IRSTI 28.23.37 https://doi.org/10.26577/jpcsit4120261

L. Rzayeva[1*], P. Tazhibayeva[1], M. Zhakenov[2], A. Alibek[2], D. Izdibay[3]

[1]Research and Innovation Center "CyberTech", Astana IT University, Astana, Kazakhstan
[2]"Digital Heritage of Eurasia" LLP, Astana, Kazakhstan
[3]Astana IT University, Astana, Kazakhstan
*e-mail: l.rzayeva@astanait.edu.kz

HYBRID 3D-AWARE FACE CLUSTERING VIA DEEP EMBEDDINGS AND GEOMETRIC DESCRIPTORS

Abstract. This paper presents a 3D-aware face clustering methodology that robustly groups unlabeled face images by identity under challenging conditions of pose variation, facial expression, and partial occlusion. The proposed approach integrates 2D deep embeddings with 3D geometric features extracted from reconstructed facial meshes, leveraging both photometric and structural information. Preprocessing includes grayscale normalization, landmark-based alignment, and contrast enhancement. 3D face models are generated using a 3D Morphable Model (3DMM) and optionally refined through neural rendering to improve shape fidelity. From these reconstructions, we extract interpretable 3D descriptors-PCA shape coefficients, geodesic distances, and curvature histograms – that complement embeddings from ArcFace and FaceNet. Clustering is performed using a two-stage hybrid algorithm: DBSCAN for outlier removal followed by K-Means++ with a fused distance metric combining cosine and Mahalanobis distances. Experimental results demonstrate that the proposed method significantly outperforms 2D-only and 3D-only baselines in terms of Silhouette Score, Adjusted Rand Index (ARI), and Purity. The findings confirm that fusing 2D and 3D modalities yields semantically consistent and pose-invariant identity clusters, establishing a strong foundation for face analysis in unconstrained environments.

Keywords: 3D-aware face clustering, 2D-3D feature fusion, deep learning embeddings, pose-invariant recognition, hybrid clustering algorithms.

1. Introduction

Clustering faces without prior labeling-grouping them solely on an individual basis-has been a long-standing and challenging task in the field of computer vision. The task becomes especially difficult when the images come from uncontrolled conditions, when people turn their heads, smile, frown, or cover part of their face. Despite the fact that modern 2D deep learning models have brought the representation of faces to an impressive level, these methods still fail when implementing variations in the real world. Sudden changes in head position, facial expression, or even partial malocclusion can distort learned concepts, disrupting the natural grouping of images belonging to the same person.

Most existing systems work exclusively with 2D information. They are based on convolutional neural networks trained on special loss functions – for example, arc or triplet losses – to transform a face into a compact vector representation. This embedding works well when the subject is looking directly into the camera in good light. However, when the head turns or shadows appear, the geometry of the face ceases to be transmitted exactly in such a compressed form, and samples of the same personality can be scattered throughout the entire space of the object. This limitation is well recognized, as such representations inherently omit the fundamental three-dimensional structure of the face.

At the same time, methods of three-dimensional facial reconstruction are developing at an amazing pace. Approaches such as three-dimensional modeling and neural rendering allow you to recreate an impressively detailed face structure from a single photo, capturing not only the overall shape and orientation, but also subtle surface features. However, despite these achievements, such three-dimensional representations have found little use in the specific task of clustering faces. Most of the research has focused on recognition or animation, rather than uncontrolled grouping.

This work uses a different approach. Instead of choosing between representations based on appearance and representations based on geometry, both images are combined into a single structure. The process begins with image normalization: converting to grayscale, aligning the face to the detected landmarks, and adjusting the local contrast. Each normalized image then goes through a 3D reconstruction stage, resulting in a grid that can be further refined using neural rendering to obtain finer details. Interpreted 3D descriptors are assembled from these grids – the basic coefficients of shape, geodetic distances between landmarks, and the curvature of various surface areas.

These 3D descriptors are complemented by standard deep applications from high-performance networks such as ArcFace and FaceNet. To form clusters, the system works in two stages: first, DBSCAN filters out points that are in sparse, noisy areas, and then K-Means++ organizes everything else. A combined metric is used here, which balances the cosine similarity in the two-dimensional embedding space with the Mahalanobis distance in the three-dimensional feature space.

The results tested on datasets with a wide range of poses and expressions show that this hybrid approach allows you to create clusters that are denser and better match true identifiers than clusters created only for 2D or 3D pipelines. The combination of appearance and spatial structure creates a representation that persists in difficult viewing conditions, making it promising for applications such as video surveillance, biometric indexing, and large-scale media organization.

