entropy

MDPI

Article

Three-Dimensional Face Recognition Using Solid Harmonic Wavelet Scattering and Homotopy Dictionary Learning

Yi He [1], Peng Cheng [2,*], Shanmin Yang [3] and Jianwei Zhang [2]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China
[2] College of Computer Science, Sichuan University, Chengdu 610065, China
[3] School of Computer Science, Chengdu University of Information Technology, Chengdu 610065, China
* Correspondence: chengpeng_scu@163.com; Tel.: +86-1388-072-9172

Abstract: Data representation has been one of the core topics in 3D graphics and pattern recognition in high-dimensional data. Although the high-resolution geometrical information of a physical object can be well preserved in the form of metrical data, e.g., point clouds/triangular meshes, from a regular data (e.g., image/audio) processing perspective, they also bring excessive noise in the course of feature abstraction and regression. For 3D face recognition, preceding attempts focus on treating the scan samples as signals laying on an underlying discrete surface (mesh) or morphable (statistic) models and by embedding auxiliary information, e.g., texture onto the regularized local planar structure to obtain a superior expressive performance to registration-based methods, but environmental variations such as posture/illumination will dissatisfy the integrity or uniform sampling condition, which holistic models generally rely on. In this paper, a geometric deep learning framework for face recognition is proposed, which merely requires the consumption of raw spatial coordinates. The non-uniformity and non-grid geometric transformations in the course of point cloud face scanning are mitigated by modeling each identity as a stochastic process. Individual face scans are considered realizations, yielding underlying inherent distributions under the appropriate assumption of ergodicity. To accomplish 3D facial recognition, we propose a windowed solid harmonic scattering transform on point cloud face scans to extract the invariant coefficients so that unrelated variations can be encoded into certain components of the scattering domain. With these constructions, a sparse learning network as the semi-supervised classification backbone network can work on reducing intraclass variability. Our framework obtained superior performance to current competing methods; without excluding any fragmentary or severely deformed samples, the rank-1 recognition rate (RR1) achieved was 99.84% on the Face Recognition Grand Challenge (FRGC) v2.0 dataset and 99.90% on the Bosphorus dataset.

Keywords: solid harmonic wavelets; scattering representation; 3D face recognition; sparse dictionary learning

check for updates

Citation: He, Y.; Cheng, P.; Yang, S.; Zhang, J. Three-Dimensional Face Recognition Using Solid Harmonic Wavelet Scattering and Homotopy Dictionary Learning. *Entropy* **2022**, *24*, 1646. https://doi.org/10.3390/e24111646

Academic Editors: KC Santosh, Ayush Goyal, Djamila Aouada, Aaisha Makkar, Yao-Yi Chiang, Satish Kumar Singh and Alejandro Rodríguez-González

Received: 30 August 2022
Accepted: 8 November 2022
Published: 13 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

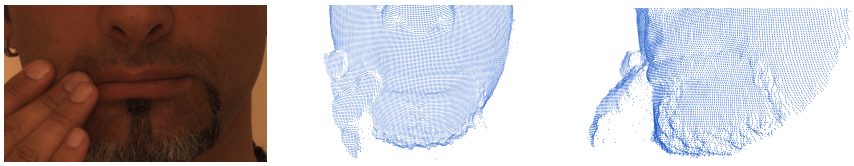
Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, point cloud and other metrical data have been utilized in multiple artificial intelligence applications; however, in the 3D face recognition research community, holistic-model-based methods still encounter issues: (1) there is a requirement of abundant training data to enable the capture of feasible features to form the large variations caused by the presence of rare and perturbing events, including occlusion/illumination/expression; (2) since the metrical data, e.g., raw point cloud, is generally sampled with structured high-resolution scanners with a restricted observing angle, some isometric deformations caused by exterior disturbances, e.g., pose variation or viewpoint variation, will inevitably confuse the inherent facial shape with sampling process noise and eventually, this issue will make raw point cloud representation behave like a non-uniform signal with dramatically varying intervals among points.

Solutions using the (semantic) model-matching process have achieved very high accuracy [1]; more recently, approaches using a 3D morphable model (3DMM) [2] have also achieved very high accuracy; however, the feasibility of key point detection and alignment relies on the constraints of large expression/pose variations.

Approaches without key point detection utilize intermediate planar models such as depth image [1,3] to resample point clouds onto the regular domain and then apply 2D deep frameworks to extract salient features. However, these methods rely on extra information—including texture/RGB channels—to enhance discrimination, which tends to be more vulnerable/unpredictable and is suspected of wasting metrical resolutions. The study in [4] was an early trial of utilizing scattering representation to solve 3D face recognition problems. Note that a specific prepossessing procedure was applied to implicitly project raw point clouds into canonical multidirectional depth maps (e.g., X/Y/Z—normal map components used as independent input channels); this approach then learned the patterns of each component, respectively, with 2D scattering convolution and concatenated the filters' responses as the global feature. This approach merely relied on geometric information to discriminate patterns from raw point clouds, but in more general scanning scenarios, complex and extrinsic disturbances, e.g., random 3D head poses/expressions/occlusions, may bring variations that affect both the signals on each predefined plane and the topology and geometry of the domain itself. For instance, the occlusion in a 2D raster image will result in some pop-up misplacements of the unrelated pixel values that may be highly diverged from the interested object, whereas in a 3D representation, the coarse shape frame is preserved with "additive" distortion/fragmentation in the local area (see Figure 1 for an illustration).



(**a**) (**b**) (**c**)

Figure 1. The different appearances of 2D and 3D representations: (**a**) the occlusion near the mouth in a raster image, (**b**) the same issue in a point cloud scan, (**c**) the second viewpoint of the occluded local area; the sample is from the Bosphorus dataset.

The benefit of geometric representation is that the integrity of the underlying structure can be expected and novel issues can be acting in multiple forms, as illustrated in Figure 2, where we picked one identity from the Bosphorus dataset with its coupled local variants. The first column shows the nose area in different (sampling) pose angles; it is clear that the integrity has been demolished and the frequency has been shifted (e.g., the nose tip area). Similar randomness emerged in the intrinsic deformations from expressions/occlusions, as demonstrated in the second column (the same subject's right eye area, with a neutral expression at the top and a surprised expression at the bottom) and the third column, where the mouth area was rendered both with and without occlusions. The above-mentioned methods, in this case, can spoil the surface assumption of faces, which leads to over-smoothed representations.

The last element that degrades the performance of the above-mentioned regular domain-based method is related to non-uniform sampling in relaxed scenarios, as illustrated in Figure 2, which shows that the extrinsic deformation also causes complicated local frequency shifting. As a result, the discriminative feature for recognition becomes crumpled into a wide frequency range. If we directly apply the (statistical) model-based methods (e.g., 3DMM) and compute the integration of the (dis)similarity measure along the predefined vertex paths or even along the regular domains (e.g., depth map), without the indispensability of homogeneity or the stability of such underlying geometry, the discrimination will probably become lost; this is because such predefined metrics will probably

regress to represent identity-unrelated properties such as (spectral) energy. Otherwise, if we eliminate these high-frequency variances through holistic denoising methods, the inter-subject dissimilarity will tend to be blurry since every face has an approximately consistent sketch shape.

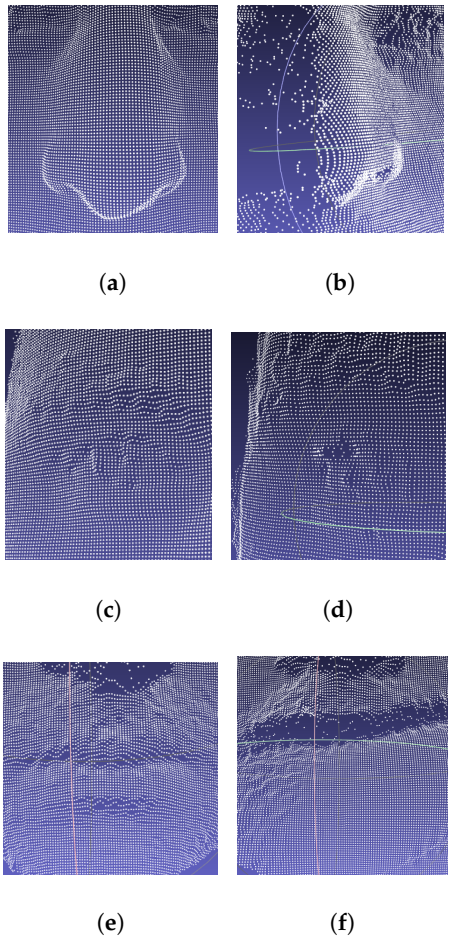


Figure 2. Low-fidelity local areas caused by pose variance/expression/occlusion. Top: (**a**) frontal; (**b**) posture (45°). Middle: (**c**) neutral; (**d**) surprised. Bottom: (**e**) non-occlusion; (**f**) occlusion.

