Ain Shams Engineering Journal 17 (2026) 103824

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Full Length Article

Intelligent recognition of visual speech based on custom CNN and Optimized Quaternion Charlier Moments using PSO

Omar El Ogri [a,b], Jaouad EL-Mekkaoui [a], Mohamed Benslimane [a], Amal Hjouji [c], Abdelali Saidi [d], Musheer Ahmad [e,*], Hela Elmannai [f], Ahmed A. Abd El-Latif [g,h]

[a] Laboratory of Sciences, Engineering and Management (LSEM), Higher School of Technology, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[b] Laboratory of Sustainable Agriculture Management (LSAM), Higher School of Technology of Sidi Bennour, Chouaib Doukkali University, El Jadida, Morocco
[c] Laboratory of Information, Signals, Automation, and Cognitivism (LISAC), Dhar El Mahrez Faculty of Science, Sidi Mohamed Ben Abdellah-Fez University, Fez, Morocco
[d] Laboratory of Optimization, Emerging Systems, Networks and Imaging, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco
[e] Department of Computer Engineering, Jamia Millia Islamia, New Delhi 110025, India
[f] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[g] EIAS Data Science Lab, College of Computer and Information Sciences, and Center of Excellence in Quantum and Intelligent Computing, Prince Sultan University, Riyadh 11586, Saudi Arabia
[h] Jadara University Research Center, Jadara University, Jordan

ARTICLE INFO

Keywords:
Visual Speech Recognition
Optimized Quaternion Charlier Moments
Shallow Convolutional Neural Network
Lip-reading
Particle Swarm Optimization

ABSTRACT

This paper presents a novel model termed as Optimized Quaternion Charlier Moments Convolutional Neural Network (QCMs-PSO-CNN) to develop an intelligent method of recognizing the visual speech accurately that is phrase or word that a person spoke in the video through Lip-reading. The proposed method put forward the Optimized QCMs using our enhanced particle swarm optimization (PSO). The suggested PSO is improvised on linear inertia weights and a sine–cosine learning factor, which effectively strengthens the optimization capability. This method suggested a custom-built shallow CNN which incorporates optimized QCM as a filter in the first layer in a novel way for better recognition ability of the method. The study shows that this method is a good solution for reducing the high video image dimension and gaining time for training. The proposed architecture QCMs-PSO-CNN found to classify the digits, letters, or words effectively. Three standard video datasets such as GRID, LRW, and GLips are chosen to evaluate the performance of the proposed method. The experimental analyses show that the proposed method attains excellent recognizing performance results compared to the other recently investigated approaches which made use of complex models and deeper architecture.

1. Introduction

Speech recognition from video can be used in many applications: speech recognition in a video without sound, re-dubbing, message recognition in noisy environments, even in spying systems. Lip reading, also called visual speech recognition (VSR), is the skill of understanding spoken language by visually analyzing a speaker's lip movements. Thanks to considerable progress in computer vision, signal processing, and pattern recognition, lip reading is now widely used. Its importance is particularly evident in many fields, including medicine and security. Lip-reading systems significantly aid individuals with hearing loss or disability by improving their communication abilities and facilitating greater participation in social interactions. Lip-reading is essential in the security sector, especially for evaluating surveillance footage to discern verbal communications at precise instances. Implementing speech recognition systems presents problems, including the extensive range of phonemes (ranging from 45 to 53) and their intrinsic similarities, as noted by Fisher [1]. Furthermore, when two or more words are pronounced the same way, for example"right" and"write". These difficulties result from the fact that all people don't talk the same way and the differences between the shapes of the speaker's mouths when they are opened, widely opened, closed, or tightly closed, the appearance of teeth, and tongue articulations. Extracting features and recognition require a robust lip reading system that can approach all these

* Corresponding author at: Associate Professor, Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India.
E-mail addresses: musheer.cse@gmail.com (M. Ahmad), a.rahiem@gmail.com (A.A. Abd El-Latif).

https://doi.org/10.1016/j.asej.2025.103824
Received 12 May 2025; Received in revised form 17 August 2025; Accepted 14 October 2025
2090-4479/© 2025 The Author(s). Published by Elsevier B.V. on behalf of Faculty of Engineering, Ain Shams University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

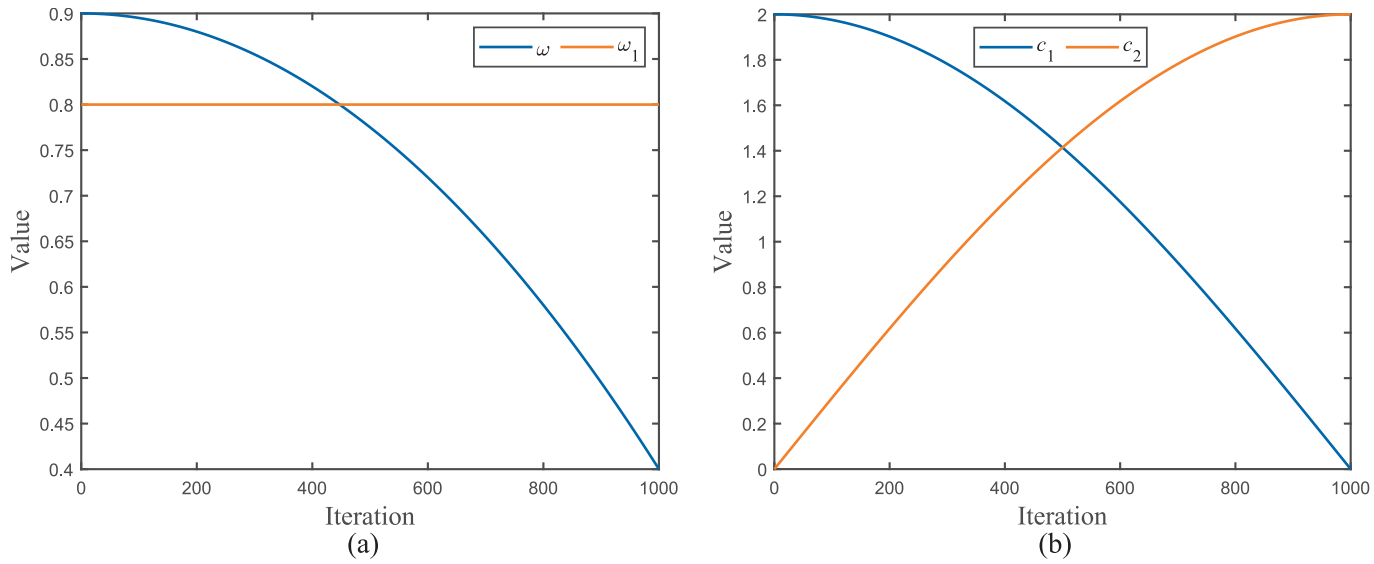


Fig. 1. (a) Weight of inertia. (b) Factor for learning sine–cosine.

Table 1
Initial parameter settings of all algorithms.

Algorithm	Parameters	Value
Proposed PSO	nPop (Population size), *T*, coefficients(C_1,C_2), Inertia constants ω	30, 500, [0,2], [0.4,0.9]
BBO	Keep Rate (Kr), parameters (α), Mutation (p)	0.2, 0.9, 0.1
ACO	Intensification Factor (q),Deviation-Distance Ratio (ζ)	0.5, 1
ICA	Selection Pressure (α), Assimilation Coefficient (β), Revolution Probability (p), Revolution Rate (μ), Colonies Mean Cost Coefficient (ζ)	1, 1.5, 0.05, 0.1, 0.2
IWO	Number of Seeds(S), Variance Reduction Exponent (E), Value of Standard Deviation(σ)	[0, 5], 2, [0.5,0.001]
SFLA	Memeplex Size (nPM), Number of Memeplexes (nM)	10, 5
CA	Sensitive parameters (α, β)	0.3, 0.5
DE	Bound of Scaling Factor (β), Crossover Probability (pCR)	[0.2, 0.8], 0.2
FA	Mutation Coefficient (α), Attraction Coefficient Base Value (β), Light Absorption Coefficient (γ)	0.2, 2, 1
HS	Harmony Memory Size (HMS), Number of New Harmonies (n), Harmony Memory Consideration Rate (HMCR), Pitch Adjustment Rate (p)	5, 20, 0.9, 0.1

variations. In order to know who performed better, an automated system or human lip-readers, Hilder et al. in [2] have shown by the experiment that automated systems outperform human lip-readers. Adebayo, Bakare Mustaphaa et al. in [3] introduced a Comparative Analysis of Deep Learning Models for Part of Speech Tagging. Therefore, to approach these issues, an automated lipreading system is required. In the direction to build a visual speech recognition system (VSRS) that can face these issues, many approaches were suggested and evaluated on various datasets, German Lipreading (Glips), Lip Reading in the [4], OuluVS2 [5], AVLetters [6], AVLetters2 [7], AVICAR [8], AVTIMIT [9], CUAVE [10], Grid [11], IBMIH [12], IBMSR [13], and XM2VTSDB [14], such as the work of Ha, Nicole Yah Yie, et al. [15], proposed a deep learning approach for visual speech recognition. Matthews et al. [6], advocated for the use of Active Shape Models (ASM) and Active Appearance Models (AAM), utilizing feature extraction and training a classification model through Hidden Markov Models (HMM). Their approach achieved an accuracy rate of 44.6 %. Hedayati-Dezfooli, M., et al. [16], developed an optimization method for propeller injection molding using soft computing, fuzzy evaluation, and the Taguchi method. Zhao et al [17], developed a robust predictive method for recognizing isolated sentences, exploiting the extraction of local spatio-temporal binary patterns in the oral region. This method, in conjunction

Table 2
Test functions for multimodal, unimodal, and multimodal with fixed dimensions.

Functions		Description	Dimensions	Range
Unimodal functions	F3	$f(x) = \sum_{i=1}^{d}\left(\sum_{j=1}^{i} x_j\right)^2$	30, 100, 500,1000	[-100,100]
	F4	$f(x) = \max_i(\lvert x_i\rvert,\ 1 \le i \le n)$	30, 100, 500,1000	[-100,100]
	F7	$f(x) = \sum_{i=1}^{n} ix_i^4 + random[0,1)$	30, 100, 500,1000	[-128,128]
Multimodal functions	F9	$f(x) = \sum_{i=1}^{n}\left[x_i^2 - 10\cos(2\pi x_i) + 10\right]$	30, 100, 500,1000	[-512,512]
	F13	$f(x) = 0.1\left(\begin{array}{l}\sin^2(3\pi x_1) + \sum_{i=1}^{n}(x_i-1)^2\left[1+\sin^2(3\pi x_1+1)\right]+\\ (x_n-1)^2\left(1+\sin^2(3\pi x_n)\right) + \sum_{i=1}^{n} u(x_i,5,100,4)\end{array}\right)$	30, 100, 500,1000	[-50,50]
Multimodal functions with a fixed dimension	F14	$f(x) = \left(\frac{1}{500} + \sum_{j=1}^{25}\left(j + \sum_{j=1}^{2}(x_j - a_y)^5\right)^{-1}\right)^{-1}$	2	[-65,65]
	F15	$f(x) = \sum_{i=1}^{11}\left[a_i - \frac{x_1\left(b_1^2 + b_1x_2\right)}{b_1^2 + b_1x_3 + x_4}\right]^2$	4	[-5,5]
	F19	$f(x) = \sum_{i=1}^{4} c_i exp\left(-\sum_{j=1}^{3} a_{ij}\left(x_i - p_{ij}\right)^2\right)$	3	[0,3]
	F23	$f(x) = -\sum_{i=1}^{10}\left[(X-a_i)(X-a_i)^T + c_i\right]^{-1}$	4	[0,10]