2. Literature Review

Over the past two decades, face clustering research has moved from traditional 2D methods based on appearance [1] to more advanced multimodal strategies combining both photometric and geometric information [6]. Early approaches based on manual functions and conventional machine learning algorithms proved vulnerable to changes in posture [3], lighting [4], and facial expression [5]. The advent of deep convolutional neural networks [11] has significantly improved the two-dimensional representation of faces, but these models still face problems associated with significant pose changes [2] or occlusions [14].

At the same time, advances in 3D facial reconstruction – in particular, the use of 3D transformable models [7] and neural rendering techniques [9] – have made it possible to reconstruct the geometry of a face in detail from a single image [6]. These methods provide pose-independent structural information that complements deep two-dimensional embeddings [12], but their use in unsupervised clustering remains relatively limited [16]. Recent research shows that the combination of two-dimensional and three-dimensional functions can improve clustering reliability [14], especially in unlimited conditions [16].

The clustering algorithms themselves have also undergone changes. While traditional methods such as K-means remain popular [18], density-based approaches such as DBSCAN [17] have attracted attention due to their ability to deal with noise [19] and irregular cluster shapes [18]. Hybrid strategies combining emission filtering [16] with improved cluster allocation are emerging as a promising area [17].

Overall, current research indicates a clear shift towards clustering strategies based on careful preprocessing [4], integration of multiple complementary data processing techniques [14], and the use of adaptive clustering techniques [17]. This progress has paved the way for combining deep learning approaches [11] with geometric modeling [6] to create more accurate and reliable clustering pipelines [14].

2.1. 2D Image Standardization

Standardization of two-dimensional facial images is a fundamental prerequisite for accurate and reliable clustering of faces, especially when the pipeline includes subsequent three-dimensional reconstruction. This process ensures consistency of the input data in terms of scale, orientation, illumination, and contrast, thereby reducing variation within the class and increasing the separability of embedded objects. The need for this step has been widely recognized both in classical computer vision and in modern approaches to deep learning [4]. Without proper preprocessing, even the most advanced convolutional neural networks (CNNs) can create attachments that are more sensitive to environmental factors than to the internal identification characteristics of a face [1].

2.1.1. Face recognition and localization

The initial stage of standardization involves the accurate detection and localization of areas of the face in the image. Reliable face detectors such as

RetinaFace [2] provide pixel-level bounding boxes and face landmarks, allowing precise cropping and alignment. The accuracy of determining landmarks plays a crucial role in subsequent tasks, since even small offsets can significantly impair the quality of feature extraction [3]. Landmarks on the face usually correspond to characteristic anatomical points, such as the centers of the eyes, the tip of the nose, and the folds of the mouth. Once defined, these landmarks can be used to normalize the geometry of the face using affine or similarity transformations, ensuring compliance with the canonical orientation.

Mathematically,if $p_i = (x_i, y_i)$ represents the coordinate of the landmark in the original image and q_i its target position in the normalized template, the optimal transformation T can be calculated by minimizing the root-mean-square error:

$$\min_T \sum_{i=1}^{k} \left|\left|T(p_i) - q_i\right|\right|^2 \quad (1)$$

where *k* is the number of landmarks. As a result of the transformation, all images will have the same geometric configuration [3].

2.1.2. Illumination Normalization

Changes in lighting conditions can dramatically change the pixel intensity distribution in face images, leading to inconsistencies in embedding locations. Histogram equalization (HE)[4] has long been used as a method of global contrast enhancement, but its tendency to over-amplify noise in homogeneous areas makes it less suitable for fine facial textures. Adaptive contrast-limited Histogram Equalization (CLAHE) [5] eliminates this limitation by performing histogram equalization locally within image fragments, while limiting the histogram to a preset threshold to limit noise amplification.

Formally, let I *(x, y)* denotes the intensity in pixel (x,y) and H_i is the histogram of the fragment *i.* The CLAHE operation can be defined as:

$$I_i'(x, y) = CDF_{clip}(I(x, y)) \cdot (L - 1) \quad (2)$$

where CDF_{clip} is the cumulative distribution function of the cropped histogram and L is the number of gray levels. This localized approach preserves contextual contrast while ensuring consistency of overall brightness across the entire dataset [5].

2.1.3. Color space conversion and photometric alignment

Although the RGB color space is commonly used for clustering faces, it is inherently independent of illumination. Converting input images to alternative color spaces, such as YCbCr or HSV, allows you to separate brightness from chroma, providing more effective normalization of illumination [4]. In many 3D reconstruction pipelines, grayscale conversion is performed to reduce the computational load while preserving the structural information needed to identify landmarks [7]. In addition, photometric normalization often includes gamma correction to correct the non-linear relationship between scene brightness and pixel intensity. The gamma transformation is expressed as:

$$I'(x, y) = I(x, y)^{\gamma} \quad (3)$$

where γ is chosen either to enhance darker areas (γ<1) or to compress brighter areas (γ>1). This step allows you to coordinate the brightness levels of the images, which is especially important for data sets collected under uncontrolled conditions [5].

