evaluated. We propose several performance metrics to evaluate effectiveness of the DGBQA scores. These performance metrics can be applied to other biometric scoring frameworks as well.

In summary, the key contributions of this work are:

- We propose the DGBQA framework for estimation of biometric characteristics of hand gestures. This is achieved by formulating a gesture and identity-aware feature space while allowing estimation of biometric characteristics of multiple gestures at a time. To the best of our knowledge, this is the first work which quantifies biometric characteristics for hand gestures.
- We uncover identity sharing which induces biometric characteristics of one gesture into another gesture.
- To address identity sharing we propose the identity-level cross-gesture disentanglement loss ($\mathscr{L}_{ICGD}$) which encourages representations of different gestures belonging to a given identity to be separated in the gesture space.
- We propose the DGBQA score for estimating the biometric characteristics of hand gestures by capturing both uniqueness and variability parameters.
- We propose several performance metrics for evaluating the proposed DGBQA framework. We demonstrate efficacy of the proposed framework through extensive experiments on four benchmark datasets.

II. Proposed DGBQA Framework

The proposed DGBQA framework for quantification of biometric characteristics of hand gestures is a two-step procedure. In the first step, an appropriate feature space is constructed, while in the second step, biometric scores are computed. The proposed overall framework is illustrated in Fig. 1. We next describe the proposed framework and the evaluation metrics.

A. Gesture and Identity-Aware Representation

Gesture-Aware Representation: The feature space must capture natural representations such as motion patterns of gestures. This will result in similar gestures being placed closer in the feature space resulting in gesture clusters. Since hand gestures also contain identity details [13], [7], we expect the formation of identity clusters within the gesture clusters. Hand gesture recognition (HGR) task requires the ability to learn motion representations and can serve as a proxy for natural representation modeling of gestures. Hence, at a preliminary level, we rely on hand gesture recognition for the feature space formulation. We use the cross-entropy loss and refer it as $\mathscr{L}_{HGR}$. However, $\mathscr{L}_{HGR}$ will attempt to cluster all the instances of a given gesture irrespective of their identity. This can result in loss of the identity-aware characteristics in the feature space.

Joint Gesture and Identity-Aware Representation: To avoid loss of the identity details, we introduce the identity recognition loss ($\mathscr{L}_{ID}$) along with $\mathscr{L}_{HGR}$. Thus, for feature space construction, we consider a multi-task objective comprising identity recognition (ID) and HGR. This will allow the feature space to capture gesture understanding as well as identity details. Since, the ID task is relatively more fine-grained, identity details are more challenging to extract. This would allow formation of gesture clusters along with identity clusters within each gesture cluster. Similar to $\mathscr{L}_{HGR}$, $\mathscr{L}_{ID}$ also uses the cross-entropy loss.

Let $\mathscr{L}_{Obj}$ denote the optimization objective function and $X_i \in \mathbb{R}^{T\times H\times W\times C}$ be the input to a network $f_\theta(.)$. Here, T represents the number of frames, H and W represent the spatial dimensions, and C represents the channel dimensions. Let $f_i \in \mathbb{R}^d$ represent $f_\theta(X_i)$, where d is the dimensionality of the output embeddings.

$$\mathscr{L}_{Obj} = \mathscr{L}_{HGR} + \lambda_{ID}\mathscr{L}_{ID} \tag{1}$$

$$\mathscr{L}_{Obj} = \frac{1}{N}\sum_{i=1}^{N}\frac{\exp(W_{HGR_{y_i}}^T f_i)}{\sum_{j=1}^{G}\exp(W_{HGR_j}^T f_i)} + \lambda_{ID}\frac{1}{N}\sum_{i=1}^{N}\frac{\exp(W_{ID_{y_i}}^T f_i)}{\sum_{j=1}^{I}\exp(W_{ID_j}^T f_i)} \tag{2}$$

Here, λ_{ID} is the weighting factor for $\mathscr{L}_{ID}$, and N, G, and I represent the number of samples, gestures, and identities, respectively. Furthermore, $W_{HGR} \in \mathbb{R}^{G\times d}$ and $W_{ID} \in \mathbb{R}^{I\times d}$ are weight matrices of the task-specific fully-connected layers. It must be noted that $\mathscr{L}_{ID}$ will try to cluster identities