We observed that the raw face scan samples naturally contained multiple species of noise/perturbations relating to sampling processes and other stochastic behaviors. Furthermore, by considering different facial identities as a family of stochastic processes, the identity-related properties were considered the underlying processes that merely related to the neutral and canonical appearance of the characterized physical face. In this setting, point cloud face recognition is similar to high-resolution texture discrimination, where Julesz gave the classical textons hypothesis [5,6]: by defining a finite collection of realizations of a texture, the statistical difference underlies preattentive discrimination. Accordingly, if we construct global measurements with desired invariants of unwanted variables, e.g., head poses/expressions, then the identity can be restored by comparing the "3D textons" [7] of a high-resolution human face (scan).

Through such intuition, face scans have several significant differences: (1) point clouds have no inherent correlation to the illumination/texture material, which shows the potential to directly obtain identity information merely from their geometry; however, the illumination/expression variations will lead to local deformations and (2) point clouds have no regular underlying cardinality, which disables the convolution-like operations and makes them an under-defined representation, so the results using convolution will be rather unstable with the order permutation. Therefore, we searched for a stable representation

against global rotations/order permutation and spatial translations that should still be capable of encoding sufficient high-resolution geometry; however, if accompanied by similar dictionary learning to find the co-occurrences of filter responses across diverged illumination and pose conditions, a sparser feature can be restored from such redundant representations.

To accomplish the above configuration, we first needed an alternative descriptor to capture features at the local scale. The micro-structures lie in scattered point sets so the representation should be regularized into a more canonical manner and the non-uniformness, e.g., frequency shifting, should be mitigated with a stabilizing operation.

Secondly, to build the 3D texton vocabulary (or more specifically, dictionary) with limited observations, we needed an efficient framework with the potential to disentangle variations according to an induced operational path, where certain desired invariance, stability, and consistency properties can be expected. Specifically, we proposed the use of a solid harmonic scattering transform [8] as a stabilizer and a local reference, lattice-based density estimation descriptor to capture such features for raw point cloud face recognition. Furthermore, we implemented a sparse scattering deep convolutional neural network based on [9] to build a local dictionary of the microstructures of a human face.

The proposed approach is illustrated in Figure 3.

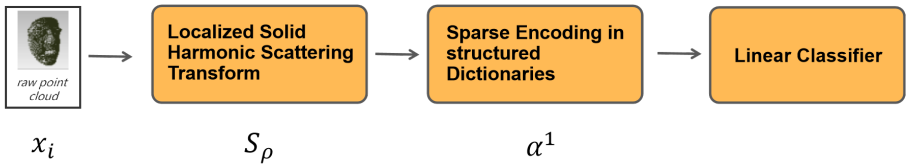


Figure 3. The sparse dictionary learning framework for 3D face recognition: the original solid harmonic wavelet scattering was adapted to piece-wise operations; with the sparse approximation coding from a union of learned dictionaries, the irregular raw point cloud faces can be classified merely through shapes.

In our approach, the raw point cloud faces are reorganized and down-sampled first with a generic resampling method such as farthest point sampling (FPS) [10]; then, one must consider the down-sampled points as entries to compute the nearest neighbors as the original local spatial signals. After having such subsets, we apply a reference lattice that originates at each entry point with a bump Gaussian function to estimate the probability of the appearance of a sampled point at such canonical grid-like positions and imitate a continuous function. After that, we apply a solid harmonic scattering transform to each point subset and obtain the representation of the 3D translation and rotation invariant scattering coefficients. With this stable representation, we then apply an iterative sparse thresholding coding (ISTC) [9] block to learn a discriminative feature space that reduces intraclass variability while preserving class separation through projections over unions of linear spaces.

Our major contributions are summarized as follows:

- Inspired by the high-resolution texture discrimination, this paper proposes a novel process–model approach to obtain the discriminative and stable facial features from pure point coordination representation for automatic face recognition; here, the facial shape clues are enhanced in a regularized domain.
- We modify the original holistic solid harmonic wavelet scattering transform approach into a windowed integral function to provide a higher-resolution representation.
- We learned a 3D facial texton dictionary, which is specifically based on the co-occurrences of filter responses across different extrinsic perturbances, e.g., head pose/illumination variation/occlusion, and succeed in achieving competitive recognition accuracy compared with alternative currently available methods.

This paper is organized as follows. In Section 2, we review the related works that have been published in recent years. In Section 3, we briefly recall solid harmonic scattering and sparse dictionary learning and then present our spatial construction for extracting the localized features of point cloud faces. The experimental results are stated in Section 4 and we conclude the paper in Section 5.

2. Related Works

From scene modeling to molecular imaging, metrical data such as point cloud/mesh data have been applied to the representation of multiple scaling geometries/structures of multitudinous objects or abstract concepts (e.g., graphs/manifolds), and the fast-growing field of geometric deep learning [11] has spawned multiple promising offspring; among these, here, we solely focus on those that directly utilize unstructured point cloud data.

2.1. Point Cloud Deep Learning

As immediate metrical results from the sensors, the points in the representations can carry intrinsic features in their natural manner, whereas regular, grid-like imaging sampling leads to the inevitable entanglement of environmental variants and objective signals; hence, this requires object detection/background segmentation, lengthy supervised feature abstraction, or data argumentation. The approach to consuming raw point clouds starts by resampling the points into continuous spatial voxels with statistic features that tend to lose higher-frequency components [12]; otherwise, an adapted multiple scaling strategy would be needed, which would require an excessive computation cost to cover the varying spectrum because of the sparsity and non-regularity of the general point cloud representation [13].

Methods for extracting order-invariant features on point clouds were discussed for Pointnet [14] and Pointnet++ [15], which provide point-wise feature and hierarchical representations. The authors of [2] utilized this idea to solve point cloud face classification; however, this approach requires the learning of a Gaussian process morphable model [16] to encode the holistic features of real face samples to mitigate the intraclass variances from the face-scan phase.

Alternative improvements involve the construction of more flexible underlying affinity structures and learning features through them; multiple challenging problems in metrical data have been solved—with satisfying results—using these techniques, including semantic scene segmentation/3D object classification [14,17,18].

2.2. Dictionary Learning on Scattering Coefficients

A scattering transform has the ability to mitigate undesired group-structured operations in well-defined domains, e.g., image/voice signals, with predetermined wavelet filter banks [19–21]. Predefined harmonic wavelets can bring translation–rotation invariances and linearize isometric deformations and have been utilized as a superior tool for representing molecules' fine 3D geometry, namely the solid harmonic wavelet scattering transform [8].

However, unlike atoms' orbital positional distributions, a raw face point cloud yields a coarse underlying smooth surface but behaves with non-guaranteed differentiability. The Euclidean learning methods would probably fail to converge. The authors of [22] proposed an approach for the overlap of multiple smoothed position signals, allowing for the representation of periodic structures. The idea of combining a scattering transform approach with a deep network has also been developed [23,24]; additionally, a recent attempt at solving complicated classification problems with scattering representation achieved a remarkably fast-converging performance with potential in mathematical analytic implementations [9]. Furthermore, their work relies on a classical, active technology—sparse dictionary learning [25]. We utilized this idea to select discriminative signal components to prevent our model from regressing to a representation of irrelevant deformations.

2.3. 3D Face Recognition

As discussed in [2,26], research on the 3D facial recognition problem has developed in several major directions: (local) feature-based, (holistic) model-based, and matching-based methods. From a general perspective, we can see an underlying trend of reducing the necessity of the registration phase/domain matching—on account of the development of acquisition techniques that enforce more regular raw scanning results—while stronger computation methods and facilities progressively enable parallel processing on high-throughput data streams. As a characteristic indicator method, the iterative closest point (ICP) [27] matching scheme played a significant role in [28–31], where the above isometric deformations have been eliminated by spatially aligning faces into a common direction.