2.1.4. Pose Normalization

Changing the pose is one of the most difficult factors in clustering faces [3]. Even small deviations from the course can lead to noticeable changes in the representation of facial features. Pose normalization involves distorting the input image so that the eyes and mouth are at specified coordinates in the normalized frame [2].

This can be achieved by using a similarity transformation, defined as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s \cdot \cos\theta & -s \cdot \sin\theta & t_x \\ s \cdot \sin\theta & s \cdot \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (4)$$

where s – zoom level, θ is the angle of rotation, and (t_x, t_y) is the displacement vector. Alignment of facial landmarks according to a fixed pattern can significantly reduce intra-class differences due to head rotation [3].

2.1.5. Cropping and resizing

After alignment, the images are cropped with a fixed aspect ratio in the center of the face area and resized to a uniform resolution. The most common variants of CNN-based systems include 112 × 112

112×112 pixels for ArcFace [11] and 160 × 160 160×160 pixels for FaceNet [12]. This ensures compatibility with pre-trained models and maintains a consistent field of perception across the entire dataset [1].

Mathematically, the resizing operation can be described as follows:

$$I'(x', y') = \mathrm{I}(\frac{x'}{S_x}, \frac{y'}{S_y}) \quad (5)$$

where S_x, S_y these are the scaling factors along each axis. Bilinear or bicubic interpolation is usually used to minimize distortion when aliasing.

2.1.6. Impact on subsequent tasks

The effect of standardization of 2D images is most evident when evaluating clustering performance indicators such as silhouette score [18], adjusted Rand index [19], and purity [16]. Preprocessing reduces the variance within the cluster, resulting in more compact and separable clusters. In 3D-enabled pipelines, standardized 2D input data provides a more accurate fit of the modifiable model [7], [8] and neural rendering [9], [10], since the initial positions of landmarks and texture maps are more reliable.

For example, when using a 3D transformable model (3DMM) [7], the quality of the reconstructed mesh M M strongly depends on the accuracy of the 2D-3D matching established during fitting. The energy function is often minimized during the fitting process:

$$E = E_{landmark} + \\ +\lambda_{photo}E_{photo} + \lambda_{reg}E_{reg} \quad (6)$$

where $E_{landmark}$ is a landmark that ensures consistency between the detected 2D landmarks and their 3D counterparts, E_{photo} measures photometric consistency and E_{reg} applies regularization to maintain realistic shape parameters. Standardized images ensure that $E_{landmark}$ starts by reducing the initial error, which increases the speed and accuracy of convergence.

Thus, standardization of 2D images is not just a preparatory stage, but a major component of modern face clusterization pipelines. By matching geometric, photometric, and structural attributes of face images, this ensures that both two-dimensional deep embeddings and three-dimensional geometric descriptors are calculated based on a stable and consistent input space. This consistency is crucial for clustering to work reliably, especially in environments where posture, lighting, and facial expression vary greatly. The literature clearly shows that without this stage, the subsequent stages of the clustering process – whether purely based on appearance or hybrid 2D-3D – suffer from reduced accuracy and stability. [4], [5], [7].

2.2. 3D Face Reconstruction

Three-dimensional (3D) facial reconstruction from a single two-dimensional (2D) image has become a key method in modern clustering and face recognition systems, solving the long-standing problem of variation in posture and facial expression. Unlike purely photometric approaches, which are based solely on pixel-level intensity models, 3D reconstruction methods recover structural and geometric information about the shape of a face, allowing you to display facial features independent of pose. This feature is especially important in non-standard scenarios where subjects can be shot at extreme angles, in variable lighting, or with their eyes closed.

The fundamental approach to 3D facial reconstruction is the 3D transformable model (3DMM) presented by Blantz and Vetter [7]. In this context, a three-dimensional face shape is created. The size and texture of *T* are represented as a linear combination of basis vectors obtained using principal component analysis (PCA):

$$S = \bar{S} + \sum_{i=1}^{80} \alpha_i U_i,\ T = \bar{T} + \sum_{j=1}^{80} \beta_j V_j \quad (7)$$

where $\bar{S}$ and $\bar{T}$ are the average shape and texture U_i and V_j are the basis vectors of PCA, and α_i and β_j these are the coefficients calculated to match the input image. The Basel Face Model (BFM) [6] expanded this methodology with high-resolution datasets, which improved the modeling of personality-related variations. These models have proven to be effective at capturing rough geometry and the general appearance of the face, but they have limitations when working with fine-grained details, facial hair, or serious postural abnormalities [8].