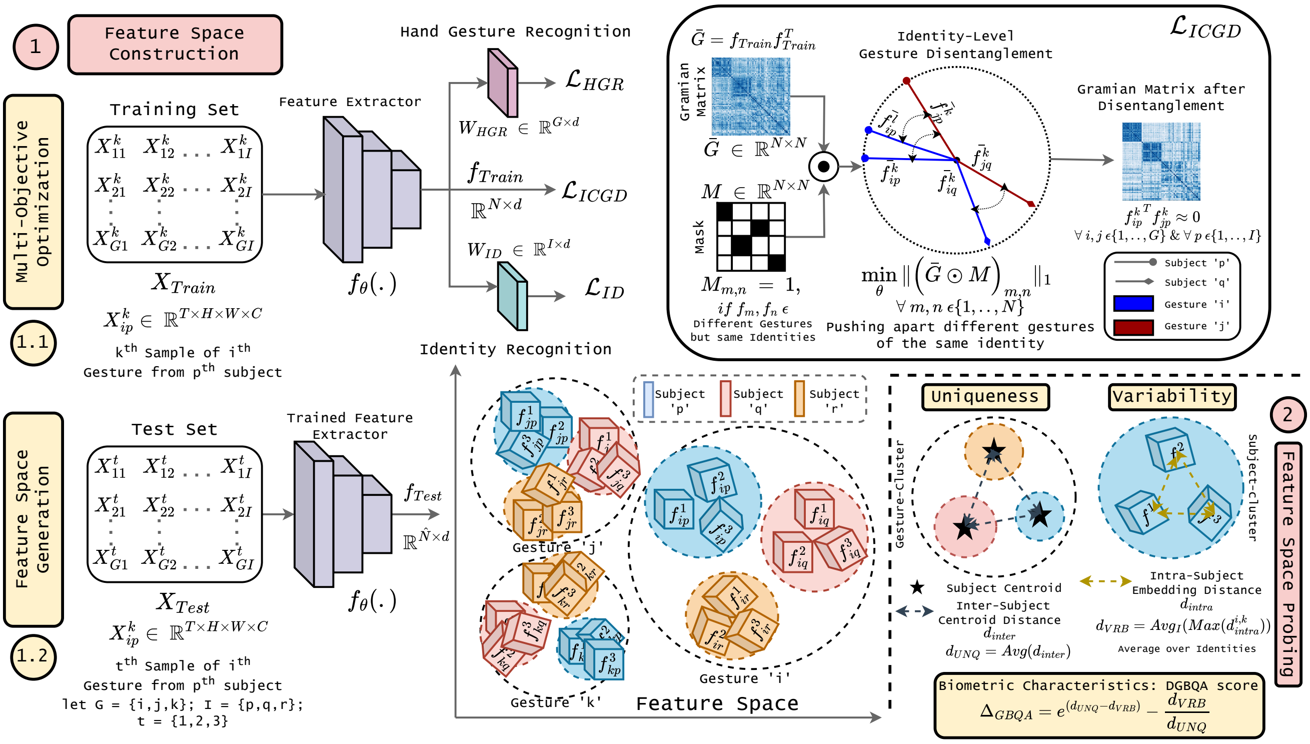

1
Feature Space Construction
Multi-Objective Optimization
1.1
Training Set
X_{11}^k X_{12}^k . . . X_{1I}^k
X_{21}^k X_{22}^k . . . X_{2I}^k
X_{G1}^k X_{G2}^k . . . X_{GI}^k
X_{Train}
$X_{ip}^k \in \mathbb{R}^{T \times H \times W \times C}$
kth Sample of ith Gesture from pth subject
Feature Extractor
$f_\theta(.)$
Hand Gesture Recognition
$\mathcal{L}_{HGR}$
$W_{HGR} \in \mathbb{R}^{G \times d}$
f_{Train}
$\mathbb{R}^{N \times d}$
$\mathcal{L}_{ICGD}$
$W_{ID} \in \mathbb{R}^{I \times d}$
$\mathcal{L}_{ID}$
Identity Recognition
$\bar{G} = f_{Train} f_{Train}^T$
Gramian Matrix
$\bar{G} \in \mathbb{R}^{N \times N}$
$M \in \mathbb{R}^{N \times N}$
Mask
$M_{m,n} = 1,$
$if\ f_m, f_n\ \epsilon$
Different Gestures but same Identities
Identity-Level Gesture Disentanglement
$\bar{f}_{ip}^l$
$\bar{f}_{jp}^k$
$\bar{f}_{jq}^k$
$\bar{f}_{ip}^k$
$\bar{f}_{iq}^k$
$\min_\theta \| \left(\bar{G} \odot M \right)_{m,n} \|_1$
$\forall\ m, n\ \epsilon \{1, .., N\}$
Pushing apart different gestures of the same identity
Gramian Matrix after Disentanglement
${f_{ip}^k}^T f_{jp}^k \approx 0$
$\forall\ i, j\ \epsilon \{1, .., G\}\ \&\ \forall\ p\ \epsilon \{1, .., I\}$
Subject 'p'
Subject 'q'
Gesture 'i'
Gesture 'j'
Feature Space Generation
1.2
Test Set
X_{11}^t X_{12}^t . . . X_{1I}^t
X_{21}^t X_{22}^t . . . X_{2I}^t
X_{G1}^t X_{G2}^t . . . X_{GI}^t
X_{Test}
$X_{ip}^k \in \mathbb{R}^{T \times H \times W \times C}$
tth Sample of ith Gesture from pth subject
let G = {i,j,k}; I = {p,q,r}; t = {1,2,3}
Trained Feature Extractor
$f_\theta(.)$
f_{Test}
$\mathbb{R}^{\hat{N} \times d}$
Subject 'p'
Subject 'q'
Subject 'r'
Gesture 'j'
Gesture 'i'
Gesture 'k'
Feature Space
Uniqueness
Gesture-Cluster
Subject Centroid
Inter-Subject Centroid Distance
d_{inter}
$d_{UNQ} = Avg(d_{inter})$
Variability
Subject-Cluster
Intra-Subject Embedding Distance
d_{intra}
$d_{VRB} = Avg_I(Max(d_{intra}^{i,k}))$
Average over Identities
2
Feature Space Probing
Biometric Characteristics: DGBQA score
$\Delta_{GBQA} = e^{(d_{UNQ} - d_{VRB})} - \frac{d_{VRB}}{d_{UNQ}}$

Fig. 1. Proposed DGBQA framework: The first step is feature space construction which is illustrated in part (1) of the figure. This step consists of two stages: (1.1) Multi-objective optimization and (1.2) Feature-space generation. In (1.1), the feature extractor f_θ is trained for hand gesture and identity recognition. To address identity sharing, we introduce the $\mathscr{L}_{ICGD}$ term in the objective function. After training f_θ, in (1.2), embeddings are extracted which constitute the feature space. In the second step, referred to as feature space probing and illustrated as part (2) of the figure, the DGBQA score (Δ_{GBQA}) is computed using uniqueness and variability parameters to capture the biometric characteristics.