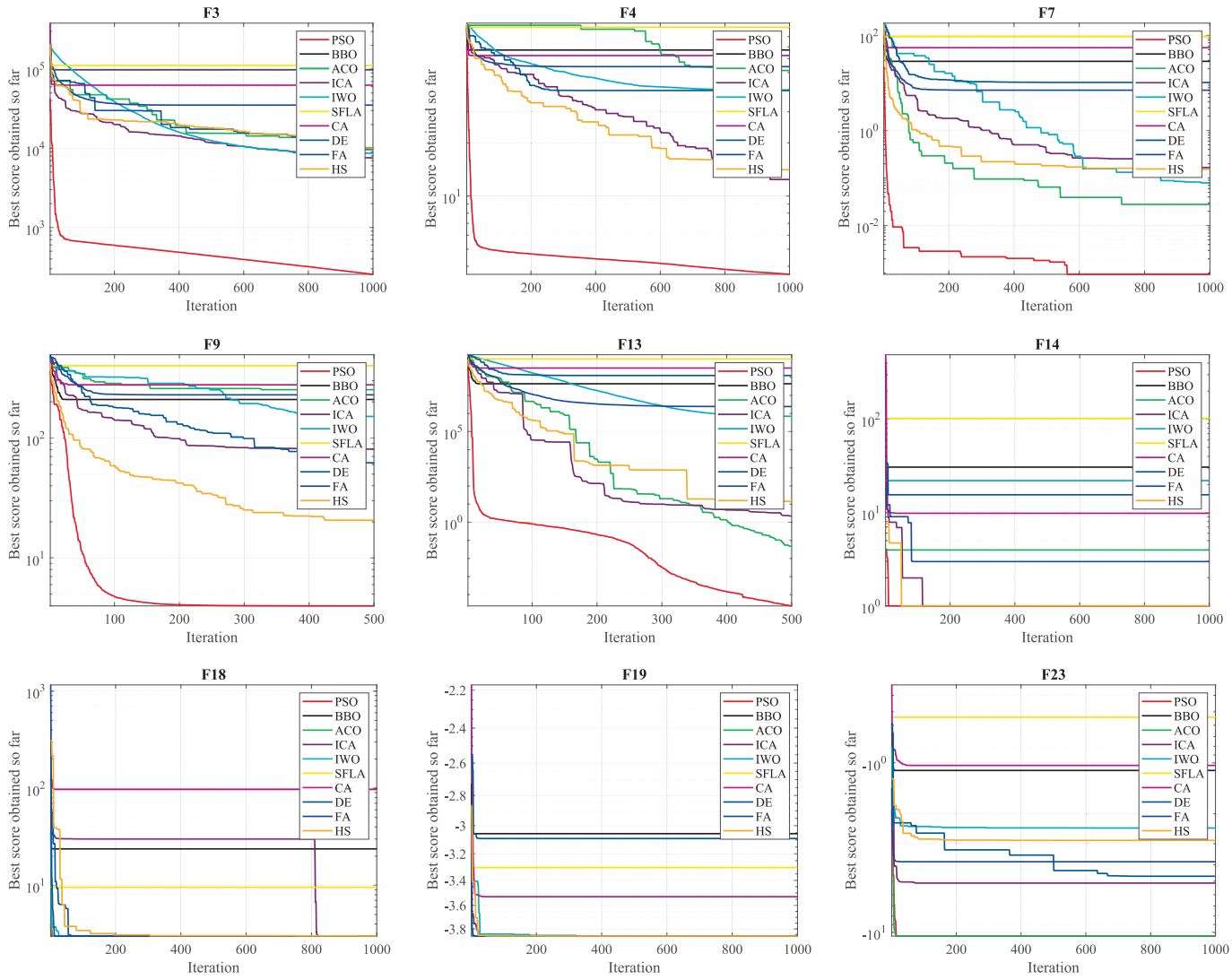


Fig. 2. Convergence curves of each algorithm for the studied problems F3-F4-F7-F9-F13-F14-F18-F19-F23.

with SVMs, attained an accuracy of 58.85 %. Petridis and Pantic [18] employed an autoencoder to extract bottleneck features from pixel data and subsequently trained an LSTM-RNN, attaining an accuracy of 58.1 %. Ultimately, Bakry and Elgammal [19] presented an MKPLS approach that assesses several kernels within their framework. This approach has been applied to the OuluVS and AVLetters datasets using distances between local binary patterns (LBP) and images to extract features. Tian and Weijun [20] proposed a method of Auxiliary Multimodal LSTM (am-LSTM) to fuse audio-visual data at the same time. To extract features from images and reduce their dimension, they used the pre-trained VGG-16 and PCA whitening, then extracted features of audio as a spectrogram with 10 ms overlap and a 20 ms Hamming window. The training was in video and audio, but they used the video only for the test. The performance of this method was 88.83 %. To recognize words from continuous speech, Chung and Zisserman [21] generated for the first time Lip Reading in the Wild (LRW). The VGG-M architecture was chosen as the foundation of their methodology due to its durability and classification efficacy. Employing a multi-turn (MT) architecture, they attained an accuracy of 61.1 %. The CFI-based CNN method, designed by Saitoh et al [22], introduces an innovative representation of sequential images combining spatio-temporal data for lip-reading. It achieved an accuracy of 59.3 % with AlexNet without DA and 87.5 % with GoogLeNet using DA, evaluated on the OuluVS2 corpus for a 90° profile view. The CNN using Hahn moments, developed by A. Mesbah et al [23]. By exploiting Hahn moments as filters in the CNN architecture, this method achieved an accuracy of 59.23 % on the AVLetters dataset. A new approach for automatic lip reading was introduced by Lu. Y and Li. H [24], in which they extracted keyframes from video s to locate the area of the mouth. In the subsequent phase, the VGG19 network was employed to extract characteristics from oral images, while an LSTM network assimilated temporal data. Recognition outcomes were produced utilizing a SoftMax layer and two fully connected layers. M. Kim et al. [25] developed the multi-head audio-visual memory (MVM) approach to address issues like as homophones and inadequate visual lip information. This integrates a singular memory for audio features with multiple key memories for visual features, with an accuracy of 88.5 % on the LRW corpus. Baaloul, Ali et al. [26] proposed a method for the visual detection of Arabic speech, integrating CNNs with vision transformers for lip-reading. This method attained an exceptional 98 % accuracy on their dataset, illustrating the efficacy of deep learning-based visual speech recognition (VSR) systems.

In this work, we introduce a novel approach called the Quaternion Charlier Moments Convolutional Neural Network, optimized using the PSO algorithm (QCMs-PSO-CNN) to implement a system capable of recognizing the phrase or word that a person says in a video (lip reading). This approach is based on quaternion algebra, Charlier moments, and the proposed PSO algorithm, which can extend the optimized quaternion Charlier moments to the color image in a holistic manner;

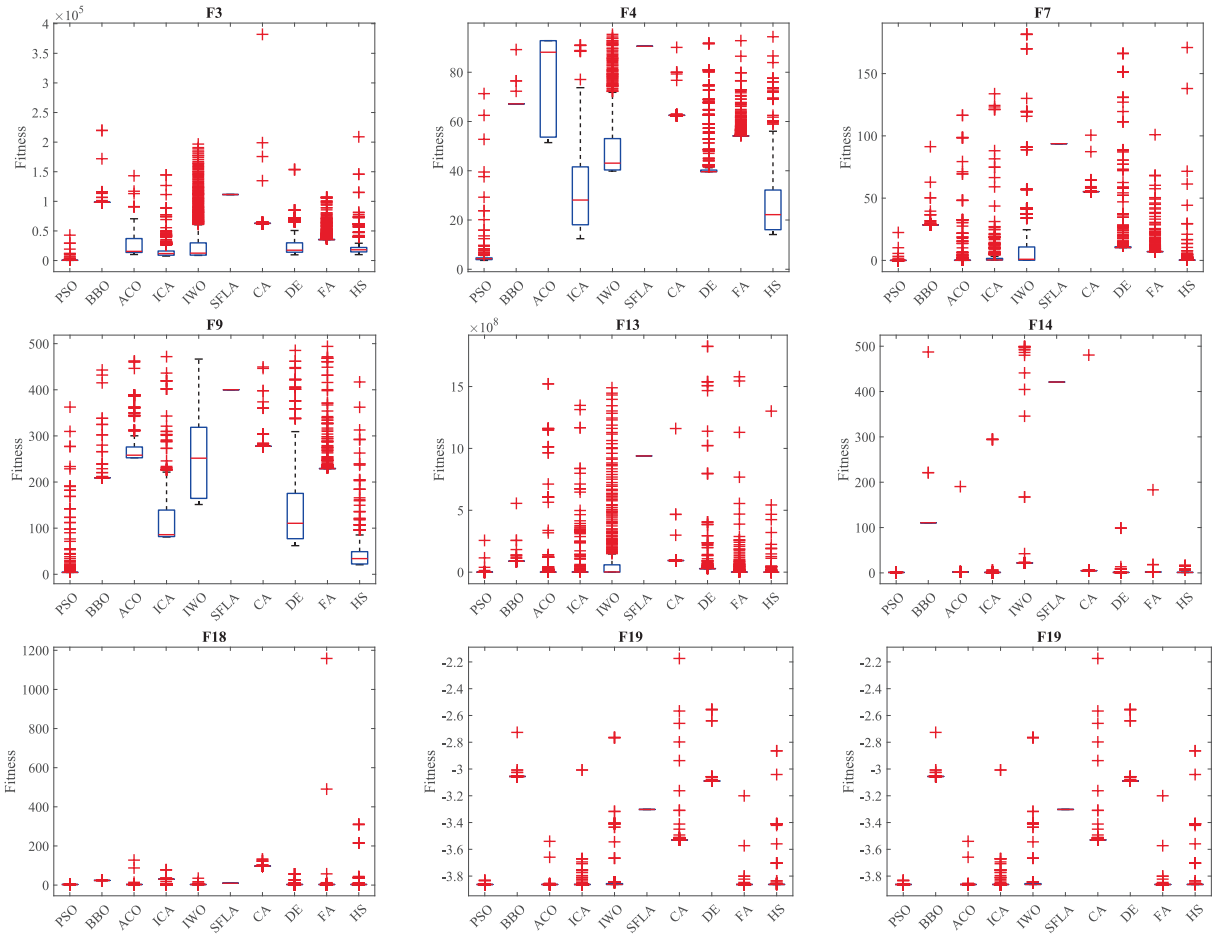


Fig. 3. Boxplots of each algorithm for the studied problems F3-F4-F7-F9-F13-F14-F18-F19-F23.

quaternion Charlier moments. In addition, the combination of CNNs and optimized quaternion Charlier moments, capable of identifying and extracting valuable information from images, are particularly effective in solving various challenges related to pattern learning and image classification. In our approach, as a first layer, we use the optimized quaternion Charlier moments to extract the useful features from color images, which we then supply to the CNN. As far as we know, this is the first time the optimized quaternion Charlier moments based on the developed algorithm PSO are employed as optimized descriptors in the architecture of CNNs applied to the intelligent recognition of visual speech. In this work, we explore an approach to the lipreading task that aims to solve several issues that result in the difficulties mentioned above. To summarize, the key contributions of this research.

are as follows:

(i) Introduce a new method called QCMs-PSO-CNN that can extract useful features from the input large-size color images by adding the optimized quaternion Charlier moments filter and consequently improve the performance of the CNN architecture.
(ii) A PSO algorithm is developed, based on the combination of linear inertia weights and a sine–cosine learning factor, which effectively strengthens the optimization capability of PSO.
(iii) A QCMs-based QCMs-PSO approach is constructed to extract local and global features of the lips images using optimized quaternion Charlier moments with the developed algorithm PSO.
(iv) Proposed PSO and 7 most advanced algorithms are carried out for the reconstruction and classification experiments of lips images, and evaluated by metrics such as fitness, MSE, PSNR, and accuracy, which indicate that this method demonstrates excellent recognition performance and stable adaptability under different dataset, and it exhibits a large potential in the field of visual speech.
(v) The deep learning method proposes a convolutional neural network (CNN) model, which excels at pattern recognition and classification using quaternion Charlier moments and the utilized PSO algorithm.
(vi) Give an effective and practical solution to address all variations of the VSR problem.

The paper is in the following organization: it presents some preliminary information about the quaternions and the Charlier moments in section 2. The optimized quaternion Charlier moments using the introduced particle swarm optimization algorithm are described in section 3. We also present the QCMs-PSO-CNN architecture and explain the procedure of its learning in section 4. The ablation experiments and results are given in section 5. Finally, concludes and prospective work are discussed.

2. Some preliminaries

This section outlines the recall of the discrete classical Charlier moments and quaternion representations of the color image.