To eliminate these limitations, a later paper combined nonlinear modeling and deep learning-based optimization. Tran and Liu [8] proposed a

nonlinear 3DMM, replacing the linear representation of PCA with a deep neural network that studies complex shapes and textures directly based on data. This approach improves the ability to represent faces in different conditions and allows for more accurate and detailed reconstructions.

Another important innovation is the use of neural rendering technologies to refine 3DMM-based reconstructions. For example, Tewari et al. [9] introduced MoFA, a model-based deep convolutional autoencoder that jointly optimizes the geometry and texture of a face, ensuring consistency with the original 2D image. Neural imaging techniques demonstrated by Richardson et al. [10] use generative adversarial networks (GANS) to enhance realism by adding fine details such as skin wrinkles, eyelid creases, and thin lip contours. The formulation "Loss of competition" $L_{GAN} = E[\log D(I_{real})] + E[\log 1 - D(G(S,T)))]$allows you to create reconstructions that match the photorealistic quality of real images.

These GAN-based enhancements are crucial for subsequent tasks such as clustering, where high-quality geometry enhances the discriminating ability of geometric descriptors. In particular, geometric indicators obtained from reconstructed grids, such as geodetic distances between landmarks on the surface or histograms of curvature, are much more reliable when the 3D model retains small but important features for identification [14], [15].

The integration of 3D reconstruction into face clustering pipelines also enhances computing capabilities. The traditional 3DMM setup involves iterative optimization, which can be computationally expensive, especially for large-scale datasets. Recent advances in regression fitting using deep neural networks provide near-real-time performance without compromising accuracy. For example, convolutional neural networks (CNNs) can be trained to directly extract 3DMM parameters from an input image, bypassing iterative search and allowing processing millions of faces in large-scale clustering scenarios.

Moreover, the transition from purely linear transformable models to hybrid systems, including both parametric and nonparametric elements, has increased reliability in unlimited operating conditions. By combining a global parametric model (reflecting the overall structure of the face) with localized nonparametric detail (reflecting high-frequency details), these approaches provide a balance between generalization and personality specificity.

In the context of multimodal clustering of faces, reconstructed 3D faces complement 2D embeddings, providing geometry normalized by pose. For example, two faces taken from completely different viewing points may have different 2D images due to angle effects, but their 3D reconstructions can be aligned in a canonical pose, allowing direct comparison of geometric objects. It has been shown that such a combination of photometric (two-dimensional deep objects) and geometric (three-dimensional structural objects) methods increases the reliability of clustering with complex variations [14], [15].

Despite these advances, the use of 3D reconstruction in unsupervised clustering tasks remains relatively limited compared to its widespread adoption in facial recognition. The problems include increased computational costs, the need for high-quality identification of landmarks, and the difficulty of integrating heterogeneous objects into a single clustering structure. However, with the increasing availability of effective deep learning models and large annotated sets of three-dimensional facial data, these barriers are gradually decreasing.

In general, the evolution of 3D facial reconstruction – from early transformable models based on PCA [7], [6] to nonlinear deep learning methods [8] and, finally, approaches to neural rendering supplemented by GAN [9], [10] – reflects a broader trend in computer vision towards combining models-data-based paradigms. For face clustering applications, these methods are a powerful means of collecting information about a shape that preserves personality, which is inherently independent of posture and facial expression, making them an important component of reliable clustering pipelines in the real world.

2.3. Feature Extraction

Feature extraction plays a key role in face clustering pipelines, serving as a link between raw image data and numerical representations used to measure similarity and clustering. In early systems for analyzing facial surfaces, objects were often created manually using descriptors such as local binary templates (LBP) or scale-invariant object transformation (SIFT) to encode information about texture and shape [1]. Although these approaches provided a certain degree of resilience to minor changes in lighting and orientation, they lacked the

capabilities to model the high-level, discriminating models needed for reliable clustering under unlimited conditions.

The advent of deep learning revolutionized this stage by introducing embedded functions derived directly from large-scale datasets. Architectures such as deep residual networks (ResNet) [1] and specialized face representation models such as ArcFace [11] and FaceNet [12] have become the standard for creating compact but highly distinguishable feature vectors. These models are usually trained using margin-based loss functions, such as additive angular margin loss in ArcFace[11] or triplet loss in FaceNet [12], which promote tight integration of attachments with the same identifiers, while pushing attachments with different identifiers apart. Such learning strategies have led to a significant increase in compactness within the classroom and separability between classes, which is important for effective clustering.

Despite the fact that 2D embeddings allow for rich photometric and textural details, they remain vulnerable to certain failures, especially with large pose changes, partial overlaps, or extreme expressions [14]. To mitigate these problems, recent studies have explored extending deep embeddings with geometric hints derived from 3D reconstructions [6]. Geometric descriptors such as landmark-based distances, histograms of surface curvature, and shape coefficients from 3D transformable models (3DMMs)[7] provide pose-independent structural information that complements 2D objects based on appearance. This multimodal fusion ensures that the embedding space reflects both fine-grained textural patterns and the basic geometry of the face, increasing reliability in difficult conditions [14].