Holistic methods have been developed in recent years to reduce the necessity of registration; for example, [32] proposed the Markov random field (MRF)-model-based approach to select discriminant features on posterior marginal probabilities. Later works, such as [1,3], concentrated on adapting a 2D learning framework for 3D scans, where a deep range image has commonly been utilized for medium representation, which might partially ignore the indifferentiability in spaces, e.g., 3D rotation by mass-supervised training. The authors of [2] provided an a priori model-based argumentation strategy to avoid the above-described question.

Owing to space constraints, we only named a few of the relevant works that studied the 3D face recognition problem; as we stated, a study on the construction of more intrinsic representations for facial recognition is necessary.

3. Materials and Methods

In this section, we first present our approach to modeling identities under noisy environments and a method for constructing stable representations; additionally, we describe the sparse dictionary learning structure for feature selection. Finally, we present our overall framework for effective jointly learning discriminative representations.

3.1. The Stochastic Process Model on Point Cloud Faces

Let one identity be denoted as X_i, with i ranging through the different identities; we consider point cloud faces as realizations of a modeled noisy observation process as follows:

$$x_i \in L_\theta(X_i) \tag{1}$$

where each sample only provides spatial coordinates and can be expressed as a vector $x_i = \{\mathrm{r}_k \in \mathbb{R}^3 : k = 1 \dots K\}$, where K is usually in 10k magnitudes. The L_θ is a function that models the above geometric transformation, illumination, and occlusion variances. In addition, the θ can be seen as a low-dimension random vector encoding the global illumination and rigid affine transformations [33], though in general, the ergodicity of L_θ is hard to be satisfied.

As a possible solution, a graph-based method [14] applied a dynamic approximation procession to transform raw point clouds into uniform point/vertex sequences with lengths of thousands; then, it was used to compute the corresponding embedded features in order to imitate a universal characteristic representation.

This method shed light on defining signals with near-independent distributions between global and local variables; however, it required heavy training to realize the asymptotic stability, which is neither available nor necessary in modeling more consistent structures, e.g., faces, where geometric deformation and/or illumination/pose variances will not lead to large universal interferences.

Upon these observations, we built Euclidean lattices and learned the spatially aware features using solid scattering transform-based local operators; by applying subsequent sparse dictionary learning in the scattering domain, the uncorrelated signal components of the identity in question were jointly learned and inhibited.

The local 3D lattice operator: We used p_0 to denote the centroid of x_i, which was easy to obtain, and from p_0 we constructed the 3D global coordination, M_{xyz}. The succeeding step used farthest point (FP) [10] sampling on x_i; note that we only needed to draw the countable $C < 200$ points as query points, $\mathrm{P}_0 = \{p_c\}_{c \leqslant C}$, for the subsequent spatial thresholding nearest-neighbor searching. From each p_c, we drew the N nearest points from their ambient spaces to form a leaf subset:

$$\mathrm{P}_c = \{r_{c,n} \in \mathrm{N}_{R_c}(p_c) : diam(p_c) < R_c\}_{n \leqslant N} \tag{2}$$

where we picked a threshold radius—$R_c = \min\|p_c - P_0 / \{p_c\}\|$—to dynamically assure coverage. Furthermore, within each ambient space, we associated a 3D local lattice coordination $\mu_{xyz} \subset M_{xyz}$ and defined the overall density estimation function as the concatenation of C local areas:

$$\hat{\rho}_\Omega(\mu) = (\hat{\rho}_1, \ldots, \hat{\rho}_c, \ldots, \hat{\rho}_C) \tag{3}$$

where each $\hat{\rho}_c$ was parameterized by

$$\hat{\rho}_c(\mu) = \sum_{n=1}^{N} G(\mu - r_{c,n}) \tag{4}$$

which is a sum of the Gaussian densities centered on each $r_{c,n}$. This spatial construction sliced each x_i into C local receptive fields; we adjusted the width parameter, σ, of the Gaussian equivalent to the distance from the nearest alternative entry point, i.e., $\sigma_c \rightarrow \sup(\|r_c - r_c'\|) \forall r \in N_c$. By renominating each ρ_c with indicator function I_{ρ_c}, a raw point was transformed into a naive Borel set, which had a uniform probability measure, so we defined the global piece-wise density function as $\rho_\Omega = \bigcup \rho_{c \leqslant C}$.

This approach encoded a raw point cloud face into a more regular continuous probability density representation, with local fields being invariant to the permutations of the input order; each characteristic vector also had a corresponding length, which enabled the windowed operations (See Figure 4 for illustration).

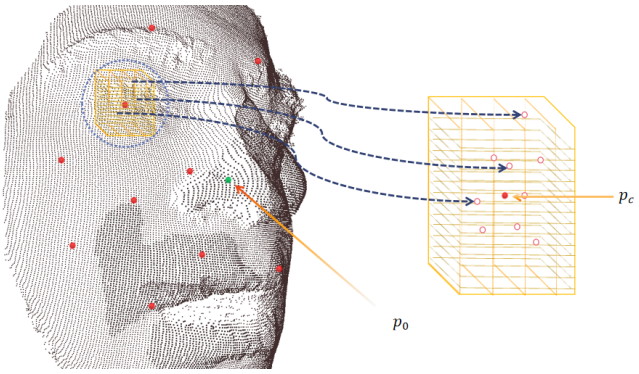


Figure 4. Left: The sparse entry points (in red) obtained by the FPS algorithm. **Right**: The local 3D lattice patch operators used to cover the ambient spaces around the entry points.

However, the above isometric deformations not only broke the order consistency but also gave rise to mixed deformations and polluted the geometric features; therefore, we needed to add a stabilizer to obtain the rotation and translation invariances.

To illustrate the operation, a piece of pseudocode is given below as Algorithm 1:

Algorithm 1 The Local Lattice Operation:

Require: $p_0 = (x_0, y_0, z_0)$ (the centroid of a raw face scan x_i), $x_i = \{\mathrm{r}_k \in \mathbb{R}^3 : k = 1 \dots K\}$

Set p_0 as the initial point of farthest point sampling and
Draw C points $\{p_c\}_{c \leqslant C}$ from x_i
for p_c in $\{p_c\}_{c \leqslant C}$ **do**
  Set p_c as the origin, and
  Compute $P_c = KNN(p_c, N)$
  Compute local lattice $\mu = \{m \cdot dx, n \cdot dy, o \cdot dz\}$
  Compute local density estimation $\hat{\rho}_c \leftarrow \sum_{n=1}^{N} G(\mu - r_{c,n})$
end for
Concatenate local densities to form the overall function as $\hat{\rho}_\Omega(\mu) = (\hat{\rho}_1, \dots, \hat{\rho}_c, \dots, \hat{\rho}_C)$
return $\hat{\rho}_\Omega$

Note that since the direction of local lattice μ was exactly covariant to the global coordination of the scanned face, the above-obtained local density feature actually exposed itself to a risk of being sensitive to rigid rotation, as well as to the order permutation of the grid position points (see Figure 5 for an illustration). Therefore, we needed to construct a stabilizer to eliminate such isometry.

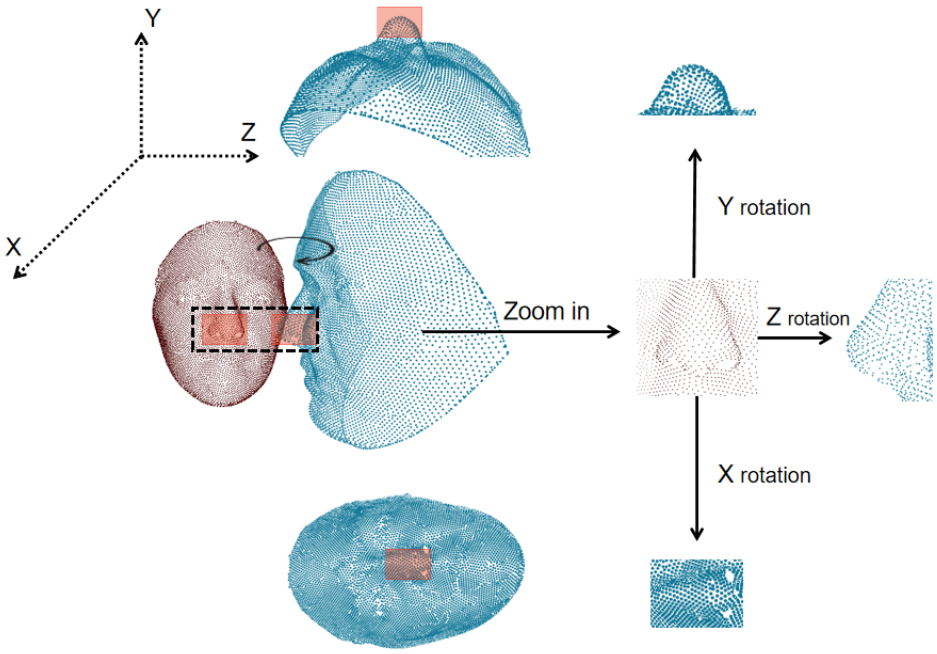


Figure 5. The rigid rotation caused by pose variations; this figure merely gives several discretized realizations. The point subsets in blue indicate the rotated version of the original local areas around the nose tip.