2.1. Computation of the Charlier moments

Charlier polynomials of order n (CP) are characterized by a hypergeometric function, where n corresponds to the polynomial order and x is a variable taking its values in $[0, N-1]$. The mathematical

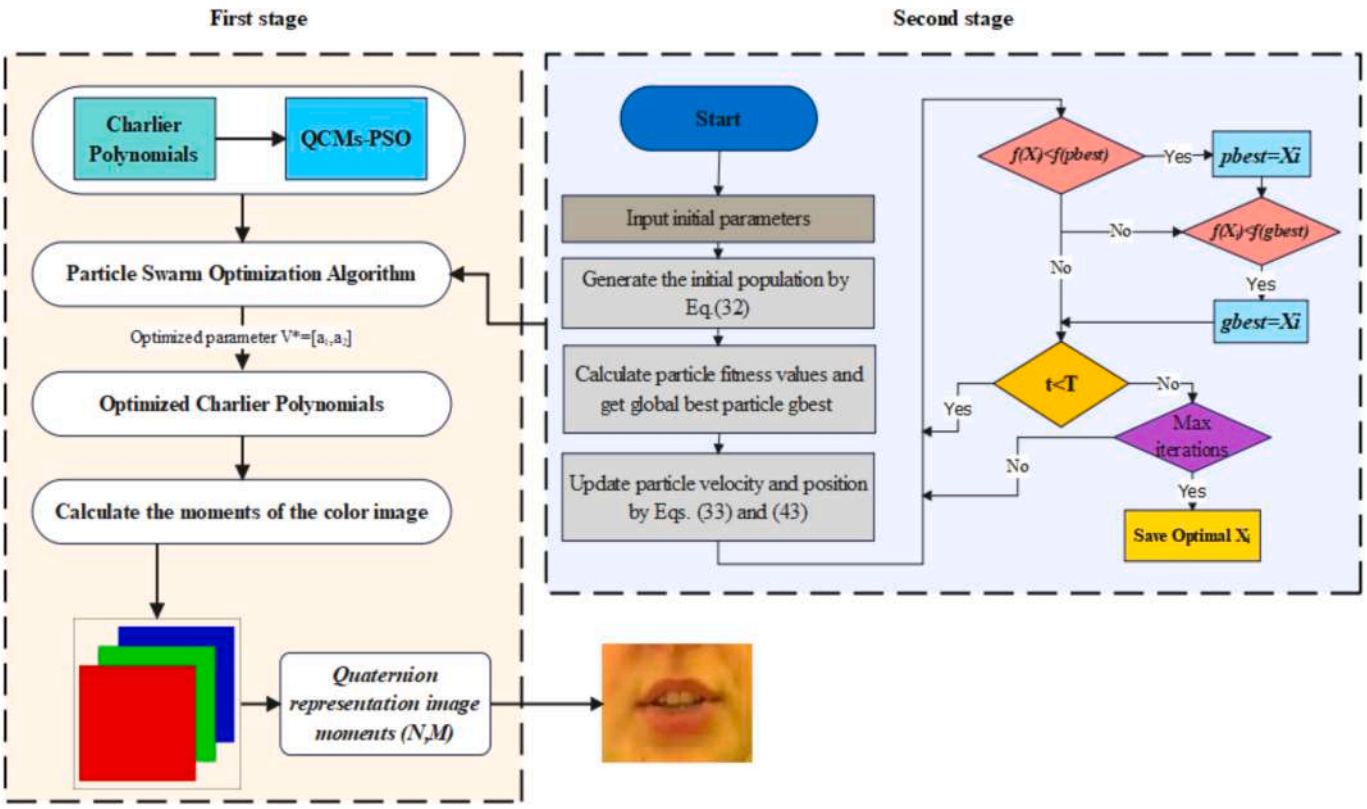


Fig. 4. Flow chart of QCMs-PSO.

formulation of these polynomials is as follows [27]:

$$CP_n^a(x) = {}_2F_0\left(-n, -x; -a^{-1}\right),\ x, n = 0, 1, 2.....\infty \tag{1}$$

where a is restricted to $a > 0$, and ${}_2F_0()$ is the generalized hypergeometric function.

The hypergeometric function of Charlier polynomials is defined as follows:

$${}_2F_0(a, b; c) = \sum_{n=0}^{m} \frac{(a)_n (b)_n (c)^n}{n!} \tag{2}$$

$(a)_n$ is the Pochhammer with $(a)_n = a(a+1)...(a+n-1)$ *if* n > 0

The orthogonality property satisfied by the Charlier polynomials is expressed as follows:

$$\sum_{x=0}^{\infty} \omega(x) CP_n^a(x) CP_n^a(x) = \rho(n)\delta_{nm}\ ;\ n, m \geq 0 \tag{3}$$

where $\rho(n)$ the squared norm and $\omega(x)$ weighting function of the CPs are as follows:

$$\rho(n) = \frac{n!}{a^n}\omega(x) = \frac{e^{-a}a^x}{x!} \tag{4}$$

The normalized form of the Charlier polynomials is given by [27]:

$$\widetilde{CP}_n^a(x) = CP_n^a(x)\sqrt{\frac{\omega(x)}{\rho(n)}} \tag{5}$$

Additionally, the Charlier polynomials can be expressed by the recurrence formula in order to accelerate squared norm and weight function's computation [28]. The recurrence relations are given by:

$$\widetilde{CP}_n(x) = -\sqrt{\frac{a}{n}}\ \frac{(x-n+1-a)}{a}\widetilde{CP}_{n-1}(x) - \sqrt{\frac{a^2}{n(n-1)}}\ \frac{(n-1)}{a}\widetilde{CP}_{n-2}(x) \tag{6}$$

with

$$\widetilde{CP}_0\ (x) = \sqrt{\frac{\omega(x)}{\rho(0)}} \text{ and } \widetilde{CP}_1(x) = \sqrt{\frac{\omega(x)}{\rho(1)}}\left(\frac{a-x}{a}\right) \tag{7}$$

The discrete Charlier moments of an image intensity function f(x, y) are given by:

$$CM_{nm} = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y) f(x, y)\ n, m = 0, 1, ...N-1 \tag{8}$$

We can also express the set of Charlier moments in matrix form as follows:

$$CM(f) = CP_1^T f\, CP_2 \tag{9}$$

while (.) T denotes the transposed matrix, and:

$$\begin{aligned} CM &= \left\{CM_{ij}\right\}_{i=0,j=0}^{i=N-1,j=N-1},\ CP_1 = \left\{\widetilde{CP}_i^a(x)\right\}_{i=0,x=0}^{i,x=N-1} \\ CP_2 &= \left\{\widetilde{CP}_j^a(x)\right\}_{j=0,y=0}^{j,y=N-1},\ \text{and } f = \left\{f(i,j)\right\}_{i,j=0}^{i,j=N-1} \end{aligned} \tag{10}$$

The reconstruction of the image using the Charlier moments is defined by the following formula:

$$f(x, y) = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y) CM_{nm} \tag{11}$$

Similarly, the image reconstruction using the matrix is defined as

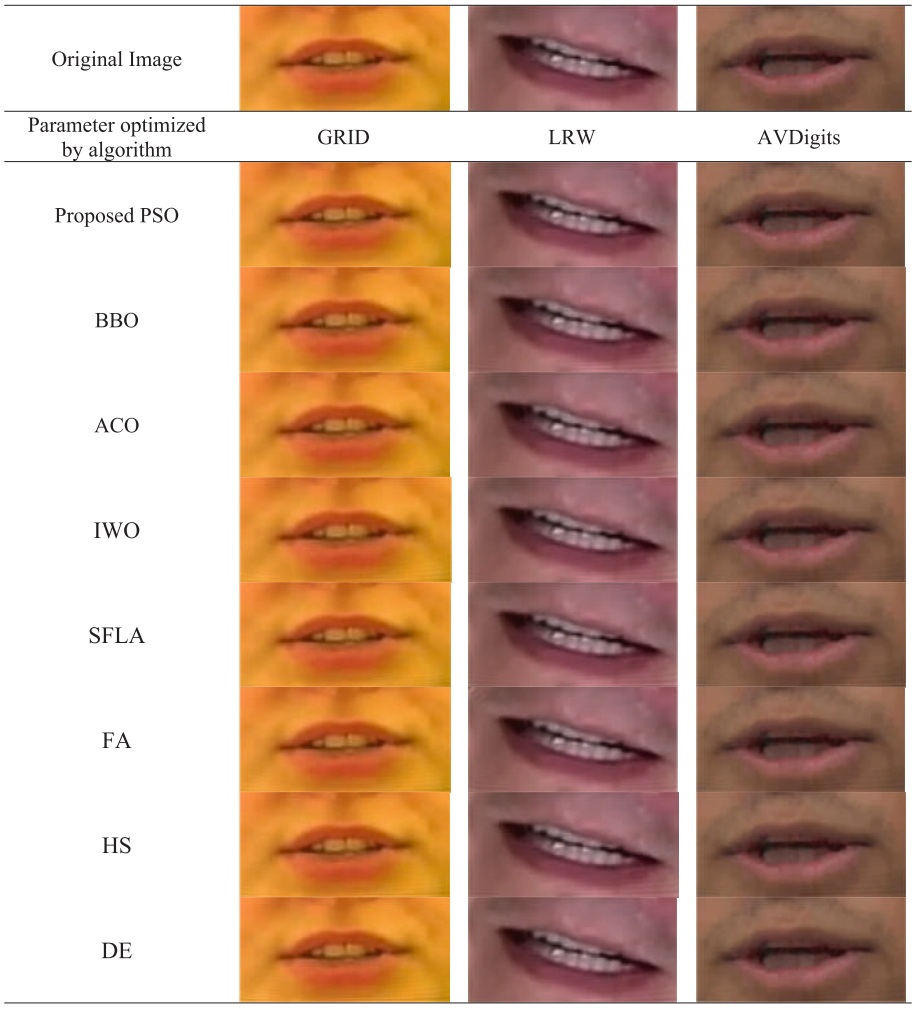


Fig. 5. Reconstructed frames using different *meta*-heuristic algorithms.

follows:

$$f = CP_2\, CM(f)\, CP_1^T \tag{12}$$

2.2. Quaternion representation of color images

Quaternions are defined as an extension of complex numbers, each of which consists of two parts: a real part and three imaginary parts [29].

$$q = q_0 + q_1\mathrm{i} + q_2\mathrm{j} + q_3\mathrm{k} \tag{13}$$

Where q_0,q_1, q_2, $q_3 \in \mathbf{R}$ are real numbers, ***i,j,k*** are imaginary units:

$$\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = \mathrm{ikk} = -1\mathrm{ij} = -\mathrm{ji} = \mathrm{kjk} = -\mathrm{kj} = \mathrm{iki} = -\mathrm{ik} = \mathrm{j} \tag{14}$$

The Following expressions are the definitions of a quaternion's conjugate and modulus, respectively.

$$q^* = q_0 - q_1\mathrm{i} - q_2\mathrm{j} - q_3\mathrm{k} \tag{15}$$

$$\|q\| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2} \tag{16}$$

At this time, the color image $f(x,y)$ is represented by a pure quaternion in the rectangular coordinate system as follows:

$$f(x,y) = f_R(x,y)\mathrm{i} + f_G(x,y)\mathrm{j} + f_B(x,y)\mathrm{k} \tag{17}$$

$f_R(x,y)$, $f_G(x,y)$, and $f_B(x,y)$ are the B, G, and R components of the color image [58].

3. The proposed optimization of QCMs using particle swarm optimization

3.1. Quaternion Charlier moments

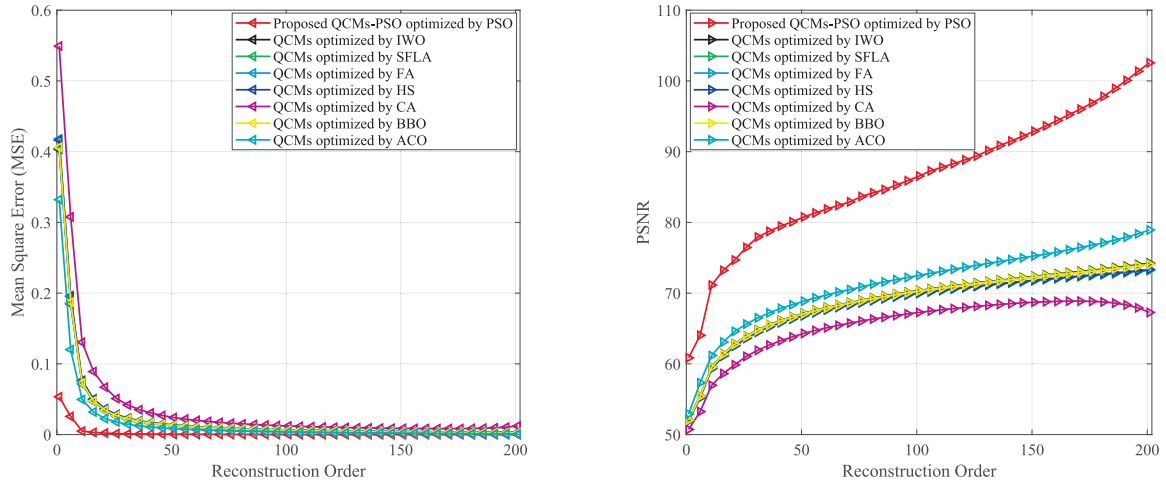
Let $f(x,y)$ represent an RGB image described in Cartesian coordinates. In light of the non-commutative characteristics of quaternion numbers, we define the right-side Quaternion Charlier Moments (QCMs) as follows:

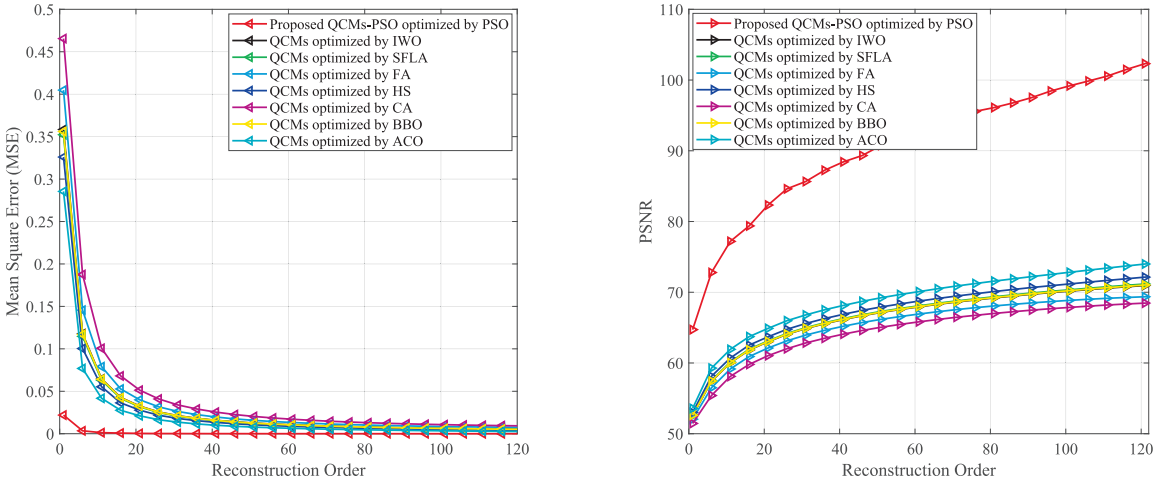
$$QCM_{nm}^R = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} f(x,y)\widetilde{CP}_n^a(x)\widetilde{CP}_m^a(x)\,\mu \tag{18}$$