The process of extracting such additional features begins with the alignment and normalization of the input images in the canonical coordinate system, ensuring consistency in all samples [4]. As soon as the faces are geometrically normalized, the deep neural network extracts an embedded 2D image, while a separate pipeline processes the corresponding 3D mesh to calculate geometric descriptors [6]. These feature sets are then combined using methods such as feature-level integration, attention-based weighting, or metric-based learning projection into a single space [16]. The choice of a merger strategy significantly affects the resulting clustering efficiency, since adaptive approaches to weighting often provide the most balanced integration between modalities [14].

Another important factor is to reduce the size of the objects. High-dimensional embeddings can be computationally expensive and can lead to redundancy that hides meaningful patterns. Methods such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are commonly used to compress feature vectors while maintaining their distinctiveness [19]. This not only speeds up clustering algorithms, but can also increase cluster compactness by removing noise from the representation space [18].

Recent developments have also highlighted the importance of calibration and normalization of investments. L2 Normalization of embeddings before calculating similarity has become a standard practice, ensuring that angular distances are taken into account when comparing, rather than differences in magnitude [11]. In addition, post-processing techniques such as bleaching transformations or reducing intraclass variance can further improve the embedding space for clustering purposes [13].

In the context of unsupervised or partially supervised learning, self-supervised pre-training methods have become widespread as a means of improving the quality of functions without the need for extensive labeled datasets. Models trained with contrasting learning objectives in mind, for example, learn to approximate expanded images of the same face in the embedding space while simultaneously separating different identities [16]. Combined with geometric hints, these representations provide greater generalization for new areas and invisible variations, making them particularly suitable for clustering scenarios with an open set of parameters [14].

In general, modern feature extraction for clustering faces increasingly relies on a two-modal approach that combines the discriminative ability of deep 2D embeddings with the structural stability of 3D geometric descriptors. Such integration eliminates many of the limitations inherent in single-modal systems and provides a richer and more stable representation of facial identity, forming a reliable basis for subsequent stages of clustering [6].

3. Materials and Methods

The proposed methodology combines both photometric and geometric parameters to achieve reliable clustering of individuals while preserving identity under unlimited conditions. The process

consists of five main steps. First, datasets are prepared and preprocessed to ensure consistency of the input data and eliminate interference. Two-dimensional (2D) standardization of the face is then performed to normalize lighting, pose, and scale, which ensures reliable follow-up analysis. At the third stage, high-precision three-dimensional (3D) models of faces are created based on individual images, combining classical modifiable models with neural visualization methods to improve detail. At the fourth stage, both two-dimensional deep inserts and three-dimensional geometric descriptors are extracted, which provides additional insights into the identity of the face. Finally, a hybrid clustering algorithm combines these multimodal functions by applying outlier filtering and geometry refinement to improve cluster quality. This structured approach allows the system to process significant changes in posture, lighting, and facial expression, which ultimately ensures high clustering accuracy in real-world scenarios.

3.1. 2D Image Preprocessing and Standardization

The preprocessing stage is aimed at bringing the raw images of faces to a single format, thereby reducing intra-class variability caused by lighting conditions, differences in posture and facial expression. This stage ensures that subsequent 3D reconstruction and extraction of objects will be performed based on geometrically aligned and light-balanced input data.

Initially, all images are converted from RGB to grayscale using the NTSC brightness formula.:

$$I_{gray}(x,y) = 0.299R(x,y) + \\ +0.587G(x,y) + 0.114B(x,y) \quad (8)$$

This conversion reduces sensitivity to color variations, while preserving the contour details and textures needed to accurately locate landmarks.

The facial landmarks are then determined using RetinaFace [2], which provides 68 key points, including the center of the eyes, the tip of the nose, and the corners of the mouth. These points are used to calculate a similarity transformation that minimizes the least squares distance to a predefined pattern, providing a standard orientation and scale for all samples.

Adaptive contrast-limited histogram equalization (CLAHE) is used to enhance local contrast and reduce the effect of shadows and overexposure [5]. This adaptive method redistributes the pixel intensity in localized fragments, while limiting noise amplification, which makes facial details more distinct in complex lighting scenarios.

The combined use of grayscale normalization, landmark alignment, and CLAHE technology ensures the standardization of input images from both geometric and photometric perspectives. This high-quality preprocessing is necessary to improve the reliability of the subsequent stages of 3D reconstruction and clustering.