Windowed solid harmonic wavelet scattering: A scattering transform is a geometric deep learning framework that replaces learning-based cross-correlation kernels with predefined filter banks. Induced stability for multiple species of isometrics and translation invariances can be prescribed with a group-invariant structure built from a deliberately configured wavelet filter bank [19,20]. For 2D signals (e.g., images), the constitutive substructure in a scattering network comprises the wavelet filters with zero integrals and yields fast decay along $\|\mu\|$; each can be parameterized by a rotation parameter, θ, and dilation parameter, j, as

$$\psi_{j,\theta}(\mu) = 2^{-2j}\psi(2^{-j}r_{-\theta}\mu) \tag{5}$$

where $r \in G$ belongs to a finite-rotation group of $\mathbb{R}^d$.

For 3D signals—as in 3D face recognition with point cloud samples—3D rotation invariance is crucial since the random pose variation may provide an alias for the local density feature obtained by our local lattice operator (see Figure 5). Accordingly, we built a stabilizer in the solid harmonic scattering approach from[8], whereby solving the Laplacian equation with the 3D spherical coordinates and replacing the exponent term in the spherical harmonic function, Y_ℓ^m, the solid harmonic wavelet can be expressed as follows:

$$\psi_{\ell,m}(r,\theta,\varphi) = \frac{1}{\sqrt{(2\pi)^3}} e^{-1/2r^2} r^\ell Y_\ell^m(\theta,\varphi) \tag{6}$$

In addition, by summing up the energies over m, a 3D covariant modulus operator can be defined as

$$U[j,\ell]\rho(\mu) = \Big(\sum_{m=-\ell}^{\ell} \Big|\rho \star \psi_{\ell,m,j}(\mu)\Big|^2\Big)^{1/2} \tag{7}$$

In short, a solid harmonic scattering transform is defined as the operation of summing up the above modulus coefficients over μ to produce translation-/rotation-invariant representation within each local field (see Figure 6 for illustration). Furthermore, by raising $Ux[j,\ell]\rho(\mu)$ to exponent q and then sub-sampling μ at $2^{j-\alpha}$ with an oversampling factor—$\alpha = 1$ to avoid aliasing, the first-order solid scattering coefficients are

$$S[j_1,\ell,q]\rho = \sum_{\mu} \Big|U[j_1,\ell]\rho(2^{j_1-\alpha}\mu)\Big|^q \tag{8}$$

Then, by iterating subsampling at intervals $2^{j_2-\alpha}$ with $j_2 > j_1$ and recomputing the scattering coefficient on the first-order output, we obtained the following second-order scattering transform:

$$S[j_1,j_2,\ell,q]\rho = \sum_{\mu} \Big|U[j_2,\ell]U[j_1,\ell]\rho(2^{j_2-\alpha}\mu)\Big|^q \tag{9}$$

These representations can hold local invariant spatial information up to a predefined scale, 2^J; in our case, this was adjusted to be equivalent to the local threshold diameter, N_{R_c}. Furthermore, we needed to extend this operation to a universal representation. Here, we defined the windowed first and second solid harmonic wavelet scattering as follows:

$$S_\cup[j_1,\ell,q]\rho_\Omega = \bigcup_{c=1}^{C} |U[j_1,\ell]\rho_c|^q \tag{10}$$

$$S_\cup[j_1,j_2,\ell,q]\rho \bigcup_{c=1}^{C} |U[j_1,\ell]U[j_2,\ell]\rho_c|^q \tag{11}$$

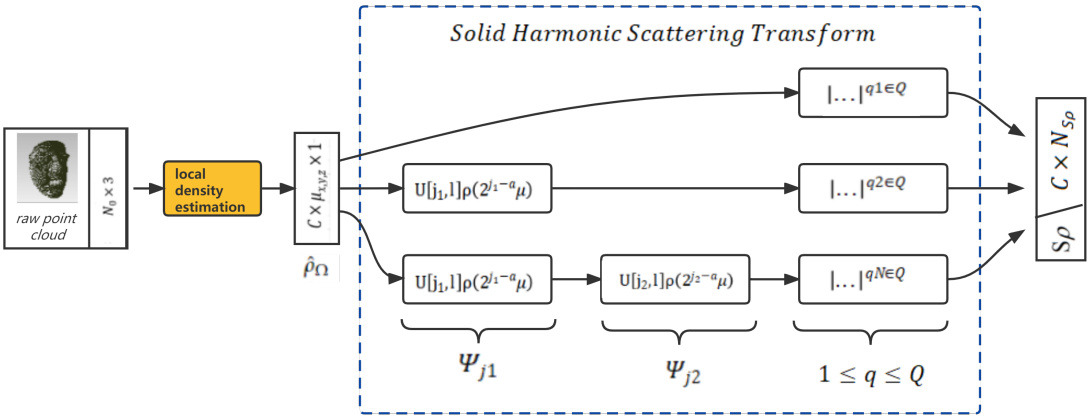


Figure 6. Illustration of the localized solid harmonic scattering transformation; the dashed blocks represent the extracted invariant representations from each local ambient space.

For a better illustration, a brief pseudo-code is stated below as Algorithm 2:

Algorithm 2 Windowed Solid Harmonic Wavelet Scattering

Require: $\hat{\rho}_\Omega$ (local density features), J (scale parameter), L (rotation phase parameter), Q (exponential parameter)

Set wavelet t$\psi_{\ell,m}(r,\theta,\varphi)$ according to a predefined parameter (J,L) with Equation [6]
for ρ_c in $\{\rho_c\}_{c\leqslant C}$ **do**
  for $0 \leqslant j \leqslant J$ **do**
    Compute the dilated modulus operation on scattering convolution $\rho_c \star \psi_{\ell,m,j}(\mu)$, e.g., Equation [7]
  end for
  Compute the first-order coefficients $S[j_1,\ell,q]\rho$ as Equation [8]
  Compute the second-order coefficients $S[j_1,j_2,\ell,q]\rho$ as Equation [9]
  Concatenate first and second coefficients as the local invariant representation S_{ρ_c}
end for
Concatenate $\{S_{\rho_c}\}_{c\leqslant C}$ as the global invariant representation S_{ρ_Ω}
return S_{ρ_Ω}

3.2. Piece-Wise Smoothed Solid Harmonic Scattering Coefficient Representation

The above strategy makes the representation stable to local deformations, and since face point clouds share a largely consistent global structure, it allows us to represent them even if no effective global embedding exists.

To balance the computation complexity and resolution in our experiments, we chose $C = 128$, $J = 7$, and $q \in Q = \{1/2, 1, 2, 3\}$; here, the above windowed operation and scattering coefficients were implemented with the Kymatio software package [34]. To simplify our notation, we wrote the scattering representation in shorthand as follows:

$$S\rho_c = \{S[p]\rho_c\}_p \tag{12}$$

where p is the union of the first and second indices $\{(j_1,\ell,q)$ and $(j_1,j_2,\ell,q)\}$, respectively, and the overall scattering coefficients of a point cloud face are

$$S\rho_\Omega = \{S\rho_c\}_{c\leqslant C} \tag{13}$$

To give an illustration of this representation, we mapped the first-order scattering coefficients of two identities (bs001 and bs070) onto the scattering indices shown in Figure 7; identity bs001 from Bosphorus had two realizations and we could see that, although there

was a significant visual difference between them, their scattering coefficients had a similar appearance.

