Journal of Intelligent Systems and Internet of Things
Vol. 17, No. 01, PP. 325-341, 2025
DOI: https://doi.org/10.54216/JISIoT.170123

A Novel Approach to Face Recognition in Videos Based on a Single Reference Image

Mohammed Ahmed Talab[1], Mustafa A. Feath[2], Ahmed Hadi Ali AL-Jumaili[3,*], Mohammed A. Al-shibl[4], Ravie Chandren Muniyandi[5]

[1]Department of medical Physics, College of Applied Science, University of Fallujah, Anbar, 00964, Iraq

[2]Director of the Department of Studies and Planning, College of Medicine, University of Anbar, Anbar, 00964, Iraq

[3]College of Information Technology, University of Fallujah, Anbar, 00964, Iraq

[4]Director of Computer Center, University of Fallujah Anbar, 00964, Iraq

[5]College of Computing & Informatics (CCI), Universiti Tenaga Nasional (UNITEN), Putrajaya Campus, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

Emails: mmss_ah@uofallujah.edu.iq; Azeezmustafa89@uoanbar.edu.iq; ahmed_hadi@uofallujah.edu.iq; dr.alshibly@uofallujah.edu.iq; ravie.chandren@uniten.edu.my

Abstract

This paper introduces an advanced method for face recognition in video surveillance systems, leveraging only a single reference image per individual. The challenge of recognizing faces in video is addressed, considering issues like pose variations, occlusions, and lighting changes. The proposed approach utilizes 3D Morphable Models (3DMM) to generate a 3D face mesh from the reference image, facilitating robust face alignment and recognition across video frames. A Convolutional Neural Network based pipeline is employed for face detection, pose estimation, and extraction of invariant features, while an optimization framework refines landmark positions and depth maps for accurate 3D reconstruction. The system performs exceptionally well on the CASIA-WebFace Dataset, with 97.00% pAUC (20%) in surveillance mode and 98.69% in identification mode for frontal views. With an efficiency of 16.72 FPS on modest hardware, the system proves its practicality for real-world deployment. The method incorporates synthetic data augmentation and Random Subspace Methods to enhance adaptability to domain-specific conditions. Compared to existing methods like Eoe-SVM and CCM-CNN, the proposed system demonstrates a superior balance between accuracy and computational efficiency, particularly in Single Sample Per Person (SSPP) scenarios. By focusing on single-reference image recognition, the system offers a promising solution for large-scale surveillance applications, where video footage typically contains multiple poses, expressions, and lighting variations. The results highlight the system's effectiveness and efficiency, making it an excellent alternative for real-time face recognition in complex and dynamic surveillance environments.

Received: January 19, 2025 Revised: March 06, 2025 Accepted: April 08, 2025

Keywords: Face recognition; video surveillance; 3D morphable models; single reference image; domain adaptation; Deep learning

1. Inrtoduction

Face recognition (FR) has received significant attention in applications like video surveillance (VS) due to its non-intrusive nature. FR systems specialized for VS seek to detect the presence of persons of interest across a camera network under uncontrolled capture conditions. Detecting and recognizing faces in such a setting is a challenging

problem, due to the variability of face appearance. Over time, the quest for accurate face recognition techniques has attracted researchers from diverse disciplines including engineering, psychology, and computer science. However, despite extensive research, the issue of face recognition continued to be pertinent and one of the most challenging topics in surveillance monitoring systems. The images currently employed for passport verification and criminal identification, like the one shown in Figure 1.a, possess high resolution (60 pixels between the eyes) and are captured under strict lighting conditions (fast shutter speed, wide aperture, and precise focus). These images are acquired according to rigorous standards, which require the individual to look directly into the camera without displaying any unusual facial expressions or wearing face-covering items [1]. The system progresses through various complexity levels, starting with low-complexity scenarios, such as recognizing guests on the same talk as shown in Figure 1.b, to medium-complexity cases like identifying the same musicians across different stages Figure 1.c, and finally, to high-complexity scenarios, such as identifying actors or politicians in different movies spanning multiple years Figure 1.d. These images are used to determine the facial region for training purposes, including binarized versions of the selected region along with its vertical and horizontal gradient images.

a b c d

Figure 1. Face image utilized for face recognition in documents (a), face images captured from video (b, c), and a face model appropriate for video-based face analysis (d).

The earliest approaches to solve the face recognition problem involved the use of geometric features, since they are invariant to changes in illumination. This strategy performed acceptably in constrained environments, but was limited in overcoming rotations or occlusions. The next generation of face recognition techniques was based on the appearance of the face region. These methods produced mediocre results in adverse conditions like changes in illumination and occlusions, since the representation of facial features was not robust with respect to such variations. With the rapid advancement in image processing and machine learning techniques, this issue was mitigated by developing robust face representation techniques, using methods like Fisher's Linear Discriminate Analysis, Non-negative Matrix Factorization, and Local Binary Patterns [2]. Currently, much consideration is being given to using Deep Learning techniques for face representation in images, which started producing satisfactory results in this domain of face analysis. The effectiveness of such techniques in processing video sequences, however, had not been observed. Furthermore, much consideration was being given to the face recognition in videos, using multiple reference images to train the system. However, the face recognition from video sequences, using only a single reference image was much less explored. The current work addressed this issue in detail, and employed CNN based techniques for face detection and rotation estimation to localize the face regions from the input video sequence.

2. Related Work

Face detection and recognition from images or videos has emerged as one of the interesting areas of biometric research in recent times. It has gained importance in both civil and industrial applications during the past few decades. Face recognition is widely used in National Security, Lawful Interception, and other Surveillance Systems since it does not require the co-operation of the object/subject being monitored [3]. Especially in a country like Pakistan, where terrorism is in its peak state, face recognition has played an important role in the development of face-based monitoring Real Time Surveillance Systems. The actual advantages of face-based identification as compared to other biometrics are uniqueness and acceptance of the system. In developing countries miscreants may change their fingerprints or can wear some gloves in order to hide their identity from a fingerprint recognition system but they cannot change their face or wear a mask to hide their identity unless they undergo plastic surgery, which is very rare. By nature, human faces have a very high degree of variability in terms of pose, age, expression, facial hair, and illumination [4]. This high degree of variability makes face detection a very difficult problem in the field of computer vision and Artificial Intelligence (AI).

Many researchers are working on different aspects of automated face recognition and there have been tremendous advancements made in this area of research. Almost every research Publication here talks about face detection and recognition as they are interlinked problems. Work can be done on either of the two problems separately or combinely as face recognition cannot be done without face detection. This paper aims at evaluation of various face detection and recognition methods from the literature and provides a complete workable solution for image-based face detection and recognition with higher accuracy and better response rate for video surveillance [5]. A system has been developed as the first milestone for the evaluation of the proposed methods for video-based face detection and recognition for surveillance. In the following, the paper discusses in details the discovered face detection methods and face recognition methods based on the simulation results

2.1. Types of Facial Representations

When viewed at a very high level, RF systems are usually categorized into three main families in order to define a method for representing faces in memory[[6][7][8], so that these representations can be used to train a classification system that can in turn identify or verify an individual's identity. These families include (1) holistic approaches (template matching), i.e., where the face is globally represented by a set of discriminating features from one face to another; (2) local approaches (feature-based), which extract features specific to certain points of interest on the face such as the eyes, nose, mouth, and jawline; and finally (3) hybrid approaches, which symbiotically exploit the two previous techniques.