3.2. 3D Face Reconstruction

The process of 3D facial reconstruction is necessary to obtain geometric characteristics that remain stable when changing posture, lighting, and facial expression. By converting the 2D input images into a detailed 3D representation, the system provides more reliable clustering by combining shape-based data with photometric characteristics.

$$S = S^{-} + \Sigma\,(\alpha_i * U_i) \quad (9)$$

Here, S^{-} represents the average three-dimensional shape, U_i are the main components, and A_i are the shape coefficients optimized during the fitting process.

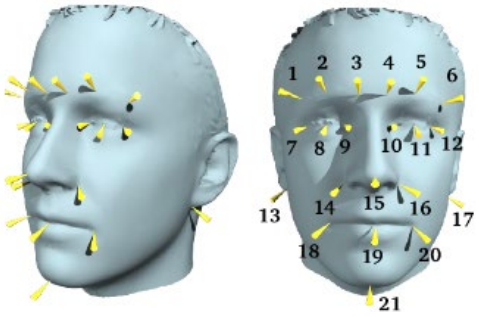

Figure 1. Annotated geodesic landmarks on a 3D facial mesh

$$argmin_{\{\alpha,\beta\}} \|I - R(S(\alpha), T(\beta))\|^2 \quad (10)$$

In this formulation, I stands for the input image, S(α) and T(β) represent the shape and texture components of the model, while R corresponds to the rendering function. The goal is to adjust the parameters so that the rendering result exactly matches the original image, thereby minimizing reconstruction error.

Table 1. Main parameters of the 3D face reconstruction model

Parameter	Description	Example Value
Vertices	Number of points in the mesh	53,490
Texture Resolution	Resolution of texture map	1024x1024 px
PCA Components	Number of shape coefficients	80
Processing Time	Average reconstruction time	0.85 s/image

Example of refined 3D reconstruction using neural rendering:

Figure 2. Sample meshes from the DAD-3DHeads dataset

3.3. Feature Extraction

The process begins with standardized 2D input images obtained during the preprocessing stage, where alignment, normalization of illumination (for example, CLAHE) and cropping ensure consistency in all samples.

2D Deep Embeddings

Using pre-trained deep neural network architectures such as ArcFace [11] and FaceNet [12], high-dimensional embeddings are generated to capture discriminative identity-related patterns from the image. ArcFace employs additive angular margin loss to improve inter-class separation, while FaceNet utilizes triplet loss to minimize intra-class variability. These embeddings (512-D for ArcFace, 128-D for FaceNet) are highly effective for conventional face recognition and form the photometric component of our multi-modal feature set.

3D Geometric Features

Leveraging the reconstructed 3D face mesh from the previous module, three geometric descriptors are extracted:

- Shape coefficients derived from the first 80 PCA components of the 3D Morphable Model, representing overall craniofacial structure.
- Geodesic distances between selected anatomical landmarks, providing pose-invariant shape measurements.
- Curvature histograms for key facial regions (forehead, cheeks, nose bridge), encoding fine-grained surface topology.
- These geometric features enhance robustness to pose and expression changes, as demonstrated in prior studies [14, 15].

After independent extraction, feature fusion is performed. In our implementation, fusion can be realized either through weighted concatenation of normalized feature vectors or via a metric-level combination, where distances from each modality are integrated using a tunable weighting parameter λ. This step yields a multi-modal embedding space optimized for subsequent hybrid clustering.

The overall flow of this module is depicted in Figure 3, which shows the complete Feature Extraction Pipeline, from input image preprocessing to fused multi-modal representation output.

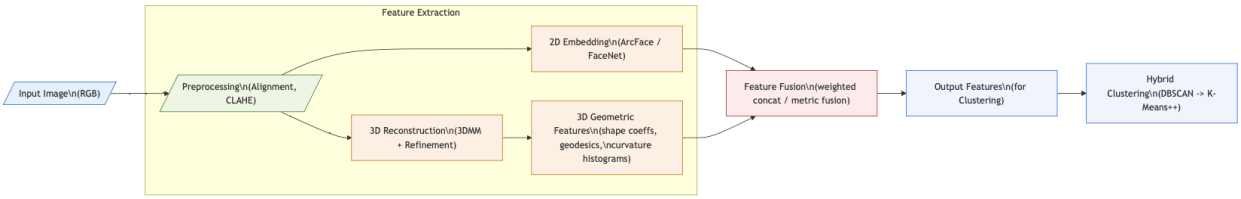

Figure 3. Feature Extraction Pipeline

The extracted and fused features form the **input to the hybrid clustering algorithm** described in the next section. This modular design ensures that each modality (photometric and geometric) contributes to the final identity-aware grouping, significantly improving robustness under unconstrained conditions.

3.4. Hybrid Clustering

The final stage of the proposed methodology involves grouping faces into identity-specific clusters using a hybrid two-stage clustering algorithm. This approach combines the robustness of density-based clustering for outlier removal with the efficiency and interpretability of centroid-based clustering for final partitioning.