(R is a reference to right-side quaternion) for $n = 0, \ldots N-1$, $m = 0, \ldots M-1$ and μ is a pure unit quaternion selected in this paper as $\mu = -(i+j+k)/\sqrt{3}$. $\widetilde{CP}_n^a(x)$ represents the nth order of Charlier 's discrete orthogonal polynomials. Additionally, the Eq. (18) can be explained by:

$$QCM_{nm}^R(f) = Q_0^R + Q_1^R i + Q_2^R j + Q_3^R k \tag{19}$$

where





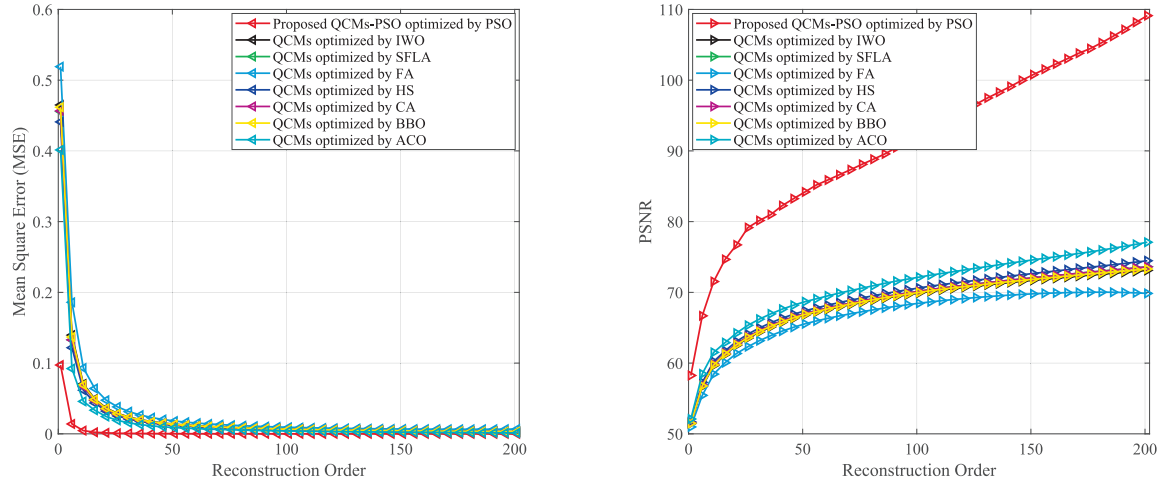


Fig. 6. The PSNR and MSE for two-dimensional QCMs using different *meta*-heuristic algorithms.

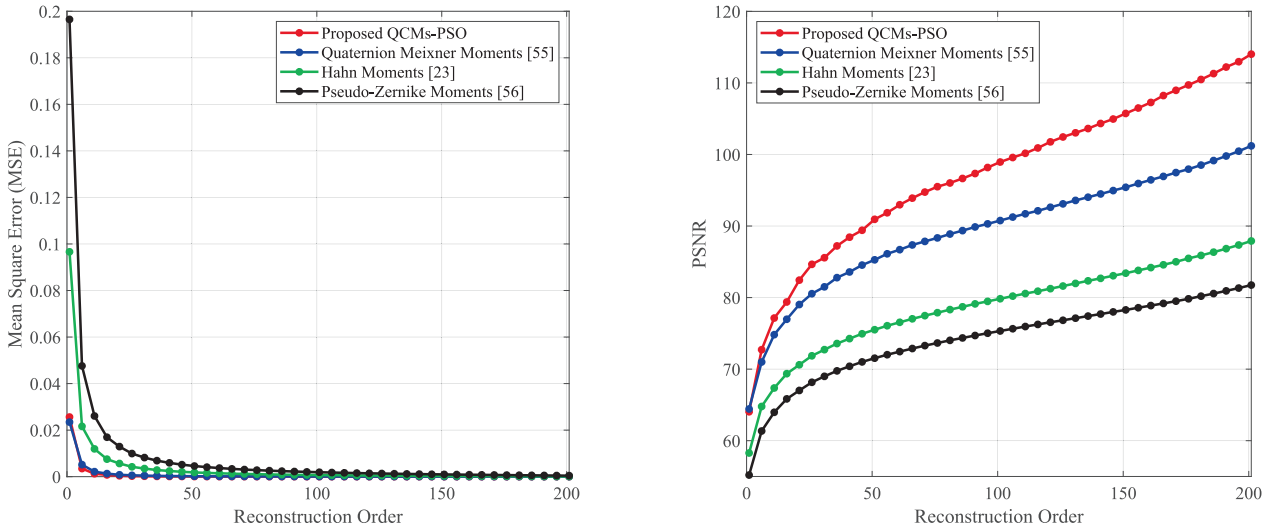


Fig. 7. (a) MSE, (b) PSNR of the proposed QCMs-PSO and the recent methods [23,53,54].

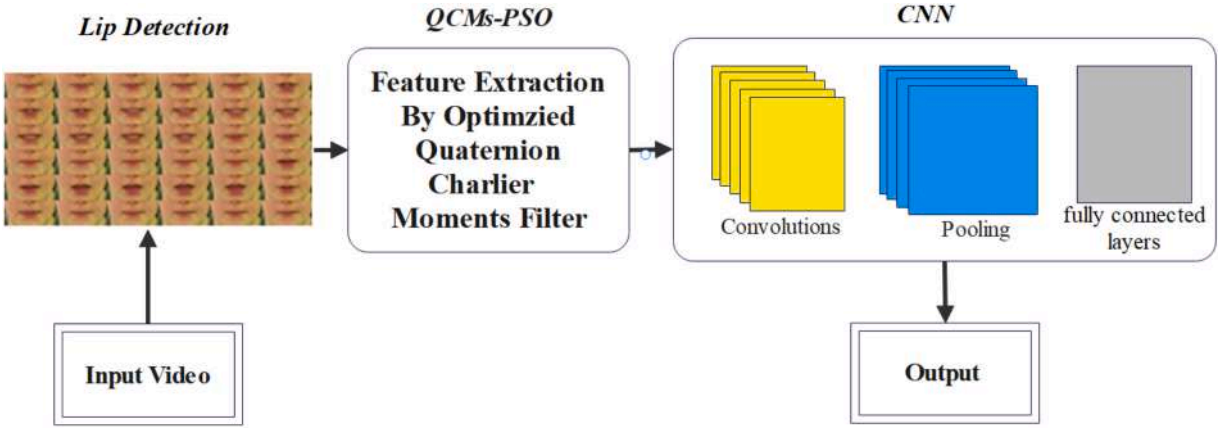


Fig. 8. The new QCM-PSO-CNN architecture for lip reading which takes an input video.

$$
\begin{aligned}
Q_0^R &= \frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_0 \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]\\
Q_1^R &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_1 \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]\\
Q_2^R &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_2 \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]\\
Q_3^R &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_3 \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]
\end{aligned}
\tag{20}
$$

with

$$
\begin{aligned}
W_0 &= f_R(x,y) + f_G(x,y) + f_B(x,y)\\
W_1 &= f_G(x,y) - f_B(x,y)\\
W_2 &= f_B(x,y) - f_R(x,y)\\
W_3 &= f_R(x,y) - f_G(x,y)
\end{aligned}
\tag{21}
$$

More explicitly, the coefficients Q_0^R, Q_1^R, Q_2^R and Q_3^R can be also expressed using the Eq. (8) by:

$$
\begin{aligned}
Q_0^R &= \frac{1}{\sqrt{3}}\left[CM_{nm}(f_R) + CM_{nm}(f_G) + CM_{nm}(f_B)\right]\\
Q_1^R &= -\frac{1}{\sqrt{3}}\left[CM_{nm}(f_G) - CM_{nm}(f_B)\right]\\
Q_2^R &= -\frac{1}{\sqrt{3}}\left[CM_{nm}(f_B) - CM_{nm}(f_R)\right]\\
Q_3^R &= -\frac{1}{\sqrt{3}}\left[CM_{nm}(f_R) - CM_{nm}(f_G)\right]
\end{aligned}
\tag{22}
$$

Owing to Charlier polynomials' orthogonality property, the inverse transformation of the right-side of QCMs can be computed easily by the following relation:

$$
\widehat{f}(x,y) = \sum_{n=0}^{\breve{N}-1}\sum_{m=0}^{\breve{M}-1} \widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\, QCM_{nm}^R(f)\mu \tag{23}
$$

where $0 \le \breve{N} \le N$, $0 \le \breve{M} \le M$. More explicitly, Eq. (23) can also given by the form:

$$
\widehat{f} = \widehat{f}_0 + \widehat{f}_1 i + \widehat{f}_2 j + \widehat{f}_3 k \tag{24}
$$

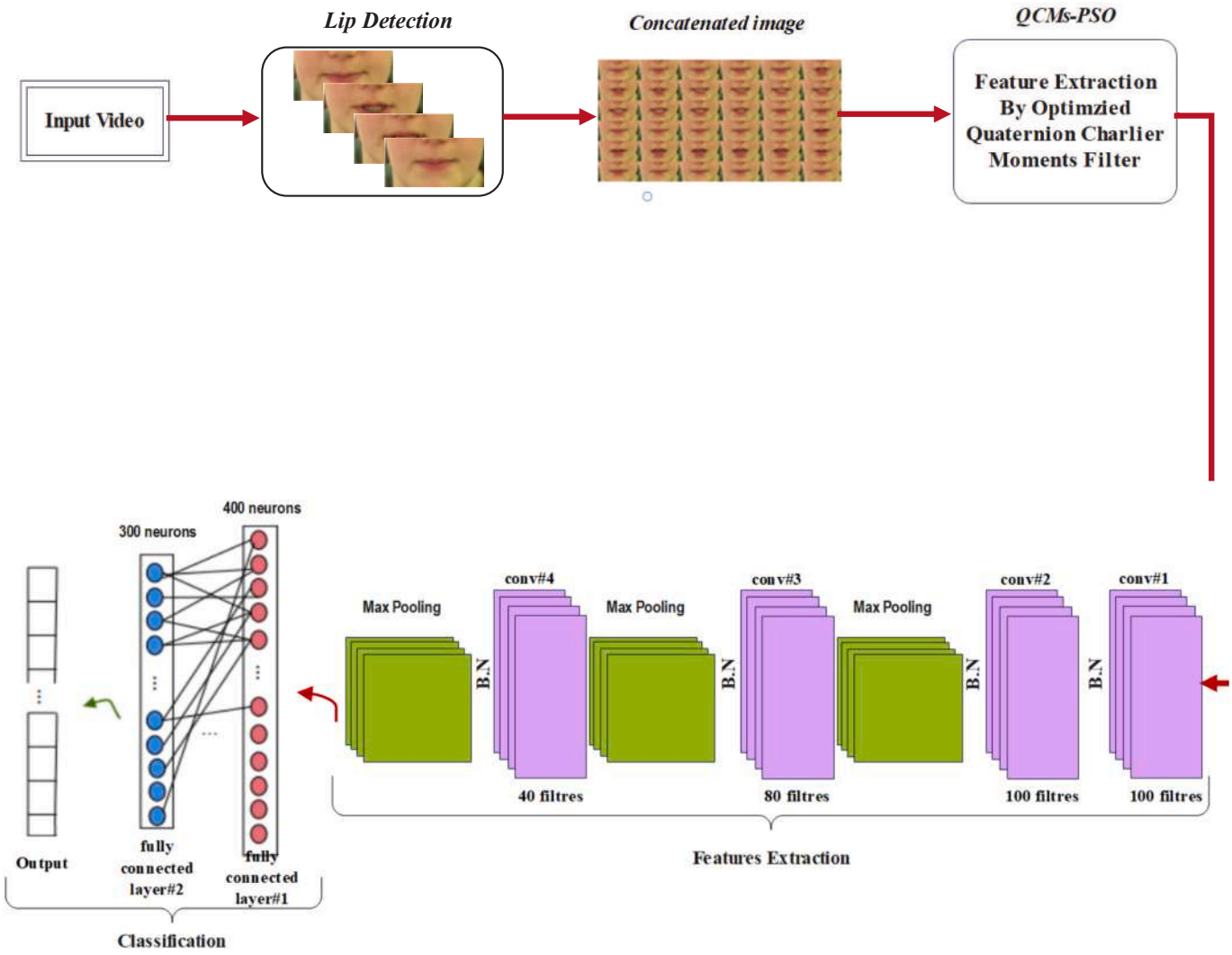


Fig. 9. QCMs-PSO-CNN model parameters: Charlier moments filter until the given order, convolution 1 (kernel and 100 filters), convolution 2 (kernel and 100 filters), max pooling 1 (pool size), convolution 3 (kernel and 80 filters), max pooling 2 (pool size), convolution 4 (kernel and 40 filters), max pooling 3 (pool size). Fully connected layer 1 (400 neurons), Fully connected layer 2 (300 neurons). Finally an output layer (26 classes).