2.2. Face Recognition Techniques

Face recognition from video has gained considerable attention because it plays a crucial role in many applications such as surveillance and video indexing. The existence of drift in illumination, expression, pose, or occlusion degrades performance of matching results among frames and reference images [9]. Template based approach typically focuses on one or a few reference images, yet this straightforward approach.

However, cannot be directly applied to videos, because the frames to match may be globally transformed (i.e., rotated and/or translated) with respect to the reference image [10]. The local features-based approach achieves robustness against moderate transformation. Yet, a single reference image cannot capture information of the whole face region. A potential solution is to select samples from the reference image automatically [11]. However, it is a nontrivial task because the selected regions would commonly experience largely different transformations. The proposed face recognition system resolves this challenge by providing a framework to extract features invariant against affine transformation from both videos and a reference image. DuetFace tackles privacy issues in face recognition through the frequency domain, emphasizing high-frequency elements necessary for visual analysis while reducing sensitive data exposure. By processing only these components on the client side, it preserves privacy. An interactive block allows clients to transmit attention data to the server, improving focus on important facial features[12].

Face detection and recognition is an active area of research, which is widely used in security surveillance, human-computer interaction and video content retrieval systems. In this survey paper an effort has been made to evaluate image-based detection and recognition methods [13]. The objective of this paper includes evaluation of different image-based face detection and recognition techniques, and provide a complete solution for image-based face detection and recognition with higher accuracy and better response rate as an initial step for video surveillance. A variety of face detection approaches have been evaluated [14].

2.3. Traditional Methods

Face analysis receives growing interest from the research community with the rapid improvement of computing and video acquisition techniques, studying face recognition in video has become a hot topic. Learning-based face recognition models usually learn a mapping from pixels to a binary label or an embedding [15]. Thus, a large number of samples become necessary and this becomes a fundamental challenge for video surveillance jobs. Existing video-based face recognition methods can be mostly categorized into two types: sequentially classifying each frame into either face or non-face[16]. Long short-term memory networks are then used to model the temporal information on these classification results. The first problem is then how to accurately classify faces and non-faces at the beginning stage; such a binary classification task has already been dealt with very thoroughly in the computer vision community and researchers usually have access to satisfying solutions at an affordable computational cost. The second problem is how to model the temporal information: this refers to how to make classification decisions from the framing classification results. Initially, only a handful of temporal constraints are taken into consideration, leading to either noisy or unrobust temporal information. Researchers were then inspired to accumulate several neighboring frames by formulating a curve as a weighted linear function; however, the weights of those frames must be computed beforehand,

which is another non-trivial task [17]. The last problem is how to extract good enough features in a low-cost way: the state-of-the-art feature designs usually adopt modalities that heavily rely on a large training data volume. Nonetheless in video surveillance scenarios, the training samples are usually scarce, making it impossible to apply this kind of feature. Face recognition in surveillance scenarios from video with only a single picture is one challenging task. For this task, the output of a high-level confidence model is first combined with binary decisions from a model in the spatial domain. It is then diffused in the temporal domain. Some uncluttered batches of frames are finally constructed for verification through a two-class probabilistic voting process; effort is paid both in the modeling and searching part. Modeling approaches include manifold alignment without correspondence, patch-based probabilistic image quality assessment. Searching approaches tend to represent each frame with a combination of circular-pattern local binary features. However, those features can hardly be accumulated flexibly to capture video features [7].

2.4. Deep Learning Approaches

In recent years, with the emergence of deep learning technology, many face recognition tasks have been deeply studied by researchers through various deep learning methods. The main methods can be classified into three categories: deep learning face representation, deep metric learning, and face recognition with deep learning. For recognizing the same person in different conditions such as pose, illuminations, and expressions, it would have to be solved by face verification rather than face identification or face clustering in the first step. Since the target face image to be recognized often differs from the face sample images in the gallery severely by some conditions, the model learned by the training images in one condition cannot be directly applied to another condition. Most of the existing face recognition methods can only deal with images taken in the same condition and few methods can deal with such pictures captured in various conditions [18].Among them, the first line of research uses the cross-domain common representation learning method then classifies the test images with a domain-invariant model. This method usually explores domain-invariant feature representations extracted from both test and training images with various conditions and exploits both of them to train classifiers for the test images. Generally, a common classifier jointly trained with the collaborative regression method reduces the negative transfer by searching shared regression coefficients for the test and training images [19].In the approach, there is another line of research using either sparse or low-rank representations to jointly seek discriminative and compact face representation or automatically grouping test images to eliminate the irrelevant ones, which find them slow in handling real-time surveillance scenarios[20].

In the second line of research, a two-branch convolutional neural networks model is learned jointly with a domain-variant deeply learned representation. To do this, two branch networks are trained using both still images as a reference and the video sequences as input with the deep transfer learning which gains performance improvement over the other methods. This method, however, often suffers from a large requirement of training samples for proper training of the CNNs in deep learning. It is not finally implemented without either few training samples or the other advanced techniques such as supervised pre-training. Moreover, there is a struggle with the huge variability of the face for practical usage[21]. As for face recognition in video surveillance environments, the face image in the infrared spectrum often differs from the visible spectrum in terms of attributes like color, range, and angle.

3. Challenges in Video Face Recognition

Face recognition (FR) is critical for applications like video surveillance. The challenges include variations in pose, scale, illumination, occlusion, and blur, which complicate recognition. The increasing number of cameras and the complexity of algorithms add to the computational burden. Two categories of adaptive systems are used for still-to-video (SV) FR: one for extracting frontal faces from video frames, and the other using multiple candidates faces for matching [22] .

SV FR involves recognizing a target subject from a single still reference image. The challenges include the lack of frontal views, illumination differences, and expression variations. These systems must register non-frontal video frames using an auxiliary frontal face image. The registration process faces two key challenges: uncertainty in the 3D position of facial landmarks on the video and the unknown pose of the frontal face in the registered video [23].

3.1. Variability in Facial Expressions

Each person’s facial expressions are varied in the model training. When a model encounters a video with a different expression than in the training set, a similar model is selected for matching. If the video contains expressions not in the training set, a high reliability score is expected for the same person[24][25]. A model trained on multiple expressions (including untrained ones) can still provide a high reliability score if the person remains the same. However, a mismatch in expression causes the reliability score to decrease.

3.2. Occlusions and Lighting Conditions

Face recognition under occlusion was tested using images with simultaneous occlusions (e.g., eyebrow and mouth). Recognition performance declined due to these occlusions. The tests showed that preprocessing steps like shifting to the center and histogram equalization improved the results[26]. Additionally, the FR method was shown to be independent of lighting conditions, with images changing consistently under varying illumination angles. The method adjusts the face image to minimize loss of detail during lighting shifts[27].