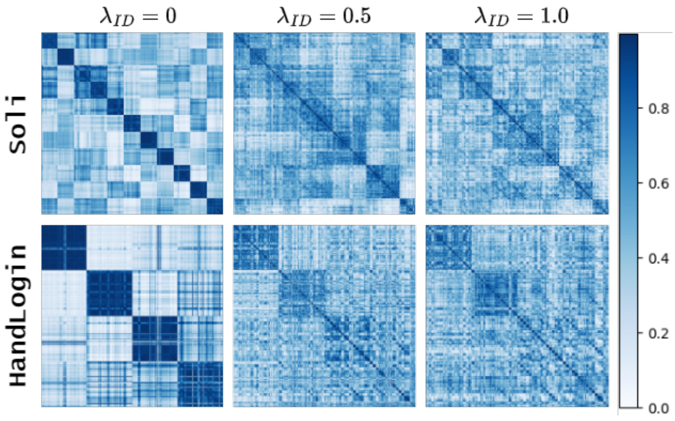


Fig. 2. Identity sharing: Correlation maps of feature embeddings for different values of λ_{ID}. The data is arranged gesture-wise, and within each gesture, embeddings belonging to each subject are placed together. The diagonal elements within the off-diagonal blocks represent correlation between different gestures of the same subject. Higher the correlation values greater is the sharing of identity details across gestures and indicates identity sharing. As λ_{ID} increases, identity sharing becomes more prominent.

irrespective of the gestures. Thus, using λ_{ID} we can control the trade-off between gesture and identity understanding.

B. Identity Sharing

Since we aim to learn the feature space representation for multiple gestures at a time during optimization, for any subject, f_θ will also learn the subject's identity details across these gestures. Thus, identity details in the output embedding will not belong to an individual gesture. Rather, it is composed of identity details captured across gestures of that subject. We know that an identity cluster within a gesture cluster is composed of embeddings of a given subject. However, these embeddings are influenced by identity details from other gestures. As a result, subject clusters as a whole exhibit some feature-level intermingling with the corresponding subject cluster within other gesture clusters. We refer to this as *identity sharing* and biometric scores for gestures derived based on such a feature space representation would be dependent on other gestures in the set. To illustrate the presence of identity sharing, in Fig. 2, we show the correlation maps of the embeddings for different values of λ_{ID}. It is seen that as we increase λ_{ID}, the correlation values between embeddings for a given subject across different gestures increases. This indicates sharing of identity details across gestures. $\mathscr{L}_{HGR}$ enables gesture understanding, however these empirical results indicate that it is unable to restrain the learning of identity details to a given gesture.

C. Identity-Level Cross-Gesture Disentanglement

Identity sharing results in embeddings belonging to a given subject but from different gestures to have some feature-level intermingling. If these embeddings are decorrelated, then we can constrain these cross-gesture interactions. To this end, we propose identity-level cross-gesture disentanglement ($\mathscr{L}_{ICGD}$) loss. The principle behind this loss term is illustrated in Fig. 1 (refer the box in top-right corner). The primary objective is to penalize those embeddings which result in significant correlation between different gesture embeddings of a given subject.

Let $X_{Train} \in \mathbb{R}^{N\times T\times H\times W\times C}$ represent a training batch with N being the batch size. Also, $\bar{f} = f_\theta(X_{Train}) \in \mathbb{R}^{N\times d}$, where $\bar{f}$ is the embedding matrix corresponding to the training batch. Each of the embeddings are l_2-normalized. Let $f_m \in \mathbb{R}^d$ represent the normalized embedding corresponding to the m^{th} sample such that $\|f_m\|_2 = 1$. Using the normalized embeddings $\bar{f}$, the Gram matrix $\bar{G}$ is obtained as $\bar{G} = \bar{f}\bar{f}^T$. Let $\bar{G}_{mn} = {f_m}^T f_n$ represent the $(m,n)^{\text{th}}$ element of the Gram matrix $\bar{G}$ and the correlation between the embeddings f_m and f_n. Thus, $\bar{G}$ contains the correlation values between any two embeddings of the training batch. Since the embeddings are normalized $\bar{G}_{mn} = \cos(\phi_{mn})$ where ϕ_{mn} is the angle between the embeddings f_m and f_n. If the correlation value is higher then it indicates that $\phi_{mn} \to 0$, and if the correlation is lower, then the embeddings would tend to be orthogonal.

After $\bar{G}$ is computed, suitable mask can be applied to it to extract the elements of $\bar{G}$ which are composed of the desired embeddings. For each subject i ($\forall i = 1,\dots,I$) considered in the training batch, we define a mask $M_i \in \mathbb{R}^{N\times N}$ as,

$$M_i = \Gamma \odot L \odot \delta^G \odot \delta_i^{ID}; \tag{3}$$

where $\odot$ represents the Hadamard product, and Γ, L, δ^G, δ_i^{ID} represents the negative mask, lower-triangular mask, gesture mask, and identity mask, respectively. Since $\bar{G}$ is a symmetric matrix, to remove the redundant values we use L to mask the lower-triangular elements to zero. We use Γ to mask the negative correlation values.

$$\Gamma_{mn} = \begin{cases} 1, & \text{if } \bar{G}_{mn} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

δ^G and δ_i^{ID} are defined as:

$$\delta^G_{mn} = \begin{cases} 1, & \text{if } y_{HGR_m} \neq y_{HGR_n} \text{ and } m \neq n \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\delta^{ID}_{i,mn} = \begin{cases} 1, & \text{if } y_{ID_m} = y_{ID_n} = i \text{ and } m \neq n \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

y_{HGR_m} and y_{ID_m} represent the gesture and subject labels of the m^{th} sample, respectively. δ^G is employed for masking elements corresponding to other gestures, and δ_i^{ID} is employed to mask elements corresponding to subjects other than subject i.

We define the $\mathscr{L}_{ICGD}$ loss as:

$$\mathscr{L}_{ICGD} = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\sum_{m=1}^{N}\sum_{n=1}^{N}(\bar{G}\odot M_i)_{mn}}{\sum_{m=1}^{N}\sum_{n=1}^{N}M_{i,mn}+1}\right) \tag{7}$$

From (7), we note that for a subject i with mask M_i, $\mathscr{L}_{ICGD}$ uses those elements of $\bar{G}$ which have: (i) positive correlation values, (ii) reside in the upper-triangular matrix, and (iii) originate from embeddings of subject i for different gestures. Then, $\mathscr{L}_{ICGD}$ minimizes the sum over the elements of the masked Gram matrix $\bar{G}\odot M_i$. The denominator is a normalization factor representing the number of elements selected by the mask M_i. A one is added to the denominator so as to prevent it from becoming zero. This is computed for each subject and loss value is the averaged value over all the subjects. The overall $\mathscr{L}_{ICGD}$ loss formulation encourages the model f_θ to construct embeddings which can decorrelate the different gesture embeddings of any given subject. This is equivalent to trying to make the angle between the selected embedding pairs closer to 90°, and thus making them orthogonal.