Fig. 10. The GRID dataset.

Fig. 11. The LRW Dataset.

Fig. 12. Example of GLips cropped to 96 × 96 pixels.

Where

$$\begin{aligned}
\widehat{f}_0 &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{\breve{N}-1}\sum_{y=0}^{\breve{M}-1}(Q_1+Q_2+Q_3)\widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]\\
\widehat{f}_1 &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{\breve{N}-1}\sum_{y=0}^{\breve{M}-1}(Q_0+Q_2-Q_3)\widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]\\
\widehat{f}_2 &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{\breve{N}-1}\sum_{y=0}^{\breve{M}-1}(Q_0-Q_1+Q_3)\widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]\\
\widehat{f}_3 &= -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{\breve{N}-1}\sum_{y=0}^{\breve{M}-1}(Q_0+Q_1-Q_2)\widetilde{CP}_n^a(x)\widetilde{CP}_m^a(y)\right]
\end{aligned} \tag{25}$$

Representing color images $\widehat{f}$ using QCMs at order $\breve{N}\times\breve{M}$ can generate reconstruction errors, these are generally assessed using the PSNR, where a higher PSNR value signifying superior reconstruction quality, as indicated below:

$$PSNR = 20\cdot\log_{10}\left(\frac{255^2}{MSE}\right) \tag{26}$$

$$MSE = \frac{1}{\breve{N}\breve{M}}\sum_{x=0}^{\breve{N}-1}\sum_{y=0}^{\breve{M}-1}[f(x,y)-\widehat{f}(x,y)]^2 \tag{27}$$

where MSE denotes root mean square error, the image size is $\breve{N}\times\breve{M}$, with the pixel values of the original and reconstructed images represented respectively by $f(x,y)$ and $\widehat{f}(x,y)$

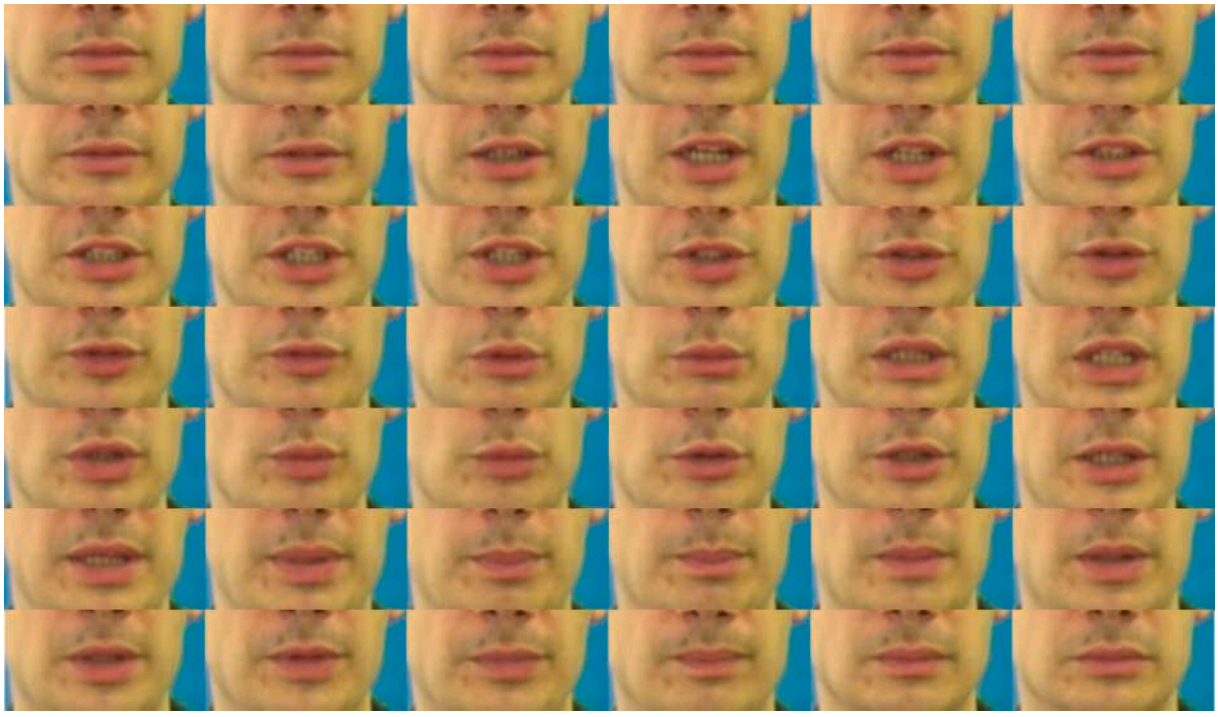


Fig. 13. CFI for the GRID dataset.

Fig. 14. CFI for the LRW dataset.

3.2. Particle swarm optimization

Particle Swarm Optimization (PSO), a metaheuristic method grounded in swarm intelligence (SI), is extensively employed to tackle continuous global optimization challenges. Drawing from the social dynamics of avian flocks and aquatic schools, as elucidated by Kennedy and Eberhart [30], Particle Swarm Optimization (PSO) functions through a population of entities termed particles, which collectively constitute a swarm. Each particle signifies a prospective solution to the optimization issue, characterized by a fitness (or objective) function. Throughout the evolutionary process, the quality of a solution is assessed according to the fitness value of each particle, determined by its present position within the search space. Within a problem's domain, each particle's position and velocity are initialized at random. Eq. (29) is used to determine the velocity of each particle i in the d^{th} dimension during the evolutionary process. This velocity is the product of the preceding velocity, the cognitive portion, and the social part. Three factors influence particle search behavior in PSO: current velocity $V_{i(t)}$, group best position *gbest*, and individual historical best position *pbest*. The PSO's comprehensive update procedure is:

$$\mathrm{X}_{i,j} = rand \times \left(u_j - l_j\right) + l_j, \;\; i,j = 1,2,\ldots,N,d \tag{28}$$

Fig. 15. CFI for the GLips dataset.

Table 3
Model's hyperparameters.

Hyperparameter	Property
Batch size	*Between 80 and 250*
Epochs	*100 for GLips*
	200 for GRID
	300 for LRW
Metric	*Area under curve (AUC)*
Loss function	*Categorical cross entropy*
Train-Validation Ratio	*80:20*
Optimizer	*Adam*
Learning Rate	*0.001*

$$V^d_{i(t+1)} = w\,V^d_{i(t)} + c1 \times rand^d_1 \times \left(pbest^d_i - X^d_{i(t)}\right) \times \left(gbest^d - X^d_{i(t)}\right) + c2 \times rand^d_2 \tag{29}$$

$$X^d_{i(t+1)} = V^d_{i(t+1)} + X^d_{i(t)} \tag{30}$$

where, d signifies the diameter of each particle, while $i \in [1, N]$ indicates the count of particles in the swarm at iteration t. The search domain for the j-th dimension is constrained by the maximum and minimum limits, u_j and l_j, respectively. The procedure includes an inertia weight w and acceleration coefficients $c1$ and $c2$, generally chosen from the interval [0,2].

Using Eq. (31), the value of w for each particle i was dynamically changed in each generation.

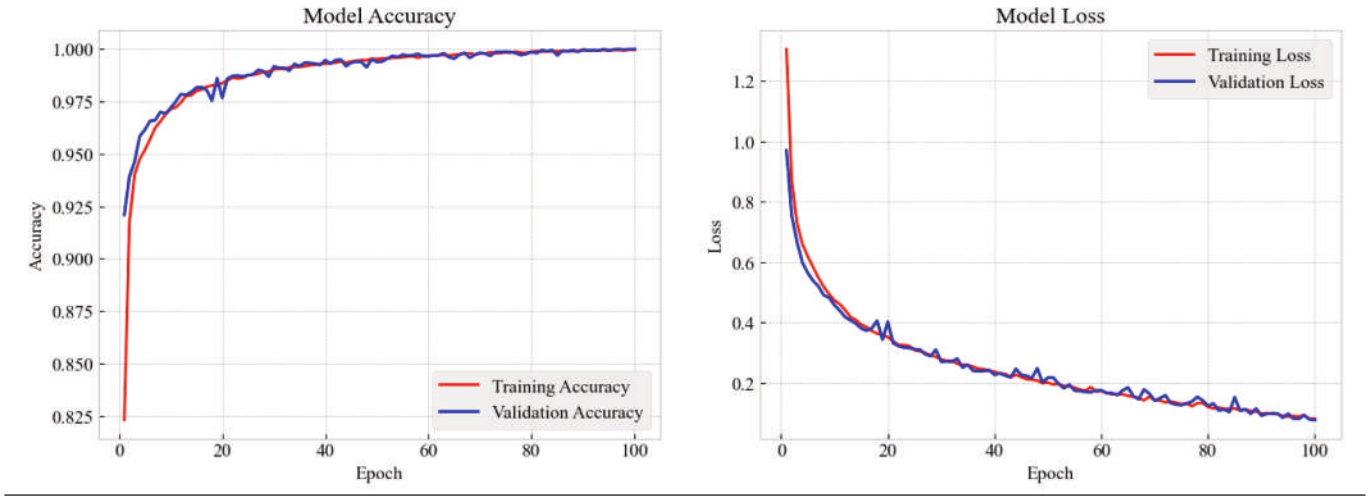
$$w_i = \frac{t^2\left[w_{fin} - \left(w_{fin} - w_{init}\right)\right]}{T^2} \tag{31}$$

The total iterations are represented by T, whereas t signifies the current iteration. The inertia weight w is initialized at w_{init} and progressively modified to w_{fin}, with typical initial and final values established at 0.9 and 0.4, respectively [31]. During the algorithm's preliminary phases, the population is extensively dispersed over the search space, as depicted in Fig. 1(a). In this phase, the adjusted parameter consistently diminishes at an accelerated rate, augmenting the algorithm's capacity for an exhaustive global search.

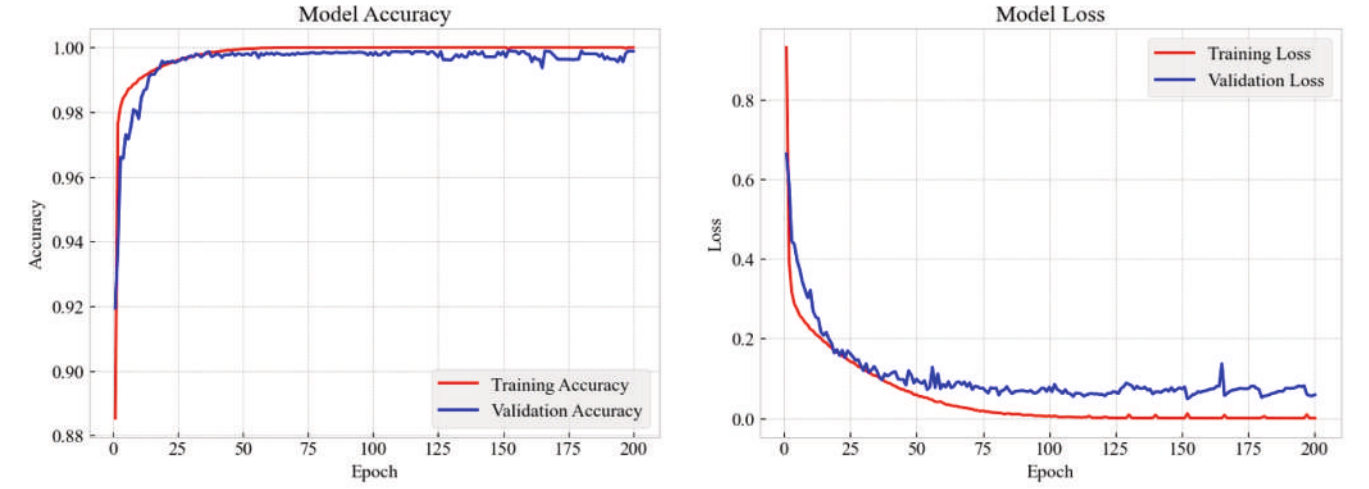
The original PSO uses 2 for both the social (c_1) and cognitive (c_2) components in order to counteract the effects of the stochastic acceleration coefficients. The particles wander excessively when the cognitive component (c_1) increases, which slows down the PSO's rate of convergence. The particle prematurely leaves the global optimum and misses the ideal solution with ease when the social component (c_2) increases. This part adjusts the cognitive component (c1) and social component

Table 4
The parameters of QCMs-PSO-CNN architecture.