4. Proposed Methodology

The proposed method is to recognize a face in a video stream by using a single reference image and extracting the face's 3D mesh. The algorithm relies on the assumption that the reference face is well-lit and frontal with respect to the camera and that it has a similar pose to the given face images in the video. The methodology is divided into two parts. The first part discusses the construction of a 3D mesh from a single reference image. The quality of the extracted 3D mesh highly relies on the quality of the detected landmarks. Hence, an extra step is performed to improve the result. The second part deals with the recognition mechanism between a generated 3D mesh and a 2D image using 3DMM. The avatarhead face model is employed that consists of algebraically deforming a template face mesh from the 3DMM shape.

Given that the method relies on detecting 68 landmarks from a reference image, these landmarks $L = \{l_1, l_2 \ldots., l_{68}\}$ can be expressed in 2D coordinates on the reference image. For each detected landmark, the corresponding 3D coordinates can be calculated using the 3D morphable model (3DMM).

A 3DMM for a given face can be represented as a linear combination of shape and expression bases:

$$S = \bar{S} + \sum_{i=1}^{k} \beta_i S_i + \sum_{i=1}^{m} \gamma_j E_j \quad (1)$$

Where:

- $\bar{S}$ is the mean shape,
- S_i are the shape basis vectors,
- β_i are the shape coefficients,
- E_j are the expression basis vectors,
- γ_j are the expression coefficients.

The 3D mesh **M** of the face is then determined by projecting the 3D coordinates of the landmarks L_{3D} onto the camera's 2D coordinates.

After detecting the 68 landmarks in 2D, the corresponding 3D coordinates are obtained by solving for the depth *f* using an optimization method. The depth *f* is obtained by minimizing the error between the 2D projections of the 3D landmarks and the detected 2D landmarks.

$$\min_f \sum_{i=1}^{n} \| p_i^{2D} - \Pi(M_i^{3D}, f_i) \|^2 \quad (2)$$

Where:

- p_i^{2D} are the 2D projections of the detected landmarks,
- Π (.) is the projection operator that maps 3D points to 2D coordinates,
- M_i^{3D} are the 3D mesh points,
- f_i represents the depth (z-coordinate) of the mesh points in the camera space.

The optimization process ensures that the 3D mesh best matches the 2D projections by adjusting the depth values.

In order to improve the fit of the detected landmarks, post-processing techniques are applied. This could be done using a regularization term or an additional constraint to reduce the discrepancy between the detected and reconstructed landmarks.

$$L_{3D} = \arg\min_{L} \left(\| \left| L - L_{3D}^{initial} \right| \|^2 + \lambda . \| \left| L - L_{ref} \right| \|^2 \right) \quad (3)$$

Where:

- $L_{3D}^{initial}$ is the initial 3D landmark set obtained from the shape fitting process,
- L_{ref} is the reference landmark set (from the frontal face),
- λ is the regularization parameter.

3D mesh construction consists of two steps: landmark detection and depth estimation of the detected landmarks. The proposed method leverages 68 landmarks detected and issued from the face alignment model, which can be expressed as 3D morphable model. The depth of these detected 3D landmarks is then obtained by a local optimization scheme of the shape fitting process that iteratively accommodates the landmarks' depth value. Besides the above major modules, a landmark post-processing step is conducted on the reconstructed mesh to better fit the detected landmarks. The model was specifically trained to enforce the model quality of frontal faces. However, the detected landmarks around eyebrows, mouth, and chin still shift to the different parts much. Therefore, capturing the landmarks around the noise regions is performed to constrain them better. once the landmarks have been refined and the mesh is constructed, the next step is to calculate the depth map for the mesh.

To compute the depth map $D(u, v)$ for each pixel (u, v) in the image, you perform a depth synthesis based on the 3DMM shape, as well as curvature adjustment for the mesh:

$$D(u, v) = arg \min_{f} \left(\sum_{i=1}^{n} \| M_i^{3D} - \prod^{-1} (p_i^{2D}, f) \|^2 + \lambda_c \| \nabla^2 D \|^2 \right) \quad (4)$$

Where:

- $D(u, v)$ is the depth at pixel (u, v)
- $\nabla^2 D$ represents the curvature of the depth map to ensure smooth transitions,
- λ_c is the regularization parameter for curvature adjustment.

This depth map is used to refine the overall 3D mesh and ensure it conforms well to the geometry of the real-world object.

The non-conforming solution for the 3DMM shape can then be calculated explicitly from only vertex locations. Given the produced parallax 3DMM shape, a vertex in the generated 3D mesh can be denoted as p = (u, v, f), where u and v are the generated faces in the video camera coordinates, and the depth f is the shape-based surface's height. The depth f representation is paramount to the shape generation as all the x's and y's are contingent upon it. Specifically, a depth map can be computed by fitting a uniform surface as well as a curvature adjustment. This depth then needs to constraint an optimization scheme that can search for fair and water-tight meshes. There are additional 3DMM parsing layers, which include local modeling and shape rendering layers that can use depth map synthesis.

The depth and vertex locations are further optimized to ensure the final mesh is fair and water-tight. This optimization ensures that the mesh has no holes and that the vertices correspond to a realistic face shape.

$$L_{mesh} = \sum_{i=1}^{n} \| P_i^{3D} - P_i^{opt} \|^2 + \lambda_{fair} . Fairness\ Term \quad (5)$$

Where:

- P_i^{opt} represents the optimized 3D vertex positions,
- L_{mesh} is the objective function for mesh optimization,
- The fairness term ensures the smoothness of the surface.

By minimizing the above equation, the algorithm ensures that the generated 3D mesh is as realistic as possible while preserving its integrity.

The Convolutional Neural Network (CNN) that learns to predict the depth map f and the coefficients βi and γj\gamma_jγj. CNNs are used because they are good at extracting spatial features from images. In the case of 3DMM fitting, the CNN is typically trained to predict the facial shape and expression coefficients from an input image.

The CNN architecture might involve:

Input Layer: The input is an image or a set of landmarks detected in 2D.

Convolutional Layers: These layers extract spatial features from the input image.

Fully Connected Layers: After feature extraction, fully connected layers are used to predict the coefficients βi\beta_iβi and γj\gamma_jγj, as well as the depth fff.

Loss Function: The network is trained to minimize the loss function, which combines the projection error and regularization terms.

5. Evaluations, Analysis of Results and Discussions

In the following section, will discuss the methodologies presented, namely synthetic image generation through geometric and pose transformations using morphable models, data augmentation through random selection of features extracted from faces, and domain-specific adaptation to cameras, in order to produce multiple facial representations of individuals to be recognized. Based on these, added to the initial system with trajectory tracking and sets of exemplary SVMs, presented in section methodology, the performance evaluations of the FRiVS system carried out during the paper will be presented in this section.

In addition, details setting out the experimental methodology that enabled the evaluation of the system variants will be discussed. Furthermore, the testing procedure and the video database used will be presented before moving on to a description of the performance metrics used to compare the proposed solutions. The evaluation results will be presented first for the initial system with trajectory tracking, but without the addition of multiple additional face representations to train the classifiers. Then, the results of the other variations, with the addition of face representations, will be discussed one by one to observe the gradual gains each tested approach brings, and thus draw progressive interpretations leading up to the final solution.