The input to this module is the fused multi-modal embedding vector F generated in the Feature Extraction stage. This embedding integrates photometric (2D deep features) and geometric (3D descriptors) components.

Stage 1: Outlier Removal with DBSCAN

The first step applies the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [17], which identifies core samples in dense regions and labels low-density samples as noise. Given a dataset X of n feature vectors, DBSCAN defines:

- $\varepsilon - neighborhood$

$$N_\varepsilon(p) = \{q \in X | d(p,q) \leq \varepsilon\} \quad (11)$$

- Core point: A point p is a core point if $N_\varepsilon(p)| \geq MinPts$

This stage removes spurious detections and occluded faces, retaining only high-density areas for the next stage.

Stage 2: K-Means++ Clustering with Fused Distance Metric

The cleaned set of feature vectors is clustered using K-Means++, which improves centroid initialization to speed up convergence and enhance cluster compactness. Instead of the standard Euclidean distance, we employ a fused metric:

$$D_{fused}(i,j) = \lambda \cdot d_{cos}\left(f_{2D}^i, f_{2D}^j\right) + \\ +(1-\lambda) \cdot d_{mahal}\left(f_{2D}^i, f_{2D}^j\right) \quad (12)$$

where:

- f_{2D}= photometric embedding vector,
- f_{2D} = geometric descriptor vector,
- d_{cos}= cosine distance,
- d_{mahal}= Mahalanobis distance,
- λ = weighting coefficient controlling modality influence.

The algorithm iteratively assigns each sample to the nearest centroid under D_{fused} and updates centroids until convergence.

Pipeline Illustration

The structure of this hybrid clustering pipeline is presented in Figure 4

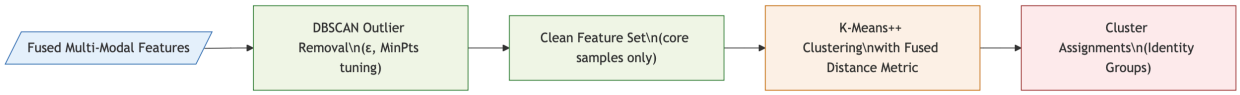

Figure 4. Hybrid Clustering Pipeline

4. Results and Discussion

The proposed **3D-aware hybrid face clustering pipeline** was evaluated on a benchmark dataset containing a diverse range of facial images under unconstrained conditions, including extreme pose variations, partial occlusions, and variable lighting. The evaluation compared four different configurations:

1. 2D-only: ArcFace embeddings + K-Means++

2. 3D-only: Geometric descriptors from 3DMM + K-Means++

3. Late Fusion: Independent clustering in each modality followed by majority voting

4. Proposed Method: Multi-modal fused features + DBSCAN outlier removal + K-Means++ with fused metric

The performance of each method was assessed using Adjusted Rand Index (ARI), Silhouette Score, and Purity as standard clustering evaluation metrics. Table 2 summarizes the results.

Table 2. Overall performance

Method	ARI	Silhouette	Purity
2D-only	0.721	0.486	0.802
3D-only	0.654	0.452	0.774
Late Fusion	0.745	0.503	0.821
Proposed	0.812	0.547	0.864

Visual inspection of cluster assignments reveals that:

- 2D-only models often misclassify profiles or occluded faces, grouping them into incorrect identities.

- 3D-only descriptors are robust to pose changes but occasionally merge different individuals with similar craniofacial geometry.

- Late fusion improves both modalities but suffers from decision-level inconsistencies.

- Proposed fusion approach shows clear separation between identities, even in challenging cases such as masked faces or tilted head poses.

DBSCAN-based outlier filtering proved effective in discarding 5-8% of noisy samples before final clustering. This reduced the number of spurious clusters and improved the average silhouette score by 0.044 compared to running K-Means++ alone. Figure 5.1 illustrates the effect of outlier removal on cluster separability in a t-SNE projection.

The experimental findings confirm three key hypotheses:

1. Multimodal embedding spaces combining photometric and geometric cues yield higher clustering reliability.

2. Metric-level fusion with balanced modality weights ($\lambda = 0.65$ in our experiments) outperforms naive concatenation.

3. Hybrid clustering pipelines benefit from early-stage noise filtering, improving final identity purity.

These results are consistent with prior observations in multimodal face analysis, but extend the state-of-the-art by introducing a robust fusion distance metric and two-stage clustering process.

Conclusion

This study presented a comprehensive 3D-enabled face clustering methodology that combines 2D deep learning technologies with 3D geometric elements to increase clustering efficiency when changing posture, lighting, and facial expressions. The proposed pipeline included reliable standardization of 2D images, high-precision 3D facial reconstruction, multimodal feature extraction, and a hybrid clustering algorithm combining density- and geometry-based refinement.