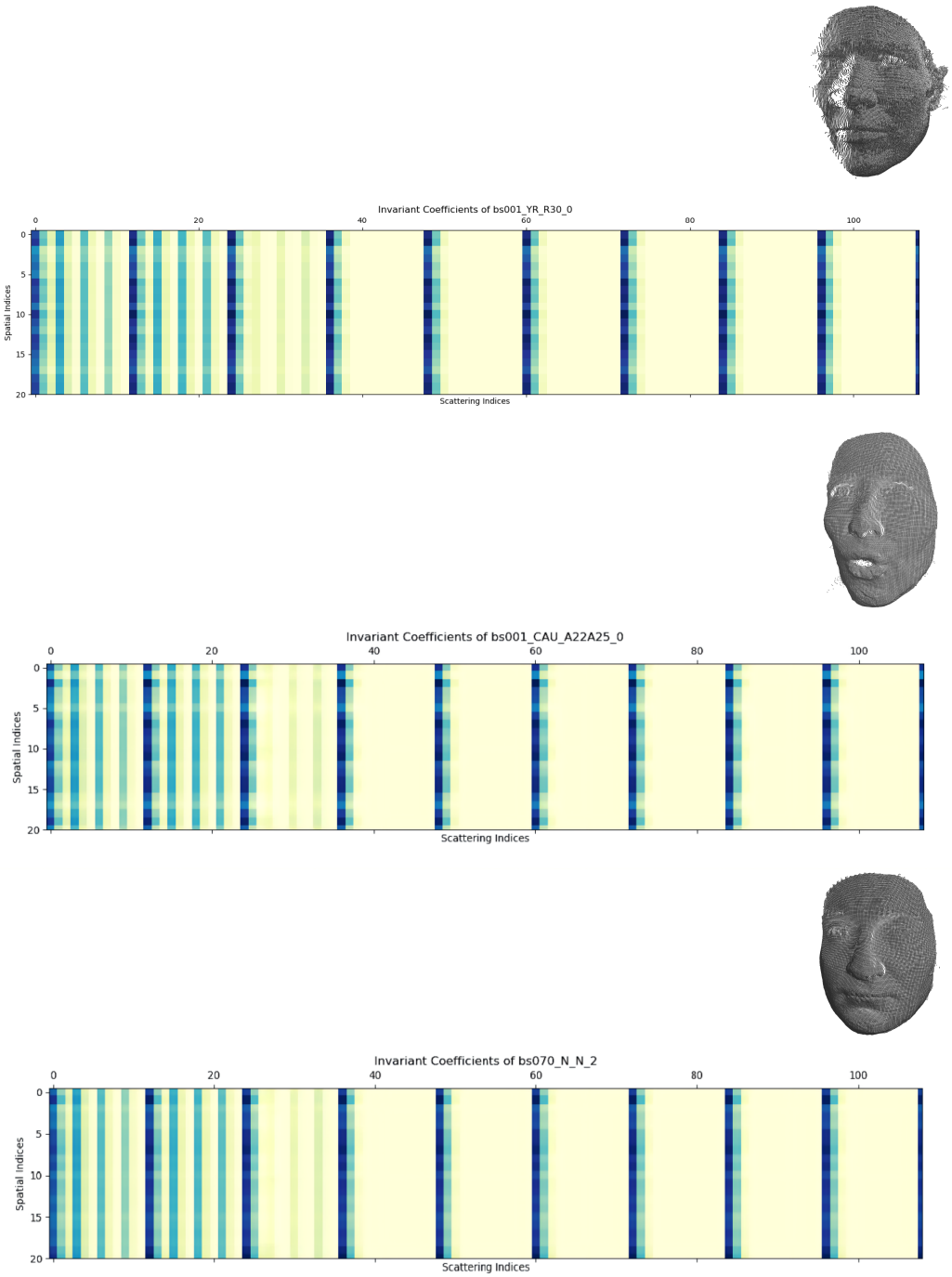


Figure 7. Scattering invariant coefficient representations: top—identity bs001 from Bosphorus with a 30° yaw rotation; middle—bs001 with an action unit combination (expression); bottom—bs070 with a neutral frontal position.

To enhance the discrimination of this representation—and inspired by [7]—in the next section, we construct a "facial scattering coefficients dictionary" from the above representation to associate multiscale properties for 3D facial recognition. Specifically, based on the good results from the 2D scattering coefficients in [9], we follow their idea for

utilizing supervised dictionary learning to select the most relative classification features from the 3D scattering coefficients.

3.3. Constructing a Local Dictionary with Semi-Supervised Sparse Coding on Scattering Coefficients

The scattering representation brings desired properties including Lipschitz continuity to translations and deformations [19]; however, the overall structure constructed using FPS and local nearest searching and normalization assumes uniform energy distribution among the realizations of each identity. In real scan scenarios, this assumption can be damaged when permutations, e.g., occlusion/rigid overall rotation, break the integrity of the face samples. In some severe cases, a certain portion of points are missing from face samples in Bosphorus. To reduce this category of intraclass variance, we imposed the homotopy dictionary learning framework presented by [9] to build a local coefficient dictionary. Then, we trained the network to select the most relative classification features from the scattering coefficients.

Supervised Sparse Dictionary Learning: The idea of selecting the sparse combination of functions from redundant/over-complete dictionaries to match the random signal structure was presented by [35] and flourishes in multiple fields related to signal processing. Supervised dictionary learning was first presented by [36] to solve the following optimization problem:

$$\min_{D,\Theta} \sum_{j} \ell(x_j, y_j, \Theta, \alpha^*(x_j, D)) \tag{14}$$

where Θ indicates a simple classifier's parameters; ℓ is the loss function for computing the penalization on the prediction (x_j, y_j); α^* is the sparse code of the input signal, x_j, with the learned dictionary, D.

In our problem, the input signal, $S\rho_\Omega$, had a union form; hence, we constructed a global dictionary with structured local dictionaries defined as:

$$D = [D_1, ...D_c..., D_C] \tag{15}$$

where $\{D_c\}_{c=1}^{C}$ are C sub-dictionaries with a certain structure—$D \in \mathbb{R}^{K\times C\times N}$. Here, K indicates the length of the local pseudo coordination, p, of $S\rho_\Omega$; the aim was to represent B input samples (B—batch size) as linear combinations of D's elements. Each D_c had $N = 512$ normalized atoms/columns—$\{d_n\}_{n=1}^{N} \in \mathbb{R}^K$. Then, the sparse approximation optimization was used to solve

$$\arg\min_{D,\alpha_i} \sum_{i=1}^{B} \|S\rho_i - D\alpha_i\|_2^2 + \lambda_* \|\alpha_i\|_0 \tag{16}$$

where α_i is the concatenated sparse codes $\alpha_i = [\alpha_{i,1}, ...\alpha_{i,c}..., \alpha_{i,C}]$. Suppose the optimized sparse coefficient matrix is $A \in \mathbb{R}^{K\times N\times C\times B}$ for a batch of input signals, $\{S\rho_i\}_{i=1}^{B} \in \mathbb{R}^{K\times C\times B}$, where each sub-dictionary has a local code $\alpha_c \in \mathbb{R}^{N\times B}$..

Expected Scattering Operation: Since we regrouped the raw point clouds and individually computed the invariant representations, $S\rho_c$, the windowed representation also had a non-expansive property; within each local field, the translation converged to being negligible by taking $J \to \infty$.

In practice, this will possibly bring ambiguity. By setting a small C, each field becomes too large and results in the loss of higher-frequency components. Yet, for a larger C, the computing complexity amounts to $O(CS)$, and the optimization of such a concatenation will lead to supernumerary consideration, e.g., vanishing gradients.