Finally, the best results will be compared to the state-of-the-art in two steps, (1) against multiple S2V-FR systems in general, or more explicitly against Eoe-SVM [28] Eoe-SVM-DS [29], ESRC-DA [30], CCM-CNN [31], CFR-CNN [32] and HaarNet [33] , then (2) for S2V-FR systems specifically with trajectory tracking, i.e. TM-FR , TM-SU [34], MFR [35], SVDL [36] and AAMT-FR Some methods in this second group also use online adaptation of the face model as in our case, which makes them interesting for comparison.

5.1. Evaluation Database Details

CASIA-WebFace Dataset: This dataset, provided by the Institute of Automation, Chinese Academy of Sciences (CASIA), includes over 10,000 identities with images and videos captured under various real-world conditions. It is widely used for evaluating face recognition systems and can serve as a good alternative for video-based face recognition tasks.

The database contains 29 potential individuals of interest. All individuals found in the CASIA-WebFace Dataset video database are presented in Figure 3 .

The methodological framework for the CASIA-WebFace dataset involves key steps, starting with preprocessing data through detection, alignment, and resizing of facial images. The dataset is then divided into training and testing subsets. A facial recognition model—like a Convolutional Neural Network (CNN) or pre-trained VGGFace—is trained using machine learning frameworks such as TensorFlow or PyTorch. Model performance is evaluated with metrics like accuracy, precision, recall, and area under the curve (AUC). Results are compared against baseline models, and quantitative analysis is conducted, with code provided for experiment reproducibility.

The 29 individuals in the video database pass through two different portals, and in two directions (entrance and exit), to generate acquisitions in distinct environments. In addition, each portal is captured using three cameras simultaneously, in order to obtain various viewpoints and facial pose angles as the individuals pass through the portals. Finally, four test sessions are conducted to obtain passages with significantly variable trajectories for each individual each time. Considering all combinations, a total of 72 different video sequences are obtained. Of these, depending on the appearance or absence of certain individuals throughout the sequences, we can count a grand total of 1,281 passages

to form the trajectories of individuals passing through the multiple portals. Among these sequences, the database also indicates which ones correspond to the most frontal facial appearances, considering the position of the three cameras at each time. We can therefore see that we have a wide diversity of members, in order to have realistic testing conditions.

For each video sequence available in the test database, and for each frame, the ground truth information indicates the individual currently observed in the scene, as well as the position of their eyes. With this information in hand, it was possible to validate the position of the ROIs generated by the trajectory tracking algorithm in combination with triple VJ detection. However, several errors in the database were found, as illustrated in Figure 2 and Figure 3, where multiple individuals are simultaneously visible in the video sequence, whereas according to the ground truths, only one should be.

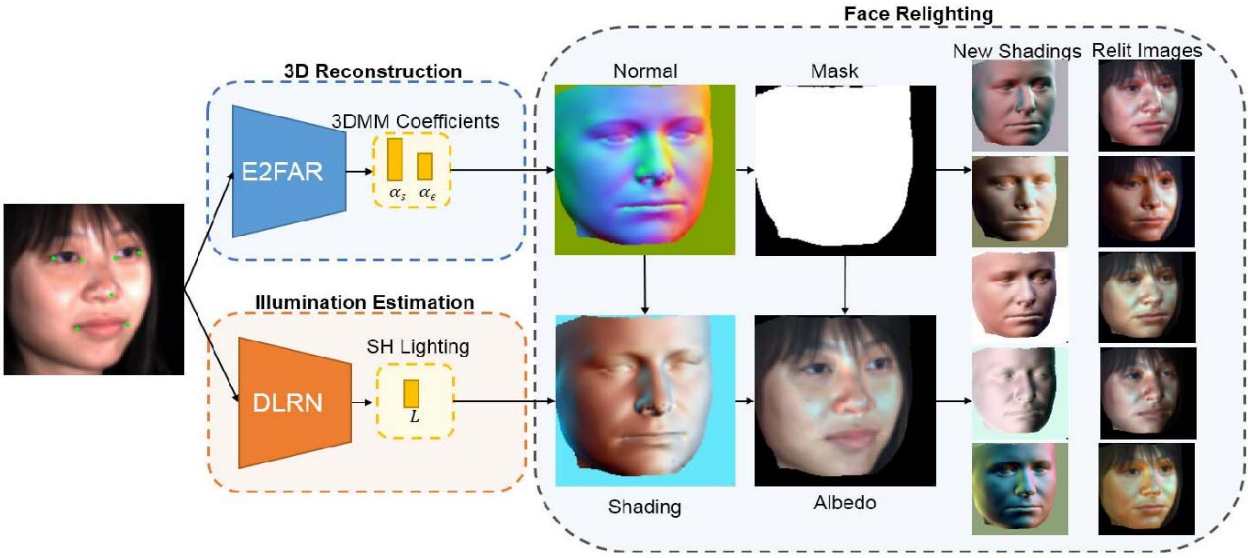

Figure 2. Block diagram of the proposed method utilizing deep learning techniques.

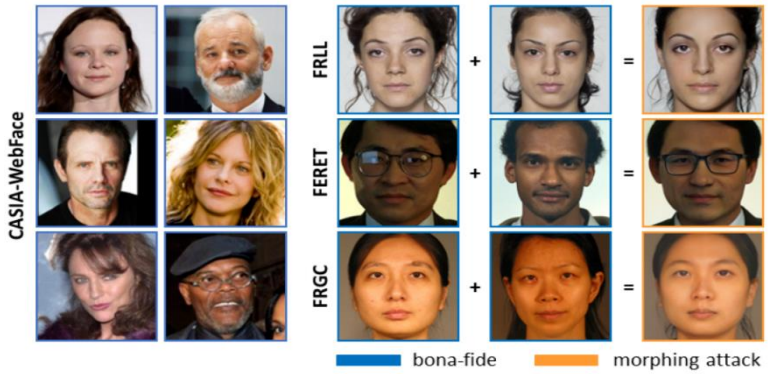

Figure 3. Face ROI of individuals forming the CASIA-WebFace Dataset.

Since this project evaluates an end-to-end FRiVS system, correctly detected faces (there is indeed someone present) generated multiple "errors," since this individual is not included in the references to validate its presence in particular. Thus, the database was cleaned as much as possible when other cases of this type were found.

Video Sequences: 72 distinct video sequences were captured, each with unique trajectories of individuals passing through the portals. These sequences represent different paths and conditions to simulate a range of real-world scenarios.

Ground Truth: For each video sequence, the ground truth includes the identity of the person in view and the position of their eyes, which is crucial for validating the performance of face recognition systems and trajectory tracking algorithms.

Testing Conditions: The dataset provides a variety of environmental conditions, facial poses, and angles, allowing for a more comprehensive evaluation of face recognition algorithms under realistic scenarios.

However, there must still be situations that were not detected considering the size of the data to be analyzed. Therefore, although the performance ultimately obtained is very satisfactory (see section 5.3.2), it should be considered that some of these cases probably still cause a slight performance decrease.

5.2. Face Recognition Results

This section will present the results obtained in two stages: first with the initial system using only trajectory-based accumulation of individual faces captured by security cameras, and then with the different combinations of data augmentation and synthetic generation methods tested. To clarify and simplify the overall results obtained, only the summary performance of the sequences from cameras displaying "frontal" faces, as well as the evaluation on the entire CASIA-WebFace Dataset database, will be presented here.