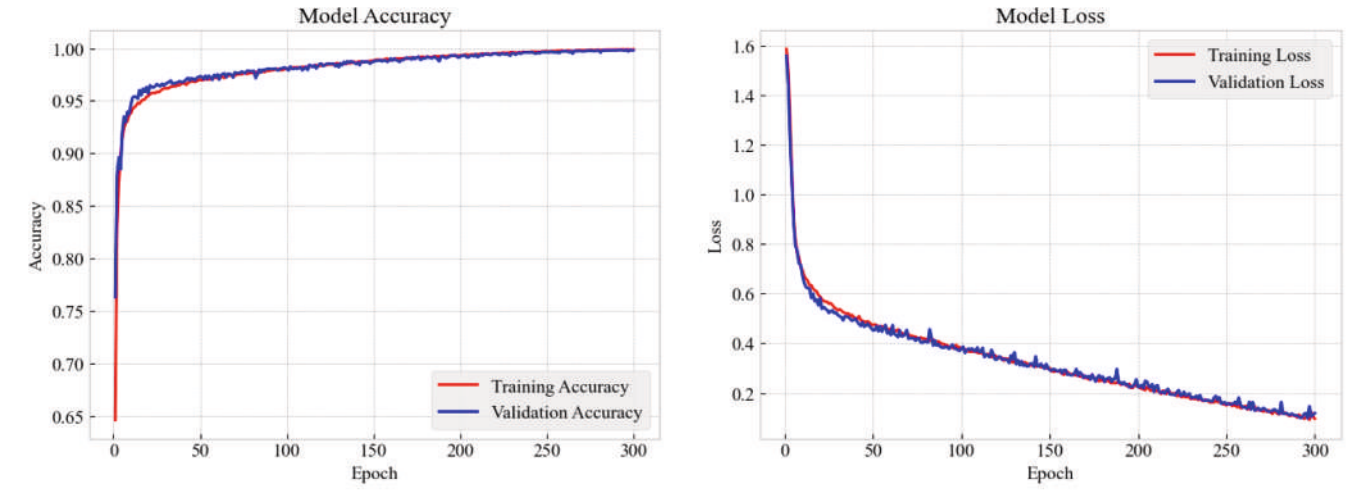
Layer number	Purpose	Filter size	Number of filters	Stride	Activation	Dropout Probability
–	Input image	–	–	–	$N \times M$	–
1	QCMs-PSO	–	–	–	$n \times m$	–
2	Conv + BN + RELU	**3 × 3**	**100**	**1**	$n \times m \times 100$	–
3	Conv + BN	**3 × 3**	**100**	**1**	$n \times m \times 100$	–
4	Maxpooling + ELU	**3 × 3**	–	**2**	$\frac{n}{2} \times \frac{m}{2} \times 40$	–
5	Conv + BN	**3 × 3**	**80**	**1**	$\frac{n}{2} \times \frac{m}{2} \times 80$	–
6	Maxpooling + ELU	**3 × 3**	–	**2**	$\frac{n}{2} \times \frac{m}{2} \times 40$	–
7	Conv + BN	**3 × 3**	**40**	**1**	$\frac{n}{2} \times \frac{m}{2} \times 40$	–
8	Maxpooling + ELU	**3 × 3**	–	**2**	$\frac{n}{2} \times \frac{m}{2} \times 40$	–
9	Fully Connected (RELU)	–	–	–	400	**0.55**
10	Fully Connected (RELU)	–	–	–	300	**0.55**
11	Softmax	–	–	–	class number	–



(a) Accuracy and Loss for GLips dataset



(b) Accuracy and Loss for GRID dataset



(c) Accuracy and Loss for LRW dataset

Fig. 16. Model Performance.

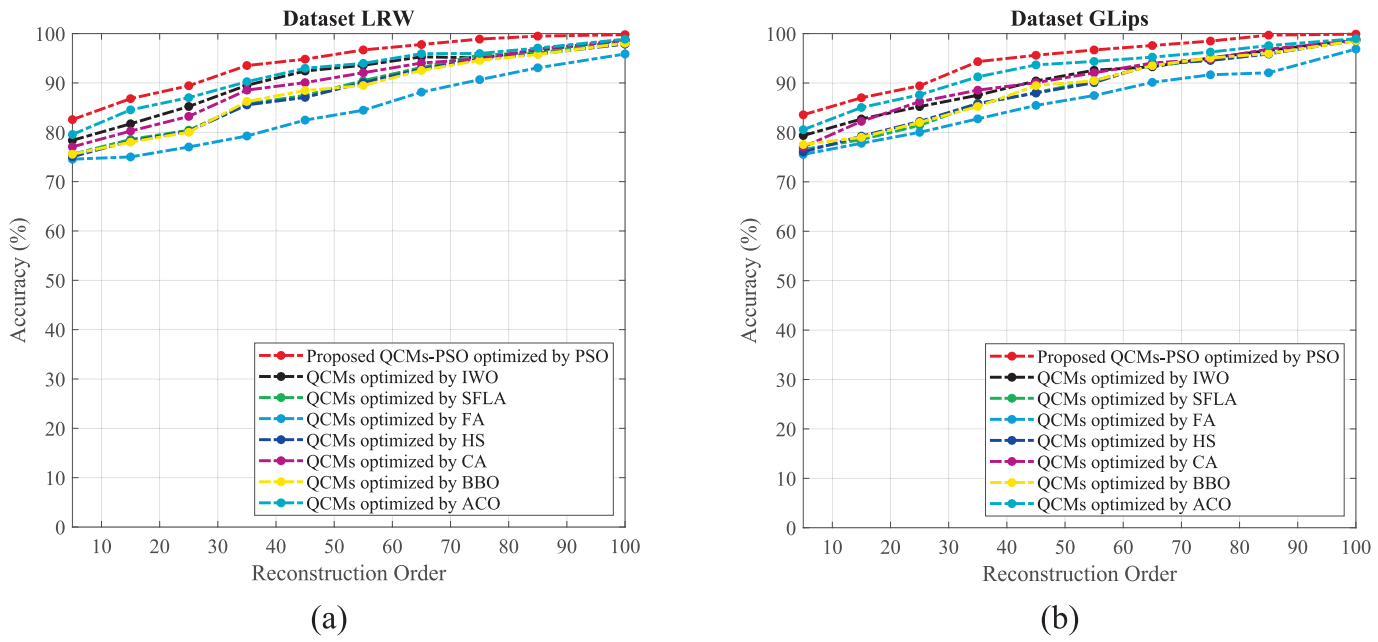


Fig. 17. The performance of QCMs evaluated through various *meta*-heuristic optimization algorithms: (a) LRW dataset, (b) GLips dataset.

Table 5
The results obtained for LRW [42] in comparison to other methods presented in the literature.

Method	Accuracy
Proposed QCMs-PSO-CNN (SI)	99.95 %
QCMs-CNN without PSO (SI)	91.42 %
HCNN [23]	59.23 %
CFI- based CNN [46]	52.47 %
LSTM-RNN [49]	59.18 %
ResNet LSTM [47]	58.93 %

Table 6
Results obtained for the GRID [41] dataset using the SI protocol, compared to alternative methods.

Method	Accuracy
Proposed QCMs-PSO-CNN (SI)	99.89 %
QCMs-CNN without PSO (SI)	92.28 %
LSTM + NNs (SI) [49]	76.61 %
CNN (SI) [52]	58.48 %
LipNet [48]	95.21 %

Table 7
Results obtained for the GLips [44] dataset using the SI protocol with DA, compared to other methods.

Method	Accuracy
Proposed QCMs-PSO-CNN (SI)	99.97 %
QCMs-CNN without PSO (SI)	94.37 %
And alm-GRU + DA [51]	85.53 %
CNN + DA [52]	57.96 %
RTMRBM [50]	71.77 %

(c_2) using a sine–cosine learning factor in order to get around this restriction. c_1 and c_2, as modified, are:

$$c_1 = 2{\cdot}\sin\left(\frac{(T-t).\pi}{2T}\right), \tag{32}$$

$$c_2 = 2{\cdot}\cos\left[\left(\frac{(T-t).\pi}{2T}\right)\right]. \tag{33}$$

As can be seen in Fig. 1(b), $c_1, c_2 \in [0,2]$, c_1 monotonically lowers and c_2 monotonically increases, both of which display the other's changing state. The quality of the solution in the exploration stage is enhanced, the range of particles wandering in the later stage is decreased, and the PSO's convergence is accelerated by the sine–cosine learning factor.

Moreover, to prevent each dimension of all particles to move beyond the search space, in the original PSO, the velocity of particles were clamped to the maximum velocity value V^d_{max}. Additionally, in the original PSO, the particle velocity was clamped to the maximum velocity value V^d_{max} in order to prevent any dimension of the particles from moving outside of the search space. Using equation (34), the user defines this maximum velocity parameter to clamp the excessive accelerations in a swarm.

$$\begin{cases} V^d_{i(t+1)} = V^d_{max} & if \quad V^d_{i(t+1)} \succ V^d_{max} \\ V^d_{i(t+1)} = -V^d_{max} & if \quad V^d_{i(t+1)} \prec -V^d_{max} \end{cases} \tag{34}$$

Every particle's performance has been assessed based on the objective function $f(\mathrm{X}_i)$. If X's fitness during the evolution process exceeds that of $pbest_i$, $pbest_i$ will be updated using Eq. (35) and $gbest^d$ will be updated using Eq. (36). Until the termination condition is met or the global optimum solution is discovered, these steps are repeated.

$$pbest^d_i(t) = \begin{cases} pbest^d_{i(t-1)} & if \quad f\left(\mathrm{X}^d_{i(t)}\right) \geq f\left(pbest^d_{i(t-1)}\right) \\ \mathrm{X}^d_{i(t)} & if \quad f\left(\mathrm{X}^d_{i(t)}\right) \prec f\left(pbest^d_{i(t-1)}\right) \end{cases} \tag{35}$$

$$gbest^d = \arg min\left(pbest^d_i(t)\right) \tag{36}$$

Algorithm 1. *provides the pseudo-code for the PSO, and Fig. 4 (Second stage) also displays the PSO algorithm flow chart, which helps to clarify the process of the suggested PSO algorithm in more detail.*

Algorithm 1:Optimization of QCMs using Particle Swarm Optimization (PSO)

Input: Intensity *f(x,y)* of image, order of moment N_{max}, Size of image N × M.

(*continued on next page*)

Table 8
Ablation study results for the proposed QCMs-PSO-CNN architecture.

Ref/Year	Approach	Dataset	Precision	Recall	Accuracy	F1-Score
Shashidhar, R., et al. [56] (2022)	DNN-LSTM	VSR	91.00 %	90.00 %	90.00 %	91.00 %
Ramadan, R. A., et al. [57] (2022)	DCNN	LRW	88.10 %	89.30 %	95.60 %	96.00 %
		GRID				
Alameen, S. A., et al. [59] (2023)	3DCNN-LSTM	YawDD	96.00 %	98.00 %	96.00 %	97.00 %
		3MDAD	90.00 %	92.00 %	91.00 %	91.00 %
Rajab, M. A., et al. [58] (2023)	CNN AlexNet	VSR	,89.70 %	89.80 %	90.00 %	88.70 %
Ameer, A. W. A., et al. [60] (2024)	CNN-GRU	GLips	_-	–	48.40 %	48.50 %
Li, Y., et al. [61]	3D CNNs + DC-TCN + CFS-DCTCN	LRW	99.50 %	99.50 %	99.50 %	99.50 %
(2024)		GRID	98.80 %	98.80 %	98.80 %	98.80 %
Vekkot, S., et al. [62] (2025)	SVM and Random Forest	MIRACL-VC1	74.00 %	75.00 %	75.00 %	74.00 %
Wang, X., et al. [64] (2025)	DBNFTLSTM	LRW	98.70 %	99.10 %	99.30 %	98.80 %
		GRID				
Baloch, A., et al. [63] (2025)	CNN-LSTM	ULRD	91.00 %	91.00 %	90.00 %	90.00 %
Our Proposed QCMs-PSO-CNN Architecture		GRID	98.91 %	99.74 %	99.89 %	99.32 %
		LRW	98.92 %	98.98 %	99.95 %	98.95 %
		GLips	99.28 %	98.77 %	99.97 %	99.02 %

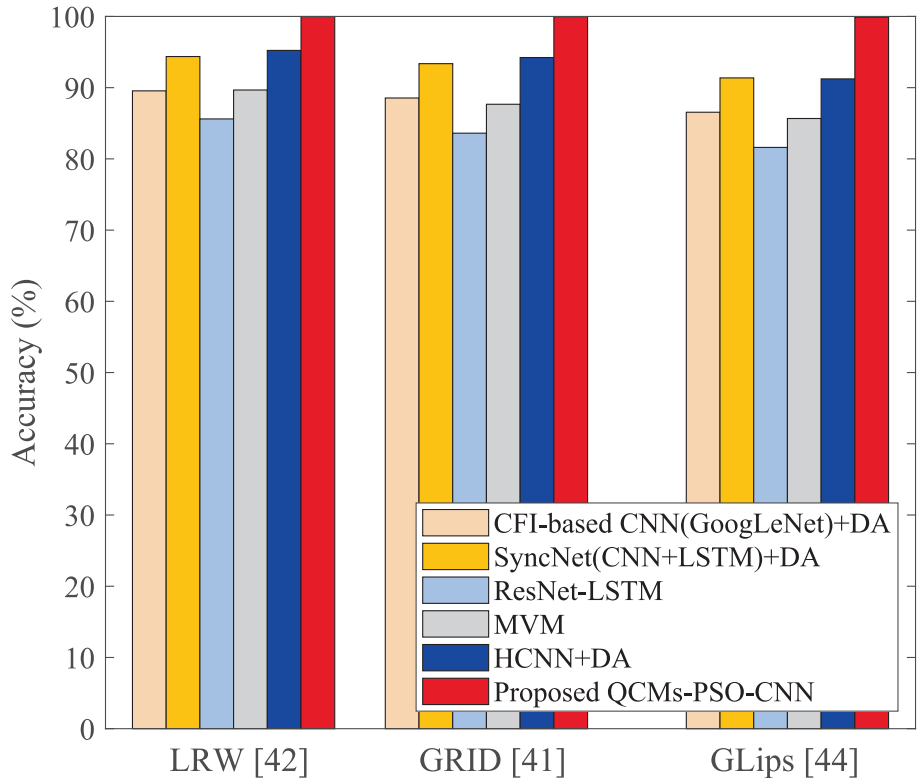


Fig. 18. Obtained results on different datasets of the new QCMs-PSO-CNN architecture in comparison with other methods.