Final Loss Function: The final loss function is formulated as:

$$\mathscr{L}_{Obj} = \mathscr{L}_{HGR} + \lambda_{ID}\mathscr{L}_{ID} + \lambda_{ICGD}\mathscr{L}_{ICGD} \tag{8}$$

where λ_{ICGD} is the weighting factor for the $\mathscr{L}_{ICGD}$ term.

D. Feature Space Probing and Delta-GBQA Score

After f_θ has been trained using $\mathscr{L}_{Obj}$, we extract embeddings of $X_{Test} \in \mathbb{R}^{\hat{N}\times T\times H\times W\times C}$ where $\hat{N}$ is the number of samples in the test set, i.e., $f_{Test} = f_\theta(X_{Test}) \in \mathbb{R}^{\hat{N}\times d}$. As illustrated in Fig. 1, we expect the feature space to be composed of gesture clusters which in turn would comprise of subject clusters. For each gesture cluster, we measure distances between the embeddings. We refer to this step as feature probing and which is expected to estimate the biometric characteristics of the gesture set. The estimation step is based on uniqueness and variability parameters. For a given gesture, the uniqueness parameter is expected to have higher value when different subjects have larger distances between their embeddings. The variability parameter is expected to have a higher value when a subject has tight clustering across its embeddings. We define $f^m_{g,i}$ as the m^{th} embedding for gesture g of subject i.

Uniqueness Parameter (d_{UNQ}): To quantify uniqueness for a gesture, we first compute the subject centroids ($\hat{f}_{g,i}$ represents the centroid of gesture g and subject i) for all the subjects in the test set. These centroids are averaged over all the embeddings of a given subject. Let P be the number of embeddings of a subject i for a gesture g, then:

$$\hat{f}_{g,i} = \frac{1}{P}\sum_{m=1}^{P} f^m_{g,i} \tag{9}$$

Distance between two subject centroids signifies the average distance between the two subjects for a given gesture. Thus, we compute the average of these distances over all the subject pairs. This will quantify, on an average, how uniquely the subjects perform this gesture. We define this as the uniqueness parameter d_{UNQ_g}. Higher the uniqueness parameter value, better are the biometric characteristics of the gesture.

$$d_{UNQ_g} = \frac{1}{\left[\frac{I(I-1)}{2}\right]} \sum_{m=1}^{I-1} \sum_{n=m+1}^{I} \|\hat{f}_{g,m} - \hat{f}_{g,n}\|_2 \tag{10}$$

Variability Parameter (d_{VRB}): For a gesture, variability measures the amount of variance in the embeddings within a subject cluster. To quantify this, we compute the maximum variance for each subject and then average it over all the subjects. Lower is the variance value, better are the biometric characteristics of the gesture. Maximum variance is measured as maximum distance between the embeddings of a particular subject. Let Q be the number of embeddings per subject for a given gesture. The variability parameter is computed as:

$$d_{VRB_g} = \frac{1}{I} \sum_{i=1}^{I} \max_{m=\{1,\ldots,Q-1\}, n=\{(m+1),\ldots,Q\}} \|f^m_{g,i} - f^n_{g,i}\|_2 \tag{11}$$

DGBQA Score (Δ_{GBQA}): We propose the DGBQA score for estimation of biometric characteristics using the uniqueness and variability parameters. The DGBQA score for any gesture g is given as

$$\Delta_{GBQA_g} = \exp\left(d_{UNQ_g} - d_{VRB_g}\right) - \left(\frac{d_{VRB_g}}{d_{UNQ_g}}\right) \tag{12}$$

We use the term 'delta' in DGBQA to refer to the fact that the proposed scoring formulation involves pairwise differences between the embeddings. Higher the Δ_{GBQA} score value, better are the biometric characteristics of the gesture. The first term contributes significantly to the score if the uniqueness parameter is higher while the variability is minimal. We use an exponential form for this term so as to assign significant value to the difference between the uniqueness and variability parameters. The second term is a penalty term that penalizes the score if d_{VRB_g} is relatively compared to d_{UNQ_g}. The Δ_{GBQA} score can also take negative values. Since the scores are derived for a set of gestures at a time, they provide a relative measure of the biometric characteristics of the gestures in the set. To compare these scores beyond the gesture set, we perform z-score normalization followed by l_2-normalization. This ensures that the DGBQA score values are in the range $[-1,1]$.

E. Evaluation Metrics

Next, we describe the proposed evaluation measures and strategies for validating the proposed DGBQA biometric score formulation. As the biometric score estimation involves formulation of a feature space and then the score computation, we need to evaluate the effectiveness of both steps.

1) Preliminaries: In general, biometric goodness is measured in terms of equal error rate (EER) computed from verification experiments. Lower EER values indicate better biometric characteristics. On the other hand, in the proposed framework, higher DGBQA scores indicate better biometric characteristics. Thus, we use $(100-\mathrm{EER})$ along with the normalized DGBQA scores. To facilitate comparison, we consider $e \in \mathbb{R}^G$ as the vector of EER values (in %). Then, let $\bar{e} = (100 \cdot \bar{1}) - e$, where $\bar{1} \in \mathbb{R}^G$ represents the vector of all ones. We perform z-score normalization, followed by l_2-normalization over $\bar{e}$ to obtain $\hat{e} \in \mathbb{R}^G$ which is considered as the ground truth biometric score.

2) Evaluation Metrics: Apart from considering the test set recognition accuracy for the HGR and ID tasks, we propose the following evaluation metrics.