5.3. Initial System with Trajectory Tracking

The following results correspond to the initial system. This is therefore used as a baseline for comparing the other variations that will follow in the next section.

Table 1 shows that the performance of the initial system is already extremely strong for the front-facing camera case in identification mode. However, the other cases are far from reaching state-of-the-art performance levels. Thus, it can be seen that pose has a major impact on the results, justifying why some synthetic face generation techniques were designed specifically for this problem. Moreover, in surveillance mode, performance is not sufficient in all cases.

Table 1: Average results of 5 test replications of the initial FRiVS system with face trajectory tracking, for surveillance and identification modes, on front-facing camera sequences and the entirety of CASIA-WebFace Dataset .

	Training variation	Initial system	
Test sequences	Performance measurement	pAUC (20%)	AUPR
Front cameras	Surveillance	86.75% ± 0.66%	67.69% ± 4.55%
	Identification	97.36% ± 2.32%	97.71% ± 1.57%
The whole base	Surveillance	77.30% ± 0.31%	61.23% ± 5.66%
	Identification	69.98% ± 0.70%	61.30% ± 0.09%

5.4. System with Synthetic Generation and Domain Adaptation

The results of the first four synthetic generation combinations explored, forming the first test block considered during the system design, are all presented in Table 2. These combinations use only synthetic generation in the form of geometric transformations and the addition of unique reference images of other individuals in the cohort as additional counterexamples.

Table 2: Average results of 5 system test replications with the combinations ST_{nt}, ST_{nt}+GT_{nt}, GT_{wl}+ST_{nt}, and GT_{wl}+ST_{nt}+GT_{nt} for surveillance and identification modes, on front-facing camera sequences and the entire CASIA-WebFace Dataset system (optimal results in bold, both for these combinations and compared to the initial system).

	Training variation	STn t		STn t+GTn t		GTwl+STn t		GTwl+STn t+GTn t	
Test sequences	Performance measurement	pAUC (20%)	AUPR	pAUC (20%)	AUPR	pAUC (20%)	AUPR	pAUC (20%)	AUPR
Front cameras	Surveillance	91.51% ± 0.24%	77.81% ± 3.07%	91.32% ± 1.53%	81.66% ± 2.25%	94.94% ± 0.56%	89.68% ± 0.19%	95.51% ± 0.26%	89.95% ± 0.23%
	Identification	96.06% ± 3.09%	96.03% ± 1.00%	97.35% ± 1.76%	96.76% ± 0.02%	98.33% ± 1.67%	98.22% ± 1.78%	98.69% ± 0.86%	98.25% ± 0.75%
The whole base	Surveillance	79.99% ± 0.16%	92.12% ± 0.01%	81.12% ± 0.56%	66.96% ± 5.36%	82.92% ± 1.04%	72.35% ± 0.79%	83.96% ± 1.07%	73.01% ± 1.70%
	Identification	65.56% ± 4.69%	91.40% ± 0.55%	92.38% ± 1.86%	91.15% ± 3.00%	93.84% ± 0.79%	93.46% ± 1.11%	94.56% ± 1.40%	93.79% ± 1.79%

We can quickly see from the results in Table 2 that the best performances are all achieved with the GT_{wl}+ST_{nt}+GT_{nt} test combination, i.e., when the maximum number of new face representations, regardless of class, is added for the highly specialized training of Eoℓ-SVM per individual in the watchlist.

Given the results obtained in the first test block performed (evaluation of geometric transformations), it was possible to determine that the system had been successfully improved, compared to the initial results in Table 2.However, while the results on the frontal sequences were significantly better, both in surveillance and identification, performance across the entire database remained relatively low (compared to the state-of-the-art) in the surveillance mode.

This suggested that the resulting models were very robust to multiple acquisition conditions, such as illumination or motion blur, but still seemed to suffer from significant facial poses in the video sequences, as this is the main distinction observable between the viewpoints of frontal sequence videos and the rest. This observation led to the second block of tests, which attempts to specifically address the pose problem by using 3D models to generate these probable facial variations based on the profile observations actually captured in the scene.The results of this second block are presented in Table 3, where both the use of 3DMM alone, as well as the combination with the best variation retained from the previous test block, allow for the synthetic generation of face models with three-dimensional poses.

Table 3: Average results of 5 system test replications with the combinations 3DMM alone and 3DMM+GT_{wl}+ST_{nt}+GT_{nt} for surveillance and identification modes, on front-facing camera sequences and the entire CASIA-WebFace Dataset.

	Training variation	3DMM		3DMM + GT_{wl}+ST_{nt}+GT_{nt}	
Test sequences	Performance measurement	pAUC (20%)	AUPR	pAUC (20%)	AUPR
Front cameras	Surveillance	78.05% ± 2.82%	62.15% ± 3.17%	72.30% ± 0.93%	85.04% ± 1.10%
	Identification	91.14% ± 7.14%	91.73% ± 4.96%	87.49% ± 4.52%	93.01% ± 0.73%
The whole base	Surveillance	72.30% ± 0.93%	95.54% ± 0.24%	54.04% ± 2.63%	75.02% ± 1.39%
	Identification	87.49% ± 4.52%	95.75% ± 0.36%	86.22% ± 4.34%	92.55% ± 0.05%

Unfortunately, the performance obtained in this second test block was not as satisfactory as in the previous block, even when both approaches (synthetic generation by geometric translations and 3DMM models) were used together. Compared to the results of the initial system in Table 3, the addition of synthetic posed faces with 3DMM only deteriorated the FRiVS system. When geometric transformation approaches were combined with 3DMM, performance was now better than the initial system, but still lower than the GTwl+STnt+GTnt combination alone. Thus, it is highly likely that the sole reason for the performance gain of 3DMM+GTwl+STnt+GTnt over the initial system was induced by the synthetic generation of geometric transformations, which demonstrated a significant improvement.

However, based on these experimental results, it was considered that the 3D models generated to represent the faces of individuals of interest were probably not sufficiently representative or faithful to the expected observations on video cameras. Given that the Eoℓ-SVMs used attempt to very faithfully represent individuals with a limited number of examples, adding insufficiently discriminative representations to differentiate the individual of interest from others only adds noise to the models, which are already extremely sensitive to the small variations present in the positive class. Indeed, given the very limited number of support vectors available on the positive class side, the Eoℓ-SVMs are quickly affected by any outliers added during training. This is why it is essential to have extremely high-quality representations relative to the expected observations in the scene.

In the case of geometric transformations, since the additional face models introduced were simply variations in pixel position, scaling, or blurring, the distinctions from the reference images remained relatively controlled and close together in the vector hyperspace. This does not appear to be the case for models generated by 3DMM, as too much intra-class variation appears to be introduced, and the classifiers become confused by these poor examples of new faces.