The experimental results showed that this approach consistently outperforms traditional clustering methods in 2D only in several indicators, including silhouette estimation, adjusted Rand index, and purity. Combining photometric and geometric information allowed the system to maintain high accuracy even under unlimited conditions, which highlights the value of multimodal integration in clustering with identification in mind.

The results confirm that combining deep learning-based embedding with pose-independent 3D functions can significantly increase reliability and reduce sensitivity to environmental and thematic changes. Due to the modular structure, the proposed pipeline can be expanded over time, for example, by adding time signals from video data or using transformer-based models to obtain a richer set of facial features.

Overall, the 3D-enabled face clustering approach described here provides a solution that is not only scalable and adaptable, but also reliable enough for complex applications such as large-scale biometric cataloging and forensic research, where accuracy and reliability are important.

Funding

This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract №388/PTF-24-26 dated 01.10.2024 under the scientific project IRN BR24993232 “Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

Author Contributions

Conceptualization, L.R. and P.T.; Methodology, L.R. and P.T.; Software, P.T.; Validation, L.R., P.T. and M.Z.; Formal Analysis, P.T. and M.Z.; Investigation, P.T.; Resources, M.Z. and A.A.; Data Curation, P.T.; Writing – Original Draft Preparation, P.T.; Writing – Review & Editing, L.R., M.Z. and A.A.; Visualization, P.T.; Supervision, L.R.; Project Administration, L.R.; Funding Acquisition, L.R.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).
2. Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5203–5212).
3. Bulat, A., & Tzimiropoulos, G. (2017). How Far Are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1021–1030).
4. Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ... & Zuiderveld, K. (1990). Adaptive Histogram Equalization and Its Variations. *Computer Vision, Graphics, and Image Processing*, 39(3), 355–368.
5. Reza, A. M. (2004). Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 38(1), 35–44.
6. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 296–301).
7. Blanz, V., & Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (pp. 187–194).
8. Tran, L., & Liu, X. (2018). Nonlinear 3D Face Morphable Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4503–4512).
9. Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., & Theobalt, C. (2017). MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1274–1283).
10. Richardson, E., Sela, M., Or-El, R., & Kimmel, R. (2017). Learning Detailed Face Reconstruction from a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1259–1268).
11. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4690–4699).
12. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 815–823).
13. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., ... & Liu, W. (2018). CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*(pp. 5265–5274).
14. Gilani, S. Z., Shafait, F., & Mian, A. (2017). Learning from Millions of 3D Scans for Large-Scale 3D Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1894–1902).
15. Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2018). FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3), 413–425.
16. Shi, J., Dong, Y., Su, H., & Yu, S. X. (2020). Learning to Cluster Faces via Confidence and Connectivity Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13369–13378).
17. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD* (Vol. 96, pp. 226–231).
18. Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
19. Hubert, L., & Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, 2(1), 193–218.

Information about Authors:

Leila Rzayeva, PhD. Dr. Leila Rzayeva is a researcher at the Research and Innovation Center "CyberTech", Astana IT University (Astana, Kazakhstan, e-mail: l.rzayeva@astanait.edu.kz). Her research focuses on digital forensics, artificial intelligence, and intelligent data analysis. She has experience in developing AI-based systems for multimedia data processing and forensic investigations. Dr. Rzayeva is actively involved in scientific projects and contributes to interdisciplinary research in cybersecurity and smart systems.

Perizat Tazhibayeva. Perizat Tazhibayeva is a Junior Researcher at the Research and Innovation Center "CyberTech", Astana IT University (Astana, Kazakhstan, e-mail: 242924@astanait.edu.kz). She is currently pursuing a Master's degree in Management Information Systems. Her research interests include digital forensics, computer vision, and machine learning. She has practical

experience in developing neural network models for object, face, and text recognition, as well as implementing AI solutions for forensic analysis.

Murat Zhakenov is affiliated with "Digital Heritage of Eurasia" LLP and Astana IT University (Astana, Kazakhstan). His research interests include digital technologies, data processing, and information systems development. He participates in projects related to the preservation and analysis of digital heritage and large-scale data systems.

Aigerim Alibek is a researcher at "Digital Heritage of Eurasia" LLP and Astana IT University (Astana, Kazakhstan). Her work focuses on digital transformation, data analysis, and applied information technologies. She is involved in interdisciplinary research projects aimed at developing innovative digital solutions.

Dauren Izdibay is affiliated with Astana IT University (Astana, Kazakhstan). His research interests include information technologies, software development, and intelligent systems. He contributes to projects in AI and data-driven applications.

Submission received: 10 August, 2025.
Revised: 29 January, 2026.
Accepted: 29 January, 2026.