(*continued*)

Algorithm 1:Optimization of QCMs using Particle Swarm Optimization (PSO)

Output: Optimized parameters for QCMs $V^* = [a_1, a_2]$.
Data: nPop = 30 (Population size), $T = 1000$ (Maximum number of iterations), $D = 4$ (Problem dimension), $VU = [N,N]$ (Maximum values of parameter), $VL = [0,0]$ (Minimum values of parameter).
Initialization of parameters
$w_{init} = 0.4$(initial Inertia Weight) and $w_{fin} = 0.9$ (final Inertia Weight)
Evaluation of the fitness function MSE provided by Eq. (27) for each particle swarm
Randomly generate the velocity and position of the initial population using Eqs. (28) and (29).
Calculate the fitness of all particles to get the global best particle *gbest*.
while $(t < T)$ **do**
Generate the linear inertia weight using Eq. (31)

(*continued on next column*)

(*continued*)

Algorithm 1:Optimization of QCMs using Particle Swarm Optimization (PSO)

Generate sine–cosine learning factors c_1 and c_2 using Eqs. (32) and (33)
for $i = 1{:}N$ **do**
Find the personal best using Eq. (35)
Update global best using Eq. (36)
end for
Calculate the fitness for each particle
for $i = 1{:}N$ **do**
if $f(X_i^{t+1}) < f(pbest)$ **then**
$pbest = X_i^{t+1}$
if $f(X_i^{t+1}) < f(gbest)$ **then**
$gbest = X_i^{t+1}$

(*continued on next page*)

(*continued*)

Algorithm 1:Optimization of QCMs using Particle Swarm Optimization (PSO)

end if
end if
end for
$t = t + 1$
end while
Return position gbest and fitness f of the best particle.
Until a termination criterion is satisfied (maximum of iteration)$V^* = [a_1, a_2]$

We compared the convergence performance of PSO algorithm with different types of comparison algorithms BBO [32], ACO [33], ICA [34], IWO [35], SFLA [36], CA [37], DE [38], FA [39], and HS [40]. Table 1 displays the initial parameters of the algorithms mentioned above. The developed algorithm is applied to 10 benchmark functions. The first three (F3, F4, F7) are unimodal benchmark functions, the next two (F9, F13) are variable-dimensional multimodal benchmark functions, and the remaining functions (F14, F18, F19, F23) are fixed-dimensional multimodal benchmark functions. Table 2 presents these functions where along with the function expression the dimension, range, and absolute minima of all the functions are mentioned.

The convergence curves for each algorithm are shown in Fig. 2 so that the convergence and stability of the suggested PSO may be examined. The addition of the sine–cosine learning factor speeds the PSO's convergence, as the figure demonstrates that PSO possesses faster convergence on the great majority of the evaluated functions. Fig. 3 displays the boxplots of each algorithm in comparison to the other algorithms. PSO's box shape is lower and narrower, indicating that the strategies' synergistic effect stabilises the results of each run close to the theoretical optimum, allowing for higher accuracy results. Therefore, PSO can thus offer a better and more reliable solution for the high-dimensional problems.

3.3. The proposed QCMs-PSO

The metaheuristic algorithm, considered as an iterative stochastic process that converges toward the global optimum of the objective function (MSE of the QCMs), together with the computation of quaternion Charlier moments to represent global properties in pattern learning and color image classification, constitute the two main strategies forming the QCMs-PSO introduced in this paper. For a more thorough explanation of QCMs-PSO, the pseudo-code of the PSO is provided by Algorithm 1, and Fig. 4 displays the QCMs-PSO flowchart.

We evaluated the efficacy of our approach by comparing it with several *meta*-heuristic algorithms documented in the literature, specifically in the context of color image reconstruction. For this assessment, we chose three photos measuring 150×300, 60×120, and 120×256 from the GRID [41], LRW [42], and AVDigits [43] datasets, respectively. Fig. 5 illustrates that our approach surpasses BBO [32], ACO [33], IWO [35], SFLA [36], CA [37], FA [39], and HS [40], in the selection of optimal parameter values for color image representation. The PSO algorithm effectively determines the optimal parameter α values for Charlier polynomials and attains an accurate depiction of images via optimized QCMs, essential for exact image classification. Additionally, The PSNR, and MSE for three distinct frames from our chosen datasets are displayed in Fig. 6. The efficacy of our optimization process is demonstrated in the MSE curves, especially for lower orders, where it outperforms alternative methods. The PSNR curves further underscore this advantage, particularly at elevated orders, illustrating the strong performance of Algorithm 1.

In this experiment, we compare the effectiveness of lipreading feature extraction using the proposed quaternion Charlier moments optimized by the PSO algorithm and applied to images from the LRW dataset, with that of approaches described in the literature [23,53,54]. Fig. 7 illustrates the difference in the curves produced by the MSE and PSNR of the proposed QCMs-PSO and the different existing methods, namely Quaternion Meixner Moments (QMMs) [53], Pseudo-Zernike Moments (PZMs) [54], and Hahn Moments (HMs) [23]. The results consistently demonstrate that the proposed QCMs-PSO method achieves high efficiency and clear superiority in lipreading feature extraction.

4. Proposed QCMs-PSO-CNN architecture

In this section, we introduce the new QCMs-PSO-CNN architecture, shown in Fig. 8, which aims to address VSR challenges by using optimized QCMs as descriptors. This shallow architecture can rapidly recognize the lips images due to the capability of optimized Charlier moments to extract features efficiently. It is designed to decrease the high computational costs and minimize the number of parameters required by CNN. Indeed, QCMs-PSO-CNN improves the quality of feature extraction and pattern assimilation in the color images. Additionally, the optimized QCMs are robust descriptors to extract the most useful information from the color images, even when dealing with large-sized images. In addition, the Charlier polynomials depend on parameter a_1,a_2 that give a wide choice in the reconstruction, so the optimized QCMs can completely cover the color images with the possibility of holding their global features. The description of parameters and the model employed in this work are presented as shown in Fig. 9. This architecture is split into two principal phases: the filter of QCMs-PSO and the CNN architecture.

Optimized Quaternion Charlier moments filter: The main role of this layer is to calculate the coefficients Q_0^R, Q_1^R, Q_2^R and Q_3^R of the input color image, then Eq. (19) is used to represent them by quaternion representations, which allows yielding a quaternion matrix noted QCMs-PSO of size depending on the moment's order value. Consequently, this layer provides an optimal representation of the color image and reduces significantly the processing's dimensionality.

Convolutional Neural Network: It provides a powerful classification due to its ability to extract high-level features. It is also divided into two principal phases: The first is to take the optimized QCMs matrix rather than the input image, then apply the different optimization functions and convolutional filters. The second one is the fully connected layer, and it is used for classification by applying several operations such as normalization, activation functions, and dropout.

5. Experiments and results

In this section, we present the results of the new QCMs-PSO-CNN architecture to validate the classification performance obtained. We first introduce some datasets, the model performance and training parameters. The experimental results are then displayed and enumerated.

5.1. Databases

In this subsection, three distinct datasets that have been chosen from the literature are introduced. They are GRID, LRW, and German Lipreading (Glips).

5.1.1. GRID

The GRID corpus dataset includes 34 speakers, divided into 18 males and 16 females, with each speaker uttering sentences 1000 times, resulting in a total of 34,000 video s. Each movie has a duration of three seconds and is captured at a frame rate of 25 frames per second. The dataset offers video s in two formats: standard quality (360×288) and high quality (720×567). Fig. 10 illustrates an example image from the GRID [41] dataset.

5.1.2. Lip reading in the Wild (LRW) dataset from Oxford and BBC

The LRW (Lip Reading in the Wild) collection has 1,000 utterances that include 500 unique words, articulated by numerous speakers from the BBC television channel. Each video in the dataset has a duration of

1.2 s, comprising 29 frames, with the target phrase located at the middle of the video. Fig. 11 illustrates an example of photos from the LRW [42] dataset.

5.1.3. German lipreading (Glips)

The GLips dataset is made up of 250,000 video s of Hessian Parliament speakers' faces, each of which is segmented into 500 distinct words with 500 instances. The format is akin to the English-language Lip Reading in the Wild dataset, wherein one word of interest is video encoded inside a 1.16-second length context. An extra metadata text file including the fields spoken word, start time in seconds, finish time in seconds, and duration in seconds is present for every video. Fig. 12 is shown an example of GLips [44].

5.2. Evaluation metrics

To assess how well the model learns, we evaluate its effectiveness using a wide range of metrics, including F1-score, accuracy, loss, precision, and recall [55]. These metrics provide specific insights into the strengths and weaknesses of the models. The following lines provide a mathematical representation of these metrics:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{37}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{38}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{40}$$

$$F1 - Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{41}$$

where TP (True Positive): Positive cases that have been properly classified, TN (True Negative): Negative cases that were properly classified, False Positive (FP): Positive instances that were wrongly categorized, False Negative (FN): Negative cases that were wrongly classified.

5.3. Preprocessing and data augmentation

5.3.1. Preprocessing

For preparing the chosen datasets, first, we begin by retrieving frames from all video s in the GRID [41], LRW [42], and Glips datasets, then we extract the ROI (lip region) in each frame corresponding only to the spoken word in all given video s. The lip regions are resized to the same pixels in all datasets. The number of frames depends on the video length in each dataset because the utterance speed is obviously different and on each speaker's content. Consequently, their manipulation is inappropriate. In the case of Glips [44] and GRID [41] datasets, to treat this situation, we proceed to take frames corresponding to the video that has the maximal number of frames nf_{max}, then we replicate the first frame until we get $\gamma = \sqrt{nf_{max} - nf_r} \in \mathbb{N}$, where nfr is the minimal number of replicated frames that we have to add to nfmax to obtain $\gamma \in \mathbb{N}$. Finally, we replicate the first frame in each video to unify the numbers of frames with the number γ^2. We chose the first frame to replicate because it is just a repetition that will not affect or add any new information to the content of the given videos. With regard to the Lip Reading in the Wild dataset, we begin by extracting the frames from the provided movies. We next utilise the meta data to segment the video and preserve just the frames that correlate to the spoken phrase. finally we crop the region of each speaker's mouth (ROI).

Finally, we adopt the CFI approach, introduced by Saitoh et al [22], to represent all the frames of a video sequence in a single image. This simple but effective method captures the spatio-temporal information of a complete sequence. In both the Glips [44] and GRID [41] datasets, the concatenated image is formed by merging γ^2 frames and organizing them into γ rows and γ columns for each movie, as depicted in Figs. 13, 14, and 15.

5.3.2. Data augmentation

In practice, sequence learning is easy to overfit when the samples of VSR are not sufficient, so we need to utilize a set of data augmentation strategies for reducing overfitting on training datasets. Additionally, Fair comparisons with other literary works require data augmentation. We used visual data augmentation in each extracted frame of each video by applying small-angle rotation, horizontal flipping, and noise injection. Moreover, the augmentation will also improve the generalization of space diversity and the difference in image quality. Consequently, it will improve the generalization of the trained neural network Krizhevsky et al. [45]. Applying many operations increases the processing and training costs since those datasets are very large, especially the Glips dataset [44].

5.4. Model performance and training parameters

This section details the parameters configured for training our model. Upon finalizing the data augmentation and preprocessing procedures previously outlined, we partitioned the dataset into two subsets: training and testing. We employed a particular splitting approach for the LRW dataset [42]. We chose five spoken utterances from each speaker for the training set and assigned the remaining utterances to the test set. This yielded 480 photographs for training and 120 images for testing. The model was trained for 300 epochs with a batch size of 80.