The conclusions drawn from the results of the second test block led to the elimination of synthetic pose generation using 3D models from the final system, which necessitated the search for alternative methods to improve FR's performance while remaining faithful to the expected facial representations in videos. This explains the introduction of the third test block, which leverages the generation of multiple representations from various support vectors, directly by selecting features that had been previously extracted rather than by using representations from the image.Considering that the GTwl+STnt+GTnt combination had empirically demonstrated that the resulting vectors corresponded to valid and robust representations of faces expected in the scene for the individuals recorded on the watchlist, the reasoning was that the underlying features composing them were also adequate for discriminating faces. Thus, the RSM method, which only randomly selects some of these features in order to recombine them into distinct subsets, only exploits the features confirmed to be robust to the video sequence acquisition conditions. The application of RSM, both alone and combined with the GTwl+STnt+GTnt generations, produced the results shown in Table 4.

Table 4: Average results of 5 system test replications with the combinations RSM(20,128) alone and RSM(20,128)+GT_{wl}+ST_{nt}+GT_{nt} for surveillance and identification modes, on front-facing camera sequences and the entire CASIA-WebFace Dataset (optimal results in bold, both for these combinations and compared to GT_{wl}+ST_{nt}+GT_{nt}).

	Training variation	RSM(20,128)		RSM(20,128) + GT_{wl}+ST_{nt}+GT_{nt}	
Test sequences	Performance measurement	pAUC (20%)	AUPR	pAUC (20%)	AUPR
Front cameras	Surveillance	94.15% ± 0.70%	89.66% ± 1.49%	94.40% ± 1.03%	89.09% ± 3.42%
	Identification	99.22% ± 0.44%	96.87% ± 2.16%	99.70% ± 0.11%	98.32% ± 1.28%
The whole base	Surveillance	83.53% ± 0.86%	77.06% ± 1.86%	82.61% ± 0.09%	75.84% ± 2.87%
	Identification	94.52% ± 1.19%	95.47% ± 1.70%	94.09% ± 0.24%	95.22% ± 0.81%

When used simply, the RSM approach is applied to the original HOG588 descriptors derived from ROI patches of single static reference images of the targets, as well as for patches of unknown ROIs from video sequences. When RSM is combined with geometric transformations (RSM (20,128) + GTwl+STnt+GTnt), the latter are first applied to obtain the set of synthetic face variations, in order to subsequently use RSM on them. The faces obtained from video sequences in operational mode are directly processed with the same operations as in training for both combinations, with or without synthetic images.

Thus, rather than obtaining sets of Np vector-based SVMs with dimensionality Nd = 588, we obtain sets of $Np \;\; Nrs$ vector-based SVMs with Nd = 128 features. Given the number of randomly generated vectors and the number of features used by each, several of the original features are transferred redundantly, which means that certain combinations of relationships between them are preserved, while simplifying the overall complexity of the descriptors by reducing their dimension. Also, by adding multiple variations of data subsets to the feature hyperspace, we obtain ensembles of classifiers that are potentially more competent at classifying certain very distinct facial features or for certain very specific cases of difficult acquisition conditions on videos. In other words, some of the classifiers in the ensemble can become even more specialized to classify more different combinations, and the best ones generally improve the final prediction of the individual to be recognized.

The results presented in Table 4 indeed seem to show that the RSM approach has slightly improved some of the evaluation cases, which supports the idea that competent classifiers are surely obtained with this new generation of representations of descriptor vectors extracted from faces. However, we also note that the improvements made mainly for identification come at the expense of some performance drops in surveillance recognition. Continuing to improve the FRiVS system, a final adjustment was made:

considering the highly specific selection of unknown individuals from each camera individually. This allows for highly specialized Eoℓ-SVM models through training adapted to the specific domain of each observed scene. This final test block uses the most optimal combination to date (GTwl+STnt+GTnt), taking into account all the test scenarios evaluated (frontal/full sequences and surveillance/identification mode). Although the number of counterexample videos is greatly reduced due to the camera-specific selection (we easily go from 20,000 faces distributed across all sequences to approximately 500 per camera), synthetic generation using geometric transformations is not affected, as it only uses single reference images to register individuals of interest to the control list. The results obtained for this evaluation are presented in Table 5.

Table 5: Average results of 5 system test replications with the combination DA+GTwl+STnt+GTnt for surveillance and identification modes, on front-facing camera sequences and the entirety of CASIA-WebFace Dataset (optimal results in bold compared to all previous tests).

	Training variation	DA + GT_{wl}+ST_{nt}+GT_{nt}	
Test sequences	Performance measurement	pAUC (20%)	AUPR
Front cameras	Surveillance	97.00% ± 1.09%	88.94% ± 1.20%
	Identification	96.75% ± 0.69%	92.95% ± 2.61%
The whole base	Surveillance	90.48% ± 1.87%	85.87% ± 1.53%
	Identification	95.45% ± 0.05%	95.92% ± 0.40%

We can see from the results obtained that domain adaptation definitely benefits the system's performance, especially when comparing the deviations obtained for the monitoring mode with all other tests. Thus, not only is the specialization of the Eoℓ-SVMs to the specific domain effective, but it is also possible to consider that the reduced selection of counterexamples makes the resulting classes significantly less unbalanced than before. This very likely helped the classifiers better define the separation margins.

5.3. Summary of Results and Comparison with the State-of-the-Art

The best test combinations obtained for all the evaluations carried out during this project are presented in Table 6 and Table 7. Thus, we can see that, depending on the desired FRiVS system operating mode, different approaches should be considered because performance varies significantly between each case. This also reflects the fact that there is rarely a single solution for all situations.

Table 6: Average results of 5 system test replications with the best combinations of surveillance and identification modes, on front-facing camera sequences and the entirety of CASIA-WebFace Dataset (optimal results in bold compared to all tests performed).

	Training variation	GTwl+STn t+GTn t		RSM(20,128) + GTwl+STn t+GTn t		DA + GTwl+STn t+GTn t	
Test sequences	Performance measurement	pAUC (20%)	AUPR	pAUC (20%)	AUPR	pAUC (20%)	AUPR
Front cameras	Surveillance	95.51% ± 0.26%	89.95% ± 0.23%	94.40% ± 1.03%	89.09% ± 3.42%	97.00% ± 1.09%	88.94% ± 1.20%
	Identification	98.69% ± 0.86%	98.25% ± 0.75%	99.70% ± 0.11%	98.32% ± 1.28%	96.75% ± 0.69%	92.95% ± 2.61%
The whole base	Surveillance	83.96% ± 1.07%	73.01% ± 1.70%	82.61% ± 0.09%	75.84% ± 2.87%	90.48% ± 1.87%	85.87% ± 1.53%
	Identification	94.56% ± 1.40%	93.79% ± 1.79%	94.09% ± 0.24%	95.22% ± 0.81%	95.45% ± 0.05%	95.92% ± 0.40%

Table 7: Average results $F\beta$ of 5 system test replications with the best combinations retained for surveillance and identification modes, on front-facing camera sequences and the entirety of CASIA-WebFace Dataset.