Rank Deviation ($\hat{r}$): This measure quantifies the average difference between the biometric goodness ranks of gestures in the set, where the ranks are arrived at using Δ_{DGBQA} and $\hat{e}$. Let r^{Δ}_g and $r^{\hat{e}}_g$ represent the ranks of g^{th} gesture based on the DGBQA score and the ground truth biometric scores, respectively. Lower rank values indicate better biometric characteristics. Thus, we define the rank deviation $\hat{r}$ as:

$$\hat{r} = \frac{1}{G} \left(\sum_{g=1}^{G} \|r^{\Delta}_g - r^{\hat{e}}_g\|_1 \right) \tag{13}$$

Rank deviation will be the key measure for comparing different biometric scoring frameworks.

ICGD Score (C_D): The ICGD score measures the residual identity sharing and is defined as:

$$C_D = \frac{1}{I} \sum_{i=1}^{I} \left(\frac{\sum_{m=1}^{N} \sum_{n=1}^{N} (\Gamma \odot L \odot \delta^G \odot \delta_i^{ID} \odot \bar{G})_{mn}}{\sum_{m=1}^{N} \sum_{n=1}^{N} (\Gamma \odot L \odot \delta^G \odot \delta_i^{ID})_{mn}} \right) \tag{14}$$

where N, G and I represent the number of test samples, gestures, and subjects, respectively. It can be seen that C_D measures the average correlation value between the embeddings that contribute to identity sharing. Lower the C_D value, less is the extent of identity sharing across gestures. We use this measure to compare the different feature spaces.

Acceptance and Normalized Acceptance Values (A_r, nA_r): For the proposed DGBQA framework to perform effectively, we require higher DGBQA scores for better ranks and the rank deviation to be minimal. To this end, we propose the acceptance value A_r. Let $\Delta[k]$ represent the DGBQA score of the k^{th} gesture. Then,

$$A_r(\Delta) = \sum_{j=1}^{G} \frac{2^{\lambda\left(\gamma\left(\frac{G-r^{\hat{e}}_j+1}{G}\right)\Delta\left[r^{\hat{e}}_j\right] + \frac{r^{\hat{e}}_j}{G}\left(1-\Delta\left[r^{\hat{e}}_j\right]\right)\right)}}{\exp(\|r^{\Delta}_j - r^{\hat{e}}_j\|_1)} \tag{15}$$

where λ and γ are scaling factors and we set them to 2. The proposed acceptance value consists of two key terms: (i) relevance term (numerator): this quantifies if the DGBQA score is higher for better ranks, and (ii) rank deviation term (denominator): this quantifies the amount of rank deviation for a gesture. The second term is relatively more important than the first as we would like the relative orders to match as far as possible. To account for this, we use an exponential term for the rank deviation. Higher the acceptance value A_r, better is the biometric scoring. This measure allows for comparing scores generated from different feature spaces. In order to compare different feature extractors, we propose a normalized acceptance value as:

$$nA_r(\Delta) = \frac{A_r(\Delta)}{A_r(\hat{e})} \tag{16}$$

III. Experimental Analysis

In this section, we evaluate efficacy of the proposed DGBQA biometric quantification framework through extensive experiments on four datasets and using the proposed evaluation metrics.

A. Experimental Protocol

1) ***Model Architecture****:* For feature space construction, we employ two architectures: Res3D-ViViT and Res3D-MF. Both architectures have a residually connected 3D-CNN backbone [5]. Following the backbone, these networks have several transformer encoder layers. Res3D-ViViT utilizes the ViViT-based encoder [1], while Res3D-MF utilizes the MotionFormer encoder [9]. We use these models as they have been used in the literature for temporal modeling. However, we would like to clarify that model architecture is not the focus of this work.

2) ***Datasets and Protocol****:* For all the experiments, we use a $60:40$ training-test split. We use following publicly available datasets for the performance evaluation.

Soli [19]: This dataset contains range-Doppler image sequences of hand gestures collected from 10 subjects and 11 gestures. It contains a total of 2750 gesture instances/samples. We obtain ground truth biometric scores by performing biometric verification experiments for each gesture. We employ 5-fold cross-validation and averaged the results obtained from three feature-extractors namely, TDSNet [13], ESNet [3], and Res3D-ViViT.

TinyRadar [11]: This is a large-scale range-Doppler HGR dataset with 11 gestures (same gesture set as Soli dataset) from 26 subjects. It contains a total of $30,300$ samples. The ground truth scores were computed using 3-fold cross-validation and averaging the results from TDSNet and Res3D-ViViT models. Note that this dataset is based on a different radar sensor type than the one used for Soli dataset.

HandLogin [20]: This dataset comprises of depth-maps for 4 gestures with 15 subjects. Since some hand-geometry details are explicitly present in the depth-maps, we obtain first-temporal difference map for the sequences [12]. We consider this pre-processing in order to generate the feature space using only the motion details of the gestures. This ensures that the biometric scores are derived only from motion details and do not consider any physiological details. This step is taken as other modalities such as range-Doppler sequences do not contain physiological details. For the ground truth scores, we use the values reported in [20].

SCUT-DHGA [7]: This is a large-scale RGB-based hand gesture authentication dataset. It comprises of 6 gestures with 143 subjects. Similar to that with the HandLogin dataset, we obtain temporal-difference maps of the sequences. For the ground truth scores, we average over the results of different models considered in the cross-session scenario in [7].

TABLE I

Ablation study on the feature space. With the introduction of $\mathscr{L}_{ICGD}$, $\hat{r}$ and C_D values (to be minimized) tend to decrease while nA_r values (to be maximized) increase compared to the case without $\mathscr{L}_{ICGD}$. This indicates mitigation of identity sharing and improvement in biometric characterization of hand gestures.