Thanks to the integral normalized scattering transform [20], which preserves a uniform norm by utilizing the non-expansive operator $\overline{S}_C$, we considered our question of structured learning for some random processes using the supposed condition. For the underlying distributions of point cloud faces yet to be established in practice, we focused on finding

a solution with the above intuition and incautiously assumed our representation to be a stationary process up to negligible higher components; thus, the metric among (a batch of) spatial realizations reduced to a summation of the mean-square distances is

$$\Delta(S\rho - DA) := \frac{1}{BC}\sum_{i}^{B}\sum_{c}^{C}\|S\rho_{i,c} - D_c\alpha_{i,c}\|_2^2 \tag{17}$$

This definition is simple but effective as a regression term with a forward–backward approximation, which is based on an operation called proximal projection [37].

$$\begin{aligned} A^* = prox_\lambda(S\rho)\forall(S\rho, A) \in \mathbb{R}^{N=KCB} \times \mathbb{R}^{N=KCB} \\ \Leftrightarrow A - S\rho = \Delta(S\rho - DA) \end{aligned} \tag{18}$$

where it encloses a solution with a forward step by computing a reconstructed $\widetilde{S}\rho$ and a backward step by putting it back into the proximal projection operator, updating λ and D. Since our aim was to implement an efficient classification model, the sparse code should be able to preserve the principle components of the input signal; additionally, with the experimental observation of the point cloud faces' solid scattering coefficients, we saw most energy being carried by its rare lower-frequency components and characterized by larger-magnitude coefficients; therefore, we picked the recent generalized ISTC algorithm [9], which adopts an auxiliary dictionary to accelerate convergence. Here, the ReLU function acts as a positive soft thresholding implementation of proximal projection. Then, the optimization can be reached in an unsupervised $n \leqslant N$-iteration-updating scheme, expressed as follows:

$$\alpha_n = \mathrm{ReLU}(\alpha_{n-1} + D^T(\beta - D\alpha_{n-1}) - \lambda_n) \text{ for } n \leqslant N \tag{19}$$

The overall architecture is shown in Figure 8.

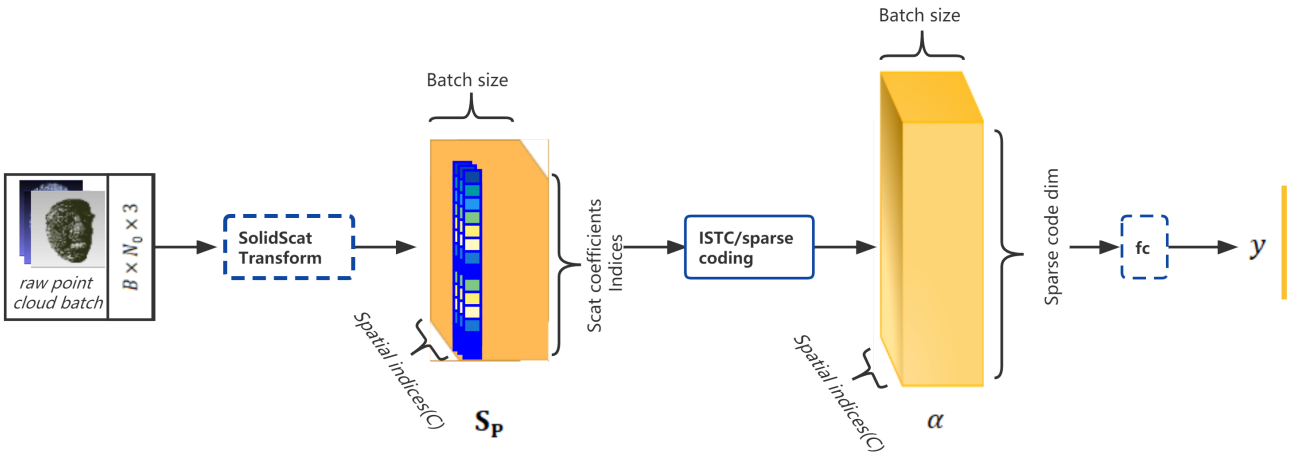


Figure 8. The architecture of the pure-spatial-coordination-based framework for 3D face recognition

To effectively demonstrate our methods, a pseudo-code is given in Algorithm 3:

Algorithm 3 Dictionary learning on local facial coefficients

Require: $\{(y, x)\}$ (training set), D (initial dictionary), λ_1 (initial Lagrange multiplier, e.g., thresholding bias), θ (classifier parameter), N (number of iterations), τ (learning rate), v (regulation parameter)

Draw a batch of samples$\{(y_j, x_j)\}_{j \leqslant Batchsize}$ and compute the scattering coefficients using the methods in Section 3.2; denote the coefficient vectors as $\{\beta_j\}_{j \leqslant Batchsize} = \{S_{\rho\Omega}\}_{j \leqslant Batchsize}$
for $1 \leqslant j \leqslant Batchsize$ **do**
  for $1 \leqslant n \leqslant N$ **do**
    Compute $\alpha_n = \alpha_{n-1} + D^T(\beta_i - D\alpha_{n-1}) - \lambda_n$ where $\alpha_0 = 0$
    Compute $\alpha_n^\star = ReLU_{\lambda_n}(\alpha_n)$
    Update $\lambda_n = \lambda_{max}\left(\frac{\lambda_{max}}{\lambda_\star}\right)^{-n/N}$
  end for
  Compute $\lambda_N = \lambda_\star$ and $\alpha_N = ReLU_{\lambda_N}(\alpha_{N-1})$
end for
Compute the classification loss $\sum_j Loss(D, \lambda_N, \theta, \beta_j, y_j)$
Update the parameters by a projected gradient step [36]
$\theta \leftarrow \prod_\theta \left[\theta - \tau(\nabla_\theta Loss(D, \lambda_N, \theta, \beta_j, y_j) + v\theta\right]$,
$D \leftarrow \prod_D \left[D - \tau(-W^T D\alpha_N + \Delta(S\rho - DA)\right]$
return Learned Dictionary D

4. Results

The goal of this paper was to construct a geometric deep face recognition learning model that relies merely on geometrical features and eliminates complicated embedding procedures, as well as hand-crafted feature alignment. We compared the performance of our approach with those of relevant methods using two well-known datasets—Bosphorus [38] and Face Recognition Grand Challenge v2.0 [39]. We found that our structure was more analytic and obtained competitive results. In Section 4.1, we describe our evaluation protocol and metrics, and in Section 4.2, we detail our implementation and present some parameter analyses. In Section 4.3, we demonstrate the parameter tuning process. The main results are presented in Section 4.4.

4.1. The Evaluation Protocol and Metrics

We applied the general evaluation protocol by comparing the rank-1 recognition rate (RR1) and the verification rate (VR) with the false-accept/positive rate (FAR/FPR) = 1×10^{-3} as the key performance metric. We compared our method with other competing methods, where the rank-1 recognition rate (RR1) was defined as the proportion of positive label predictions out of the total number of predictions for the whole test set. The total number of label predictions consisted of the sum of the positive and negative predictions. By further clarifying the positive results into the true-positive rate (TPR) and the false-accept/positive rate (FAR/FPR), the verification rate was defined as the portion of positive results under a certain false-accept/positive rate (FAR/FPR).

4.2. Implementations

The backbone of our implementation was based upon the work of [9], who achieved competitive results in 2D image classification problems with a mathematical analytical structure. In our study, the data structure had prominent differences; therefore, we needed to perform modifications, as follows.

(1) The 3D solid scattering coefficient representation: As introduced in Section 2, we transformed the raw point cloud into representative zero-, first-, and second-order cascades of the solid harmonic scattering coefficients (shown in Figure 5); the implementation was based on an open-source framework [34].

The typical size of a sample in the Bosphorus and FRGCv2 datasets ranges from 30 k to 100 k; as discussed in Section 3.1, a sufficient C of partitions can reduce the error in estimating the dimensional metric; however, a rising C requires increasing arbitrary coefficients, and we needed to balance the computation load between requesting the fineness of the representation and processing to obtain high-efficiency recognition.

For instance, the solid harmonic scattering transform on a local receptive field had $|Q|JL + |Q|JL(J-2)/2$ invariant coefficients as outputs; we fixed $J = 7$ and $q \in Q = \{1/2, 1, 2, 3\}$ as the principle settings on the solid scattering transform process and set $C = 128$ as the number of local fields. Within each field, we applied a spatial 3D grid with $8 \times 8 \times 8$ reference positions, which means we utilized $128 \times 8 \times 8 \times 8$ floats as each sample's density representation; the scattering coefficient representation had a constant dimension in the locality as 84 first-order and 252 second-order coefficients, with 128×336 floats as the overall representation.

(2) The sparse dictionary learning structure is demonstrated in Figure 6. It remained a very wide feature vector when we directly input a batch of scattering coefficients into the ISTC layer; therefore, we applied a 15×1 convolution operation with batch normalization to reduce it to 128×200. Furthermore, it included 3.8×10^5 learned parameters. The N was set to 3 since it was experimentally sufficient to allow the sparsity to reach the extremum.