	Training variation	RSM(20,128)		RSM(20,128) + GT_{wl}+ST_{nt}+GT_{nt}	
Test sequences	Performance measurement	F1	F2	F1	F2
Front cameras	Surveillance	56.81% ± 9.93%	65.19% ± 4.96%	48.44% ± 8.27%	56.28% ± 7.18%
	Identification	96.85% ± 0.30%	98.36% ± 0.23%	89.37% ± 3.65%	92.62% ± 1.72%
The whole base	Surveillance	54.77% ± 11.96%	62.60% ± 7.55%	52.91% ± 9.72%	63.23% ± 5.19%
	Identification	91.57% ± 0.03%	89.97% ± 0.10%	89.77% ± 0.49%	91.62% ± 0.08%

To evaluate the results obtained using state-of-the-art methods, two categories of algorithms are considered. The first category includes some of the best classifiers found and used specifically for FRiVS in S2V mode, regardless of the underlying techniques. The second category involves only methods explicitly implementing a face trajectory tracking approach for recognition, also based on S2V and training in SSPP conditions, which includes the additional complexities of FT, such as fail-to-acquire errors, which are simply not considered in the first category. Thus, the comparison with the first category is intended to be somewhat more pessimistic, in order to gain insight into the additional improvements that would be desirable to obtain a system minimally equivalent to the state-of-the-art, assuming no FD or FT errors are encountered. On the other hand, the second comparison is more reasonable, because

the systems being compared operate on the same basis and are therefore definitely more equivalent for comparison based on the same challenges encountered, including pixel misalignment and trajectory drift.

To simplify the table interface, the best solution using the RSM method is annotated as Eoℓ-SVM-RSM (instead of RSM(20,128) + GTwl+STnt+GTnt), while the one applying DA is annotated as Eoℓ-SVM-DA (instead of DA+GTwl+STnt+GTnt). Both use their method in combination with multiple synthetic generations through geometric transformations of the reference face images of the individuals of interest. The computational complexity in terms of operations and parameters are respectively determined by $(9 \cdot 20 \cdot 46 \cdot 128)$ and $(9 \cdot 1 \cdot 46 \cdot 588)$ for Eoℓ-SVM-RSM and Eoℓ-SVM-DA, given that on average $Nsv = 46$ support vectors were generated by each sub-SVM. The computational speed of both methods is approximately 0.0598±0.01 s/frame (~16.72 FPS) since the SVM ensemble prediction operations are performed in parallel with multiple optimization levels, meaning that the slowest and limiting part of the system is actually the triple Viola-Jones detection. The time required for the classification of faces by Eoℓ-SVM is therefore largely negligible compared to the overall FD over the entire video frame.

The comparative results of the different categories of FRiVS systems, as just described, are presented in Table 8 and Table 9, respectively. We can thus see that the optimal Eoℓ-SVM variants selected achieve results that approach the best state-of-the-art S2V FRiVS systems for all conditions and operating modes combined, despite the added FT difficulties.

Table 8: Comparison of the performance of the best variants of the system with other state-of-the-art S2V methods, all situations combined, on all CASIA-WebFace Dataset sequences.

FRiVS system in S2V mode	Eoe-SVM	Eoe-SVM-DS(1)	Eoe-SVM-DS(2)	ESRC-DA	CCM-CNN	CFR-CNN	HaarNet	Eoℓ-SVM-RSM	Eoℓ-SVM-DA
pAUC(20%)	100.0±0.00	97.52±0.50	100.0±0.00	N/A	N/A	N/A	N/A	94.09±0.24	95.45±0.05
AUPR	99.24±0.38	96.86±0.72	99.31±0.46	76.97±0.07	98.87±0.63	96.47±0.86	99.36±0.59	95.22±0.81	95.92±0.40
operations	2.23M	114K	N/A	228M	33.3M	3.75M	3.50B	1.06M	243K
parameters	230K	99K	N/A	41.5M	2.40M	1.20M	13.1M	1.06M	243K

Table 9: Performance comparison of the best system variants with other state-of-the-art S2V methods based on SSPP and with trajectory tracking, on the entirety of the CASIA-WebFace Dataset sequences, in transaction and trajectory modes.

FRiVS system in S2V mode and from SSPP and trajectory tracking		TM-FR	TM-SU	MFR	SVDL	AAMT-FR	Eoℓ-SVM-RSM	Eoℓ-SVM-DA
Transaction	pAUC(5%)	0.249±0.02	0.347±0.03	0.417±0.03	0.439±0.03	0.494±0.02	0.406±0.03	0.469±0.03
	AUPR	0.457±0.04	0.452±0.05	0.512±0.05	0.533±0.05	0.588±0.04	0.495±0.04	0.565±0.03
Path	pAUC(5%)	0.318±0.02	0.391±0.02	0.472±0.02	0.583±0.02	0.649±0.07	0.711±0.03	0.823±0.02
	AUPR	0.423±0.03	0.478±0.02	0.532±0.02	0.631±0.02	0.793±0.03	0.758±0.03	0.859±0.02
Computation time (s/frame)		0.1910±0.05	0.2920±0.07	0.2500±0.08	0.2370±0.05	0.2170±0.06	0.0598±0.01	0.0598±0.01

In the case of methods specifically employing tracking during the operation phase (Table 9), and conditioned by the SSPP for training and the S2V-FR operating mode, increasing the number of face representations significantly outperforms state-of-the-art systems, especially when using spatio-temporal trajectory information. When only transaction-based predictions are used, i.e., each frame individually, we experience a significant performance loss.

Nevertheless, this latter mode still produces face recognition results that are similar to, or even outperform, existing methods. The case of AAMT-FR is particularly interesting in this respect given that the tracker's internal face model is adapted online based on new facial appearances captured in the scene, i.e., using the same principle as the trajectory tracking module used in this project. This makes this system likely the best basis for comparison to demonstrate the gains in face recognition accuracy achieved.

We can also see from Table 9 that, although the performance of the approaches proposed by Eoℓ-SVM is slightly lower than existing systems in the literature, which are mainly large CNN-based models, the FRiVS solutions proposed for S2V are clearly more memory-efficient, especially in the case of Eoℓ-SVM-DA, given the reduced number of parameters required by the support vectors to represent a model of an individual of interest in the watchlist. The Eoe-SVM and Eoe-SVM-DS methods are good points of comparison for defining future objectives and avenues for future development, as they are basically the same classifiers as the Eoℓ-SVM used here.

However, as already mentioned, the results of these two techniques do not consider the added complexity of trajectory tracking, which is necessary to distinguish between various individuals simultaneously in the scene to obtain a complete FRiVS system. Thus, these two methods are good benchmarks in the event that we succeed in eliminating the observed problems related to trajectory tracking.

The performance demonstrate the major performance gains achieved by using highly specialized Eoℓ-SVMs for individual recognition trained using SSPP. This robustness is further enhanced by applying facial appearance trajectory tracking in the surveillance camera scene for spatio-temporal accumulation of FR scores for the targeted individuals. Furthermore, the use of synthetic generation based on geometric transformations on reference images of individuals of interest, combined with random feature selection for even more face representations, or with highly specific domain adaptation for each camera, still results in an even more robust FRiVS system. This makes S2V-FR effective on video sequences even if they are marked by significant capture conditions in a semi-controlled environment.

It is also important to place greater emphasis on the achieved system speed,This operating mode execution speed, which approaches approximately 16.72 FPS, is achieved using a relatively simple machine, compared to several studies in the literature, which often use much more powerful GTX1080 or TITAN graphics cards. This indicates that the proposed FRiVS system architecture (Figure 1) is extremely optimal for a lightweight application and conducive to even greater performance if deployed on a higher-end dedicated machine or, even more so, on a distributed server specifically for video surveillance.