Model	λ_{ID}	λ_{ICGD}	HGR Acc.	ID Acc.	$\hat{r}$	C_D	A_r	nA_r
Soli								
Res3D-ViViT	0.0	0.0	95.64	–	4.09	0.375	3.12	0.08
	1.0	0.0	89.90	75.27	1.54	0.386	10.81	0.30
		0.5	95.27	74.54	1.54	0.309	11.32	0.31
		1.5	74.18	76.00	1.90	0.218	15.21	0.42
	1.5	0.0	89.53	76.18	0.81	0.387	23.99	0.67
		0.5	93.45	72.90	0.45	0.351	27.66	0.77
		1.5	92.63	75.36	1.36	0.239	18.32	0.51
Res3D-MF	0.0	0.0	95.00	–	2.27	0.287	11.85	0.32
	1.0	0.0	92.63	76.81	1.0	0.406	25.01	0.69
		0.5	90.54	72.63	1.0	0.361	15.39	0.42
		1.5	90.36	76.18	1.0	0.229	19.49	0.54
	1.5	0.0	90.18	76.54	0.27	0.432	34.06	0.94
		0.5	89.18	75.00	0.45	0.340	24.11	0.66
		1.5	90.00	72.36	1.54	0.284	14.62	0.40
HandLogin								
Res3D-ViViT	0.0	0.0	94.53	–	1.0	0.371	6.16	0.36
	0.5	0.0	87.89	44.53	0.0	0.365	19.46	1.14
		1.0	90.23	46.48	1.0	0.265	9.12	0.53
		1.5	88.67	46.09	1.0	0.248	13.12	0.77
	1.0	0.0	84.76	56.74	1.0	0.385	6.28	0.36
		1.0	83.20	50.00	1.0	0.350	12.19	0.71
		1.5	89.94	49.65	1.0	0.284	8.78	0.51
Res3D-MF	0.0	0.0	91.02	–	1.0	0.317	9.50	0.55
	1.0	0.0	85.55	41.80	1.5	0.382	2.41	0.14
		1.0	85.93	57.81	0.5	0.296	13.77	0.81
		2.5	85.54	46.09	1.0	0.226	7.10	0.41
	1.5	0.0	76.56	49.21	1.0	0.377	7.90	0.46
		1.0	83.98	58.59	1.0	0.324	9.79	0.57
		1.5	78.51	63.67	1.0	0.284	12.40	0.72
TinyRadar								
Res3D-ViViT	0.0	0.0	90.19	–	2.72	0.357	15.20	0.41
	1.0	0.0	87.62	64.39	0.90	0.540	21.15	0.57
		1.0	86.29	57.55	1.27	0.515	21.97	0.60
		1.5	86.42	58.23	1.27	0.496	21.64	0.59
		2.5	85.30	57.13	1.45	0.541	18.93	0.51
SCUT-DHGA								
Res3D-ViViT	0.0	0.0	99.59	–	2.33	0.196	5.14	0.23
	1.5	0.0	98.27	28.14	1.33	0.576	11.05	0.51
		1.0	93.85	25.93	0.66	0.566	15.20	0.70
		1.5	95.44	9.60	1.33	0.533	6.81	0.31
		2.5	96.47	18.82	1.66	0.523	6.38	0.29

B. Results, Analysis, and Discussion:

1) Ablation Study on Feature Space: First, we validate robustness of the proposed feature space formulation. We experiment with different values of λ_{ID} and λ_{ICGD} and the results are reported in Table I. When only $\mathscr{L}_{HGR}$ is employed, the rank deviation is relatively higher. By augmenting the objective with $\mathscr{L}_{ID}$, there is a sharp reduction in $\hat{r}$ (and increase in A_r), but we observe a significant increase in the ICGD score (C_D). This highlights the existence of identity sharing. Furthermore, as λ_{ID} is increased, there is a greater intermingling of identity details across gestures resulting in increased C_D values. After introducing $\mathscr{L}_{ICGD}$, we observe a significant reduction in C_D values along with an increase in the acceptance values. This clearly indicates that: (i) $\mathscr{L}_{ICGD}$ is key in reducing the identity sharing, and (ii) once the gesture embeddings focus on identity details only within their motion patterns, we obtain improved estimation of their