The whole framework is illustrated in Figure 6, where the sub-dictionary, D_c, had 512 atoms, $\{d_i\}_i$, each with 1×1 support to provide an initial low correlation among the atoms and it met the over-complete condition for each of the locally projected scattering coefficients. This ISTC network took as input an array, $LS\rho$, of size 128×200, and output a sparse code matrix of 128×512. The total number of learned parameters in D was about 6.5×10^4. The number of parameters in L and D was around 4.5×10^5 in total. Furthermore, in order to generate a classification ability for simulating real human face recognition situations, we adopted a simple linear classifier and evaluated the performance; the results are presented in the following sections.

4.3. Hyperparameter Tuning Process

The proposed network was implemented with PyTorch Version 3.7 (Initially Released in September 2016 by Meta AI, Astor Place, New York City, US) [40] and was trained with i7-8700K CPU and a single GTX2080TI GPU.

We used the Face Recognition Grand Challenge (FRGC) v2.0 dataset and the Bosphorus dataset to evaluate the performance of the proposed face recognition method. It took about 15 hours for our hardware to train the network on each dataset, and our method obtained a high accuracy on FRGCv2 with a short training procedure. Since FRGCv2 has a relatively regular and uniform sampling process—which can be observed in Figure 9 as a general case—it helped to verify our hypothesis. In order to adequately clarify the capability of our framework, we mainly applied the rank-1 recognition rate on the full Bosphorus dataset for the ablation study.

Figure 9. One identity from FRGCv2 with three scans.

There were three major parameters introduced in the solid harmonic scattering—namely, J, L, and C—which decided the width of the scattering coefficients as the input for the dictionary learning phase; another combination of $dim_{Dictionary}$ and $dim_{LinearProj}$ decided the number of parameters for the dictionary learning. With the experimental observations, the solid harm coefficients had a stepped magnitude distribution along their support, which could be in great disparity, which led to aliasing during the normalization among local patches when we picked insufficient scaling levels, J; however, if we applied a big J, it required excessive computation in return. Given this, we first went through a wide combination space of the above parameters and obtained preliminary results, as presented in Table 1, where we can see a trend in performance improvement in the increasing J, $dim_{Dictionary}$, and $dim_{LinearProj}$. Then, we utilized two-stage training with individual parameter searching, as follows.

(1) Parameters in Solid Harmonic Scattering: Figure 10 demonstrates the rank-1 recognition rate on the Bosphorus dataset by training the network with J = 3, 4, 5, 6, 7, 8, 9 values under C = 128. It can be seen that the dimension of each sub-dictionary had to be subsequent to satisfy overcompleteness; it can be seen from the blue/green lines that when $dim_{Dictionary} > dim_{scattering}$ and $J > 6$, the recognition rate barely grows.

(2) Parameters in Sparse Dictionary Coding: We fixed $dim_{Dictionary} = 512$; here, we found that a variation in J in (5, 7, 9) reached its best spot on $dim_{Linear_Proj} \geqslant 150$. Figure 11 depicts the varying performance; we applied [J = 7, dim_{Linear_Proj} = 150, C = 128, $dim_{Dictionary}$ = 512] as the principle experimental configuration of this framework.

Table 1. Rank-1 recognition rate (RR1) under different combinations of (J, L, C) and $(\mathrm{dim}_S, \mathrm{dim}_L, \mathrm{dim}_D)$ on Bosphorus dataset.

(J, L, C)	$(\mathbf{dim}_S, \mathbf{dim}_L, \mathbf{dim}_D)$	RR1
5, 3, 128	252, 120, 512	90.42%
6, 3, 128	336, 120, 512	93.54%
7, 3, 128	432, 120, 512	88.65%
6, 3, 128	336, 150, 512	91.88%
6, 3, 128	336, 200, 512	95.42%
7, 3, 128	336, 150, 512	95.63%
7, 3, 128	432, 200, 512	99.49%

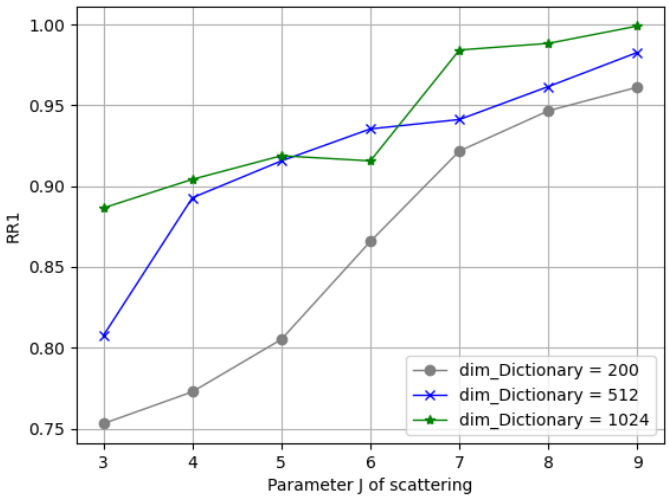


Figure 10. Comparisons of different scattering parameters.

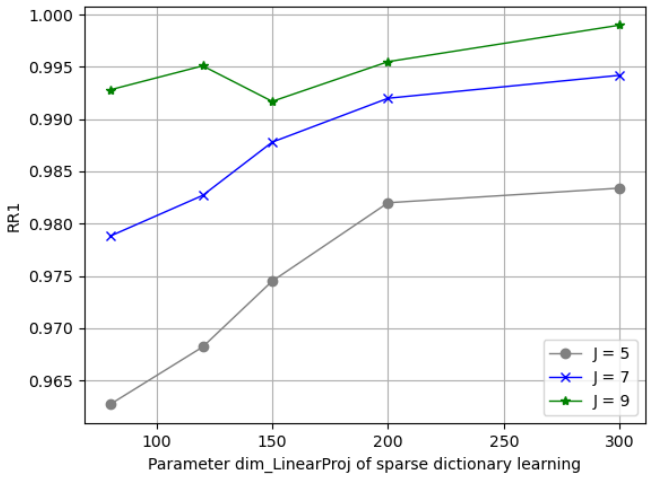


Figure 11. Comparisons of different sparse dictionary learning parameters.

4.4. Comparison with Other Methods

(1) Results on the FRGCv2 dataset: The FRGC v2.0 dataset [39] contained 4007 scans of 466 subjects in total; we followed its protocol to train on the Spring2003 partition and used the remaining data for testing. The results of running our proposed method and the state-of-the-art methods on the FRGC v2.0 dataset are shown in Table 2. The methods that used corresponding 2D photos are denoted as (2D+3D) and the ones that used a fine-tuning strategy are marked with (FT). Note that our approach required no information other than the positions of the point clouds; this property allowed for a much simpler sampling process in actual scenarios, whereas the illumination/rotation variants have been "compressed" in our representations. The recognition accuracy of our approach was also competitive with a rank-1 recognition rate of 99.84%.

(2) Results on the Bosphorus dataset: The Bosphorus dataset [38] has 4666 scans collected from 105 subjects, with very rich variants in expression, systematic variations in poses, and different types of occlusions.

Table 2. Rank-1 recognition rate (RR1) under FAR = 1×10^{-3} on FRGC V2.0 dataset.

Method	RR1	VR
Mian et al. [34] (2008)	96.10%	98.60%
Al-Osaimi. [41] (2016)	96.49%	90.00%
Ouamane et al. [42] (2017)	–	96.65%
Ouamane et al. [42] (2017) [2D+3D]	–	98.32%
Gilani and Miancite [3] (2018)	97.06%	–
Gilani and Mian [3] (2018) (FT)	99.88%	–
Cai et al. [1] (2019) (FT)	100.00%	100.00%
Yu et al. [2](2022)	98.85%	96.75%
Ours	99.84%	99.39%

To be specific, almost every subject had varying scans, with 34 expressions; 13 yaw, pitch, and cross rotations; and 4 occlusions (hand, hair, eyeglasses); this dataset was found to be the most convincing experimental context for demonstrating our method's capacity. We did not exclude any hard samples from this dataset and obtained a 99.90% rank-1 accuracy. The comparison with other methods is illustrated in Table 3.

Table 3. Rank-1 recognition rate (RR1) under FAR = 1×10^{-3} on Bosphorus dataset.