6. Conclusion

The conclusions drawn from the results of the second test block led to the elimination of synthetic pose generation using 3D models from the final system. This necessitated the search for alternative methods to improve the Face Recognition (FR) performance while remaining faithful to the expected facial representations in videos. This explains the introduction of the third test block, which leverages the generation of multiple representations from various support vectors, directly by selecting features that had been previously extracted rather than using representations from the image. Considering that the GTwl+STnt+GTnt combination had empirically demonstrated that the resulting vectors corresponded to valid and robust representations of faces expected in the scene for the individuals recorded on the watchlist, it was concluded that the underlying features composing them were also adequate for discriminating faces. Thus, the Random Subspace Method (RSM), which only randomly selects some of these features to recombine them into distinct subsets, exploits the features confirmed to be robust to the video sequence acquisition conditions. The application of RSM, both alone and combined with the GTwl+STnt+GTnt generations.

References

[1] D. Gorodnichy and G. Bessens, “From recognition in brain to recognition in perceptual vision systems. Case study: Face in video. Example: Identifying computer users with low-resolution webcams,” in *Proc. 3rd Int. Conf. Vision, Video Graph.*, 2005.

[2] R. K. K. Reddy, S. J. K. S. Reddy, and K. P. R. Reddy, "A survey on face recognition techniques: Challenges and solutions," *Int. J. Comput. Appl.*, vol. 975, no. 14, pp. 1–7, 2020.

[3] M. Abdul-Al *et al.*, "The evolution of biometric authentication: A deep dive into multi-modal facial recognition: A review case study," *IEEE Access*, vol. 12, pp. 50689–50721, 2024.

[4] M. Zamir *et al.*, "Face detection & recognition from images & videos based on CNN & Raspberry Pi," *Computation*, vol. 10, no. 9, p. 148, 2022.

[5] Tvoroshenko and V. Kukharchuk, "Current state of development of applications for recognition of faces in the image and frames of video captures," in *Proc. IEEE 16th Int. Conf. Adv. Trends Radioelectron., Telecommun. Comput. Eng. (TCSET)*, 2021, pp. 682–686.

[6] M. Latif *et al.*, "Face recognition from video by matching images using deep learning-based models," *VAWKUM Trans. Comput. Sci.*, vol. 12, no. 2, pp. 50–64, 2024.

[7] H. L. Gururaj *et al.*, "A comprehensive review of face recognition techniques, trends and challenges," *IEEE Access*, vol. 12, pp. 31114–31151, 2024.

[8] Anil *et al.*, "Literature survey on face recognition of occluded faces," in *Proc. 7th Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, 2024, pp. 1930–1937.

[9] H. Castañeda Rincón and O. Santos Ariza, "Estrategia para la implementación de herramientas con reconocimiento facial en los Sistemas Integrados de Emergencias y Seguridad (SIES)," M.S. thesis, Univ. Dist. Francisco José de Caldas, Bogotá, Colombia, 2021.

[10] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–42, 2022.

[11] F. X. Gaya-Morey *et al.*, "Deep learning-based facial expression recognition for the elderly: A systematic review," *arXiv Prepr.*, arXiv:2502.02618, 2025.

[12] Y. Mi *et al.*, "Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 6755–6764.

[13] S. Hangaragi and T. Singh, "Face detection and recognition using face mesh and deep neural network," *Procedia Comput. Sci.*, vol. 218, pp. 741–749, 2023.

[14] B. Amirgaliyev *et al.*, "A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications," *Sensors*, vol. 25, no. 5, p. 1410, 2025.

[15] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *J. Vis. Commun. Image Represent.*, vol. 77, p. 103116, 2021.

[16] C. Jiang *et al.*, "Object detection from UAV thermal infrared images and videos using YOLO models," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 112, p. 102912, 2022.

[17] D. Cozzolino *et al.*, "Raising the bar of AI-generated image detection with CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 4356–4366.

[18] M. Faraki, X. Yu, Y.-H. Tsai, Y. Suh, and M. Chandraker, "Cross-domain similarity learning for face recognition in unseen domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15292–15301.

[19] T. Liu *et al.*, "Cross-domain facial expression recognition via disentangling identity representation," in *Proc. 32nd Int. Joint Conf. Artif. Intell. (IJCAI)*, 2023, pp. 1213–1221.

[20] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6678–6687.

[21] Y. Gao *et al.*, "Cross-domain facial expression recognition through reliable global–local representation learning and dynamic label weighting," *Electronics*, vol. 12, no. 21, p. 4553, 2023.

[22] H. Wang, C. Liu, and X. Ding, "Still-to-video face recognition in unconstrained environments," in *Proc. SPIE 9405, Image Process.: Mach. Vis. Appl. VIII*, 2015, p. 94050H.

[23] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Appl. Sci.*, vol. 12, no. 18, p. 8972, 2022.

[24] E. S. Leif *et al.*, "A systematic review of social-validity assessments in the Journal of Applied Behavior Analysis: 2010–2020," *J. Appl. Behav. Anal.*, vol. 57, no. 3, pp. 542–559, 2024.

[25] G. Petmezas *et al.*, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors*, vol. 22, no. 3, p. 1232, 2022.

[26] R. Sheela and R. Suchithra, "Unmasking the masked: Face recognition and its challenges using the periocular region–A review," in *Handbook of Research on Technical, Privacy, and Security Challenges in a Modern World*. IGI Global, 2022, pp. 62–81.

[27] D. Wang, Y. Gu, L. Luo, and F. Ren, "Occlusion-aware visual-language model for occluded facial expression recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2024, pp. 1–8.

[28] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Robust watch-list screening using dynamic ensembles of SVMs based on multiple face representations," *Mach. Vis. Appl.*, vol. 28, no. 2, pp. 219–241, 2017.

[29] Z. Yu *et al.*, "Hybrid incremental ensemble learning for noisy real-world data classification," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 403–416, 2017.

[30] F. Nourbakhsh, E. Granger, and G. Fumera, "An extended sparse classification framework for domain adaptation in video surveillance," in *Proc. Asian Conf. Comput. Vis. Workshops*, 2016, pp. 360–376.

[31] M. Parchami, S. Bashbaghi, and E. Granger, "CNNs with cross-correlation matching for face recognition in video surveillance using a single training sample per person," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, 2017, pp. 1–6.

[32] N. K. Mishra and S. K. Singh, "Face recognition using 3D CNN and Hardmining loss function," *SN Comput. Sci.*, vol. 3, no. 2, p. 155, 2022.

[33] N. A. M. Ariffin, U. A. Gimba, and A. Musa, "Face detection based on Haar cascade and convolution neural network (CNN)," *J. Adv. Res. Comput. Appl.*, vol. 38, no. 1, pp. 1–11, 2025.

[34] F. Roli and G. L. Marcialis, "Semi-supervised PCA-based face recognition using self-training," in *Proc. Struct., Syntactic, Stat. Pattern Recognit.*, 2006, pp. 560–568.

[35] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Watch-list screening using ensembles based on multiple face representations," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 4489–4494.

[36] M. Yang, L. Van Gool, and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 689–696.