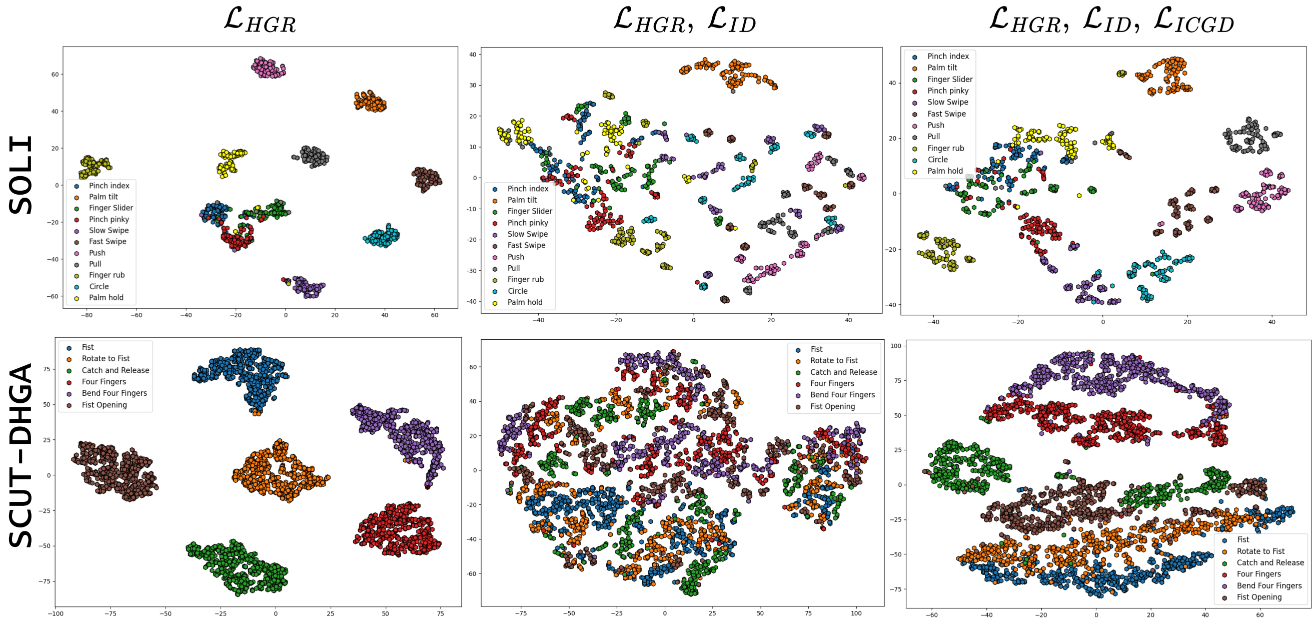


Fig. 3. Comparative study on feature space formulation: The t-SNE plots for three loss function objectives: (i) Only $\mathscr{L}_{HGR}$ (1st column), (ii) $\mathscr{L}_{HGR}$ and $\mathscr{L}_{ID}$ (2nd column), (iii) $\mathscr{L}_{HGR}$, $\mathscr{L}_{ID}$, and $\mathscr{L}_{ICGD}$ (3rd column). The proposed multi-objective optimization (3rd column) enables formulation of gesture-clusters consisting of smaller subject-clusters.With $\mathscr{L}_{HGR}$ and $\mathscr{L}_{ID}$ (2nd column) identity details are highly shared across gestures. However, the inclusion of $\mathscr{L}_{ICGD}$ is seen to mitigate identity sharing.

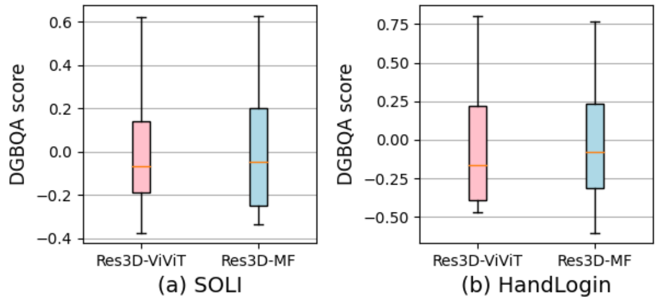


Fig. 4. DGBQA scores using two different models. This shows that the DGBQA scores using the two models are similar and statistically not very different. This indicates that the computed DGBQA scores are generic.

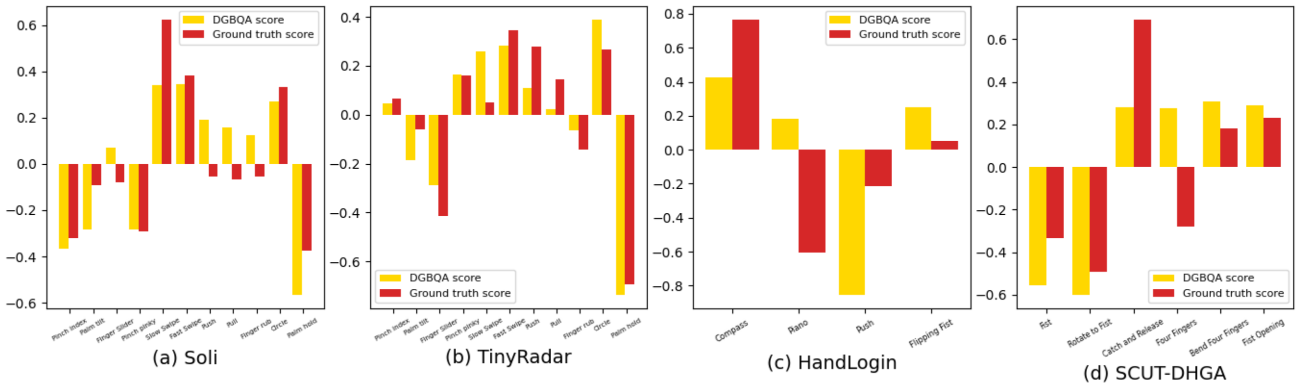
biometric characteristics.

To further validate our claims, in Fig. 3 we show the t-SNE plots for three different loss functions. It is clearly seen that using the proposed complete loss function formulation results in the most-suited feature space (refer Fig. 3, 3rd column). Using the other two loss function settings, either the identity details are shared across multiple gestures (Fig. 3, 2nd column), or the identity details are not preserved (Fig. 3, 1st column). Using larger values of λ_{ICGD}, even lower C_D values are achieved, however this also leads to disruption of the identity details. As f_θ is now further encouraged to make the embeddings orthogonal, rather than extracting gesture and identity-aware representations.

It is observed that models trained on TinyRadar and SCUT-DHGA datasets exhibit a smaller reduction in the ICGD score. This can be attributed to fewer embedding pairs being available for orthogonalization in a training batch. Specifically, these two datasets contain a higher number of identities, which reduces the possibility of a training batch containing embeddings from the same identity but with different gestures. In Fig. 4, we compare the DGBQA scores obtained by Res3D-ViViT and Res3D-MF models. It is seen that the scores obtained using the two models are similar and statistically not very different. This shows that the proposed DGBQA scores are generic and are only loosely dependent on the choice of the feature extractor.

2) Biometric Scoring Performance: We compare the DGBQA score values with the ground truth scores ($\hat{e}$). In Fig. 5, we show the Δ_{GBQA} and $\hat{e}$ scores for all the four datasets. DGBQA scores reported are based on the model that achieved the best performance in terms of nA_r while using the complete loss function (8). It is seen that the DGBQA scores are quite similar to the ground truth scores. Furthermore, relative rank positions of the gestures based on the DGBQA scores and $\hat{e}$ scores are also similar.