Method	RR1	VR
Mian et al. [34] (2008)	96.40%	–
Al-Osaimi [41] (2016)	92.41%	93.50%
Lei et al. [31] (2016)	98.90%	–
Ouamane et al. [42] (2017) [2D+3D]	–	96.17%
Gilani and Miancite [3] (2018)	96.18%	–
Gilani and Miancite [3] (2018) (FT)	100%	–
Cai et al. [1] (2019) (FT)	99.75%	98.39%
Yu et al. [2] (2022)	99.33%	97.70%
Ours	99.90%	99.55%

5. Discussion

In general, our method provides an approach to defining a representation that is invariant to isometric transformations up to an induced small scale; additionally, our method enables one to train parameters from a limited number of observations. However, alternative deformations may exist in some inherent behaviors, e.g., aging/expression, which is not isometric. For those kinds of pattern recognition tasks, modifications should be applied to the original structure to capture discriminative features on purpose. The flexibility of this framework has the potential to extend and spread to wider fields, whereas mining geometrical data can play a role in more transparent methodologies.

6. Conclusions

This work shows that mere spatial coordination in point cloud faces is sufficient to improve the performance of 3D face recognition beyond the accuracy of other current methods. In addition, the strategy of applying fast-converging sparse dictionary deep learning to select the related features while reducing intraclass variances has created the potential to develop into applications in real time and unbounded real scenarios. Our future research interest is in improving the treatment of point cloud human faces as stochastic processes; we will examine the potential of its application on larger-scale recognition tasks.

Author Contributions: Conceptualization, Y.H. and P.C.; methodology, Y.H.; software, Y.H.; validation, Y.H. and S.Y.; formal analysis, Y.H.; investigation, P.C.; resources, J.Z.; data curation, S.Y.; writing—original draft preparation, Y.H.; writing—review and editing, Y.H.; visualization, Y.H.; supervision, J.Z.; project administration, P.C.; funding acquisition, P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Sichuan Science and Technology Program (2022YFG0261) and (2021YJ0079).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cai, Y.; Lei, Y.; Yang, M.; You, Z.; Shan, S. A fast and robust 3D face recognition approach based on deeply learned face representation. *Neurocomputing* **2019**, *363*, 375–397.
2. Yu, Y.; Da, F.; Zhang, Z. Few-data guided learning upon end-to-end point cloud network for 3D face recognition. *Multimed. Tools Appl.* **2022**, *81*, 12795–12814.
3. Gilani, S.Z.; Mian, A. Learning from millions of 3D scans for large-scale 3D face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 14–16 November 2018; pp. 1896–1905.
4. Yang, X.; Huang, D.; Wang, Y.; Chen, L. Automatic 3d facial expression recognition using geometric scattering representation. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; IEEE: Piscataway, NJ, USA, 2015; Volume 1, pp. 1–6.

5. Julesz, B. Visual pattern discrimination. *IRE Trans. Inf. Theory* **1962**, *8*, 84–92.
6. Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature* **1981**, *290*, 91–97.
7. Leung, T.; Malik, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vis.* **2001**, *43*, 29–44.
8. Eickenberg, M.; Exarchakis, G.; Hirn, M.; Mallat, S.; Thiry, L. Solid harmonic wavelet scattering for predictions of molecule properties. *J. Chem. Phys.* **2018**, *148*, 241732.
9. Zarka, J.; Thiry, L.; Angles, T.; Mallat, S. Deep network classification by scattering and homotopy dictionary learning. *arXiv* **2019**, arXiv:1910.03561.
10. Eldar, Y.; Lindenbaum, M.; Porat, M.; Zeevi, Y.Y. The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process.* **1997**, *6*, 1305–1315.
11. Bronstein, M.M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18–42.
12. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3D shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1912–1920.
13. Hermosilla, P.; Ritschel, T.; Vázquez, P.P.; Vinacua, À.; Ropinski, T. Monte carlo convolution for learning on non-uniformly sampled point clouds. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–12.
14. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12.
15. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4 December 2017; Volume 30, pp. 5105–5114.
16. Lüthi, M.; Gerig, T.; Jud, C.; Vetter, T. Gaussian process morphable models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1860–1873. [PubMed]
17. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
18. Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.H.; Kautz, J. Splatnet: Sparse lattice networks for point cloud processing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2530–2539.
19. Mallat, S. Group invariant scattering. *Commun. Pure Appl. Math.* **2012**, *65*, 1331–1398.
20. Bruna, J. Scattering Representations for Recognition. Ph.D. Thesis, Ecole Polytechnique X, (CMAP) Center Mathematics Appliquées, Paris, France, 2013.
21. Oyallon, E.; Mallat, S. Deep roto-translation scattering for object classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2865–2873.
22. De, S.; Bartók, A.P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769. [PubMed]
23. Gao, F.; Wolf, G.; Hirn, M. Geometric scattering for graph data analysis. In Proceedings of the International Conference on Machine Learning, Long Beach, California, USA, 9–15 June 2019; PMLR: New York, NY, USA, 2019; pp. 2122–2131.
24. Zou, D.; Lerman, G. Graph convolutional neural networks via scattering. *Appl. Comput. Harmon. Anal.* **2020**, *49*, 1046–1074.
25. Mairal, J.; Ponce, J.; Sapiro, G.; Zisserman, A.; Bach, F. Supervised dictionary learning. In Proceedings of the 21st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 8–10 December 2008; Volume 21.
26. Patil, H.; Kothari, A.; Bhurchandi, K. 3-D face recognition: Features, databases, algorithms and challenges. *Artif. Intell. Rev.* **2015**, *44*, 393–441.
27. Besl, P.J.; McKay, N.D. Method for registration of 3-D shapes. In Proceedings of the Sensor Fusion IV: Control Paradigms and Data Structures, Boston, MA, USA, 12–15 November 1991; SPIE: Bellingham, WA, USA, 1992; Volume 1611, pp. 586–606.
28. Yue, X.; Biederman, I.; Mangini, M.C.; von der Malsburg, C.; Amir, O. Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vis. Res.* **2012**, *55*, 41–46. [CrossRef]
29. Cai, L.; Da, F. Estimating inter-personal deformation with multi-scale modelling between expression for three-dimensional face recognition. *IET Comput. Vis.* **2012**, *6*, 468–479.
30. Al-Osaimi, F.; Bennamoun, M.; Mian, A. An expression deformation approach to non-rigid 3D face recognition. *Int. J. Comput. Vis.* **2009**, *81*, 302–316. [CrossRef]
31. Lei, Y.; Guo, Y.; Hayat, M.; Bennamoun, M.; Zhou, X. A two-phase weighted collaborative representation for 3D partial face recognition with single sample. *Pattern Recognit.* **2016**, *52*, 218–237. [CrossRef]
32. Ocegueda, O.; Shah, S.K.; Kakadiaris, I.A. Which parts of the face give out your identity? In Proceedings of the CVPR, Barcelona, Spain, 6–13 November 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 641–648.
33. Basri, R.; Jacobs, D.W. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 218–233.
34. Andreux, M.; Angles, T.; Exarchakis, G.; Leonarduzzi, R.; Rochette, G.; Thiry, L.; Zarka, J.; Mallat, S.; Andén, J.; Belilovsky, E.; et al. Kymatio: Scattering Transforms in Python. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
35. Mallat, S.G.; Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **1993**, *41*, 3397–3415. [CrossRef]

36. Mairal, J.; Bach, F.; Ponce, J. Task-driven dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 791–804.
37. Combettes, P.L.; Pesquet, J.C. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*; Springer: New York, NY, USA, 2011; pp. 185–212.
38. Savran, A.; Alyüz, N.; Dibeklioğlu, H.; Çeliktutan, O.; Gökberk, B.; Sankur, B.; Akarun, L. Bosphorus database for 3D face analysis. In Proceedings of the European Workshop on Biometrics and Identity Management, Brandenburg, Germany, 8–10 March 2008; Springer: New York, NY, USA, 2008; pp. 47–56.
39. Phillips, P.J.; Flynn, P.J.; Scruggs, T.; Bowyer, K.W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W. Overview of the face recognition grand challenge. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 947–954.
40. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
41. Al-Osaimi, F.R. A novel multi-purpose matching representation of local 3D surfaces: A rotationally invariant, efficient, and highly discriminative approach with an adjustable sensitivity. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015; Volume 25, pp. 658–672.
42. Ouamane, A.; Chouchane, A.; Boutellaa, E.; Belahcene, M.; Bourennane, S.; Hadid, A. Efficient tensor-based 2D + 3D face verification. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2751–2762.