Comparison with State-of-the-Art Scoring Frameworks: To further validate the robustness of the proposed DGBQA framework, we compare with state-of-the-art scoring frameworks from other domains [21], [17], [2]. This comparison is on the basis of $\hat{r}$, while the biometric scores were computed using the same feature space from which DGBQA scores are derived. The results are listed in Table II and it is seen that the DGBQA framework achieves the lowest rank deviation in almost all the cases. This superior performance can be attributed to the uniqueness (10) and variability parameters (11) that have been captured in the proposed DGBQA scoring function (12). In contrast, the other frameworks either do not capture these parameters or



(a) Soli (b) TinyRadar (c) HandLogin (d) SCUT-DHGA

Fig. 5. Comparison of ground truth biometric scores ($\hat{e}$) and DGBQA scores (Δ_{GBQA}).

TABLE II
BIOMETRIC SCORING PERFORMANCE COMPARISON IN TERMS OF RANK DEVIATION ($\hat{r}$).

Model	Δ [21]	MasterFace [17]	Generative capacity [2]	Δ_{GBQA}
SOLI				
Res3D-ViViT	2.09	1.72	0.81	**0.45**
Res3D-MF	1.36	1.73	0.81	**0.45**
HandLogin				
Res3D-ViViT	1.00	1.00	1.00	**1.00**
Res3D-MF	0.50	0.50	0.00	0.50
TinyRadar				
Res3D-ViViT	3.63	1.45	1.27	**1.27**
SCUT-DHGA				
Res3D-ViViT	1.33	1.66	1.33	**0.66**

are based on certain assumptions which may not always hold. Nevertheless, these frameworks also attain significantly lower rank deviation. This indicates the robustness of the feature space formulation.

3) Gestures Good for Biometrics: From Fig. 5, we find that in the Soli and TinyRadar datasets, the swipe-based gestures (slow-swipe, fast-swipe) and the 'Circle' gesture achieve high DGBQA scores. While in the SCUT-DHGA and HandLogin datasets, gestures with the highest DGBQA scores were 'Catch and Release' and 'Compass', respectively. It is noted that all these gestures involve coarse but significant motion. Furthermore, they require the use of palm or fist. Thus, gestures involving coarse motions of palm or fist are more suitable for biometric applications.

IV. Conclusion

In this work, we developed the DGBQA framework for scoring the biometric characteristics of hand gestures. First, the proposed framework constructs a feature space suitable for biometric scoring. Next, this feature space is used to quantify the uniqueness and variability parameters of the gestures in order to estimate their biometric characteristics. We also presented several metrics for evaluating such scoring frameworks. Based on extensive experiments on four diverse datasets, the DGBQA scoring framework and formulation were found to perform very well. As part of our future work, we will work towards developing scoring formulations that are subject-agnostic. Furthermore, we will work on developing a universal feature extractor for biometric scoring.

References

[1] A. Arnab et al. ViViT: A Video Vision Transformer. In *Proc. International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[2] V. N. Boddeti, G. Sreekumar, and A. Ross. On the Biometric Capacity of Generative Face Models. *arXiv preprint arXiv:2308.02065*, 2023.

[3] T. Huang et al. Enhanced Spatial-Temporal Salience for Cross-View Gait Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6967–6980, 2022.

[4] A. K. Jain, D. Deb, and J. J. Engelsma. Biometrics: Trust, but Verify. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):303–323, 2021.

[5] G. Jaswal, S. Srirangarajan, and S. Dutta Roy. Range-Doppler Hand Gesture Recognition using Deep Residual-3D Transformer Network. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 311–315, 2021.

[6] H. Kong et al. Continuous Authentication through Finger Gesture Interaction for Smart Homes using WiFi. *IEEE Transactions on Mobile Computing*, 20(11):3148–3162, 2020.

[7] C. Liu et al. Dynamic-Hand-Gesture Authentication Dataset and Benchmark. *IEEE Transactions on Information Forensics and Security*, 16:1550–1562, 2020.

[8] O. Mendels, H. Stern, and S. Berma. User Identification for Home Entertainment based on Free-Air Hand Motion Signatures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(11):1461–1473, 2014.

[9] M. Patrick et al. Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 12493–12506, 2021.

[10] N. Sae-Bae, N. Memon, and P. Sooraksa. Distinctiveness, Complexity, and Repeatability of Online Signature Templates. *Pattern Recognition*, 84:332–344, 2018.

[11] M. Scherer et al. TinyRadarNN: Combining Spatial and Temporal Convolutional Neural Networks for Embedded Gesture Recognition with Short Range Radars. *IEEE Internet Things J.*, 8(13):10336–10346, 2021.

[12] X. Sheng et al. A Progressive Difference Method for Capturing Visual Tempos on Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):977–987, 2022.

[13] W. Song et al. TDS-Net: Towards Fast Dynamic Random Hand Gesture Authentication via Temporal Difference Symbiotic Neural Network. In *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021.

[14] W. Song and W. Kang. Depthwise Temporal Non-Local Network for Faster and Better Dynamic Hand Gesture Authentication. *IEEE Transactions on Information Forensics and Security*, 18:1870–1883, 2023.

[15] W. Song, W. Kang, and L. Lin. Hand Gesture Authentication by Discovering Fine-Grained Spatiotemporal Identity Characteristics. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[16] W. Song, W. Kang, and Y. Zhang. Understanding Physiological and Behavioral Characteristics Separately for High-Performance Video-based Hand Gesture Authentication. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[17] P. Terhörst et al. On the (limited) Generalization of Masterface Attacks and its relation to the Capacity of Face Representations. In *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2022.

[18] D. Thapar, A. Nigam, and C. Arora. Recognizing Camera Wearer from Hand Gestures in Egocentric Videos. In *Proc. ACM International Conference on Multimedia*, pages 2095–2103, 2020.

[19] S. Wang et al. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proc. Symposium on User Interface Software and Technology*, pages 851–860, 2016.

[20] J. Wu et al. Leveraging Shape and Depth in User Authentication from In-Air Hand Gestures. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 3195–3199, 2015.

[21] A. Zunino, J. Cavazza, and V. Murino. Revisiting Human Action Recognition: Personalization vs. Generalization. In *Proc. International Conference on Image Analysis and Processing*, pages 469–480, 2017.