In the case of the GRID corpus dataset [41], we took a testing set that included eight speakers (five males and three females) using the speaker-independent protocol. We used the 26 speakers remaining for training. Consequently, 8000 CFI images are obtained for the test and 26,000 CFI images for the train. The model was trained for 200 epochs with a batch size of 250. For the GLips dataset, we used the default split and trained the model for 100 epochs with a batch size of 188. The QCMs-PSO filter produces the optimized QCMs matrix of the CFI image, which is then provided to the CNN instead of the input CFI image. We conducted a series of incremental order changes and noted the results, ultimately determining the optimal order based on the highest accuracy rate achieved. Indeed, we conducted an ablation analysis to evaluate our model's performance by removing the QCMs-PSO and using the same architecture as in Tables 3 and 4. We adjusted the CFI size for each dataset: for the LRW dataset, the CFI was scaled to 150 × 350; for the GRID dataset, the CFI dimensions were established at 256 × 256; and for the GLips dataset, the CFI was composed of 29 frames, each measuring 29 × 29. We utilized the identical model architecture outlined in Tables 3 and 4 for the LRW [42] and GRID [41] datasets throughout training. For the GLips dataset [44], we modified the fully connected layers by augmenting the number of neurons from 400 and 300 to 2720 and 1200, respectively, to accommodate the increased dataset size and the higher number of classes.

During the training of the proposed model, the QCMs-PSO descriptor generates an optimal moment matrix, the dimensions of which depend on the orders chosen from the input lip image. The matrix is subsequently input into the CNN. We methodically adjust the order size and document the corresponding accuracy to get the optimal order that yields the maximum performance. Additionally, to analyze the influence of the optimized QCMs-PSO descriptor, we performed a comparison analysis utilizing the identical architecture specified in Tables 3 and 4, without the filter, thereby enabling us to evaluate its contribution to the model's efficacy.

The performance of the proposed automatic lip-reading classification model is evaluated in this experiment. The effectiveness of our model in visual speech recognition is assessed using two key metrics: information loss and accuracy. In this experiment, we evaluate the recognition capabilities of the proposed QCMs-PSO-CNN architecture, utilizing

features derived from quaternion Charlier moments optimized by the PSO algorithm as the input vector. The training and validation curves shown in Fig. 16 illustrate the model's performance over 100 to 300 epochs for the three datasets, focusing on accuracy and loss metrics.

5.5. Comparison methodology

The simulation outcomes, including classification rates for the LRW [42] and GLips [44] datasets, were assessed using several moment orders and juxtaposed with *meta*-heuristic algorithms including BBO [32], ACO [33], IWO [35], SFLA [36], CA [37], FA [39], and HS [40]. This comparison underscores the efficacy of the utilized PSO algorithm in our approach. Fig. 17 illustrates that the PSO algorithm consistently surpasses competing algorithms, especially at lower moment orders. Table 5 additionally contrasts our approach with prior studies on the LRW [42] dataset, encompassing CFI-based CNN [46], ResNetLSTM [47], LSTM [18], and HCNN [23]. Furthermore, we disclose the accuracy of our method in the absence of the PSO algorithm. The results indicate that our technology far surpasses current techniques, highlighting the efficacy of the optimized QCMs descriptor. This underscores the advantages of our method regarding classification accuracy and resilience, even when employing a singular lip image to represent a complete series. In addition, our experiment shows that using either the noise injection or rotation data augmentation can lead to better results than the works using rotation, translation, flipping, and color shift data augmentation compared to. Furthermore, we performed a comparison examination of our technique against two pertinent studies, as detailed in Table 6. The initial study, conducted by Assael et al. [48], presents LipNet, a sentence-level lipreading system utilizing 3D convolutions in conjunction with Gated Recurrent Units (GRUs) for visual speech detection.

The second study, conducted by Wand et al. [49], introduces a word-level lipreading system that employs Long Short-Term Memory (LSTM) networks and Neural Networks (NNs) featuring multiple feed-forward layers. Our optimal descriptor QCMs-PSO-CNN strategy demonstrates enhanced classification accuracy while markedly decreasing complexity relative to these methods. Likewise, using the GRID [41] dataset, the incorporation of the optimal QCMs descriptor results in a significant enhancement above 40 % relative to a conventional CNN employing the SI protocol with solely rotation-based data augmentation (DA). Table 7 illustrates a comparison of our technique with RTMRBM [50] and alm-GRU [51] on the GLips [44] dataset. The results indicate that QCMs-PSO-CNN surpasses these techniques, and the incorporation of DA further improves the model's efficacy.

5.6. Ablation studies

In this study, we conducted multiple ablation studies on the GRID [41], LRW [42], and GLips [44] datasets. Specifically, we designed two ablation experiments to evaluate the effectiveness of the proposed QCMs-PSO-CNN architecture.

In the first experiment, we compare the performance of the proposed architecture for visual speech recognition with that of approaches reported in the literature. We evaluate accuracy, precision, recall, and F1-score for the GRID, LRW, and GLips datasets. These metrics are summarized in Table 8, which compares QCMs-PSO-CNN with traditional machine learning classifiers, including DNN-LSTM [56], DCNN [57], 3DCNN-LSTM [59], CNN-AlexNet [58], CNN-GRU [60], 3DCNNs + CFS-DCTCN [61], SVM + Random Forest [62], DBNFTLSTM [64], and CNN-LSTM [63]. Our QCMs-PSO-CNN achieved the highest accuracy of 99.97 %, significantly outperforming baseline models such as 3DCNNs + CFS-DCTCN (99.50 %) and DBNFTLSTM (99.30 %). This analysis highlights the model's accuracy and provides additional metrics, including recall, precision, and the F1-score.

In the second experiment, as previously mentioned, we validated the effectiveness of incorporating QCMs optimized by the PSO algorithm into the CNN model, which demonstrated strong performance in addressing lipreading challenges. Fig. 18 illustrates that our straightforward QCMs-PSO-CNN design surpasses five comparable studies in classification accuracy while markedly diminishing complexity. Moreover, it is apparent that this optimized architecture yields enhanced outcomes on both the GRID and GLips datasets relative to alternative techniques. Integrating QCMs-PSO-CNN with the combination of the optimal QCMs-PSO descriptor and the CNN model yields superior performance compared to more intricate methodologies, such as LipNet and LSTM, which depend on a considerably greater number of parameters.

6. Conclusion

This paper proposed a novel small architecture QCMs-PSO-CNN for the VSR problem. This architecture consists of two parts: The Optimized Quaternion Charlier Moments by proposed particle swarm optimization algorithm and Convolutional Neural Networks. The QCMs-PSO-CNN comes to enhance the classification rate by extracting useful features from the input large-size image. Indeed, it can reduce the dimensionality of the images and therefore decrease significantly the high complexity of the CNN. As we have exhibited before, with this small architecture, we obtained better results than the other complex models on the three datasets GRID, LRW, and GLips. In this study, we developed a novel descriptor based on quaternion Charlier moments optimized using the PSO algorithm. These descriptors are used in the Convolutional Neural Networks model (QCMs-PSO-CNN) for visual speech recognition. The key elements of this contribution can be summarized as follows: (i) The study introduces a unique methodology that combines QCMs-PSO Optimized by PSO algorithm, and CNN model, with deep learning techniques. (ii) The integration of a quaternion Charlier moments optimized and a Convolutional Neural Network in the QCMs-PSO-CNN architecture allows for the extraction of both spatial and temporal features from lipreading data. (iii) The research emphasizes the importance of optimized descriptor vector by combining information obtained from QCMs-PSO-based feature extraction and deep learning. This descriptor vector enhances the ability to capture and utilize discriminative features for visual speech recognition. In future work related to lipreading models, such as extending the model to higher-dimensional systems or incorporating stochastic elements, we plan to apply and enhance the proposed QCMs-PSO-CNN method to address other complex problems in computer vision, particularly those involving augmented reality techniques and their integration into robotics. Furthermore, we aim to expand our research by investigating the stability and synchronization of inertial memristive neural networks with time delays [65], as well as the synchronization of Markovian jump neural networks for sampled-data control systems with additive delay components [66], with the ultimate goal of applying these approaches to visual speech recognition.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors have not used any Generative AI and AI-assisted tools to prepare this manuscript. The authors take full responsibility for the content of the publication.

Data availability statements.

The datasets used in this paper are publicly available as:

GRID dataset at https://spandh.dcs.shef.ac.uk/gridcorpus/.

LRW dataset at https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html.

GLips dataset at https://www.fdr.uni-hamburg.de/record/10048.

CRediT authorship contribution statement

Omar El Ogri: Writing – original draft, Conceptualization. **Jaouad EL-Mekkaoui:** Methodology, Data curation. **Mohamed Benslimane:** Visualization, Software. **Amal Hjouji:** Validation, Project

administration. **Abdelali Saidi:** Writing – review & editing, Investigation. **Musheer Ahmad:** Writing – review & editing, Supervision. **Hela Elmannai:** Validation, Formal analysis. **Ahmed A. Abd El-Latif:** Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors of this article would like to thank Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R747), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Additionally, the authors would like to thank Prince Sultan University for their support.

References

[1] C. G. Fisher. Confusions among visually perceived consonants. J Speech Hear Res 11(4), p. 796-804, déc. 1968, doi: 10.1044/jshr.1104.796.
[2] S. Hilder, R. W. Harvey, et B.-J. Theobald. Comparison of human and machine-based lip-reading.. in *AVSP*, 2009, p. 86-89. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: http://www2.cmp.uea.ac.uk/~bjt/avsp2009/proc/papers/paper-25.pdf.
[3] Adebayo BM, et al. Comparative analysis of deep learning models for part of speech tagging in the malay language. HighTech Innovat J 2024;5(2):272–81.
[4] J. S. Chung et A. Zisserman. Lip Reading in the Wild. in *Computer Vision – ACCV 2016*, vol. 10112, S.-H. Lai, V. Lepetit, K. Nishino, et Y. Sato, Éd., in Lecture Notes in Computer Science, vol. 10112. , Cham: Springer International Publishing, 2017, p. 87-103. doi: 10.1007/978-3-319-54184-6_6.
[5] Anina I, Zhou Z, Zhao G, Pietikäinen M. Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); 2015. p. 1–5. Consulté le: 20 avril 2024. [En ligne]. Disponible sur:.
[6] Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R. Extraction of visual features for lipreading. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: IEEE Trans Pattern Anal Mach Intell 2002;24(2):198–213. https://ieeexplore.ieee.org/abstract/document/982900/.
[7] Cox SJ, Harvey RW, Lan Y, Newman JL, Theobald B-J. The challenge of multispeaker lip-reading.. in *AVS*. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: Citeseer 2008:179–84. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3ccf2122a3890b062798790acc8d1d1c084a7b0f.
[8] Lee B, et al. AVICAR: Audio-visual speech corpus in a car environment. In *Eighth International Conference on Spoken Language Processing*. 2004. Consulté le: 20 avril 2024. [En ligne]. Disponible sur:.
[9] T. J. Hazen, K. Saenko, C.-H. La, J. R. Glass. A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. In: *Proceedings of the 6th international conference on Multimodal interfaces*, State College PA USA: ACM, oct. 2004, p. 235-242. doi: 10.1145/1027933.1027972.
[10] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. in *2002 IEEE International conference on acoustics, speech, and signal processing*, IEEE, 2002, p. II-2017. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/5745028/.
[11] Cooke M, Barker J, Cunningham S, Shao X. An audio-visual corpus for speech perception and automatic speech recognition. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: J Acoust Soc Am 2006;120(5):2421–4. https://pubs.aip.org/asa/jasa/article-abstract/120/5/2421/934379.
[12] Huang J, Potamianos G, Connell J, Neti C. Audio-visual speech recognition using an infrared headset. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: Speech Comm 2004;44(1–4):83–96. https://www.sciencedirect.com/science/article/pii/S0167639304001116.
[13] P. Lucey, G. Potamianos, S. Sridharan. Patch-based analysis of visual speech from multiple views. in *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, AVISA, 2008, p. 69-74. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: https://eprints.qut.edu.au/15247/.
[14] Messer K, Matas J, Kittler J, Luettin J, Maitre G. XM2VTSDB: the extended M2VTS database. in *second international conference on audio and video-based biometric person authentication*. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: Citeseer 1999:965–6. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b62628ac06bbac998a3ab825324a41a11bc3a988.
[15] Ha, Nicole Yah Yie, Lee Yeng Ong, and Meng Chew Leow. Slowfast-tcn: a deep learning approach for visual speech recognition. Emerg Sci J 8.6 (2024): 2554-2569.
[16] Hedayati-Dezfooli M, et al. Optimizing Injection molding for propellers with soft computing, fuzzy evaluation, and Taguchi method. Emerg Sci J 2024;8:2101–19.
[17] Zhao G, Barnard M, Pietikainen M. Lipreading with local spatiotemporal descriptors. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: IEEE Trans Multimed 2009;11(7):1254–65. https://ieeexplore.ieee.org/abstract/document/5208233/.
[18] S. Petridis, M. Pantic. Deep complementary bottleneck features for visual speech recognition. in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, p. 2304-2308. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/7472088/.
[19] A. Bakry, A. Elgammal. Manifold-Kernels Comparison in MKPLS for Visual Speech Recognition. 21 janvier 2016, *arXiv*: arXiv:1601.05861. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1601.05861.
[20] C. Tian, W. Ji. Auxiliary Multimodal LSTM for Audio-visual Speech Recognition and Lipreading. 17 mars 2017, *arXiv*: arXiv:1701.04224. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1701.04224.
[21] J. S. Chung, A. Zisserman. Out of Time: Automated Lip Sync in the Wild. in *Computer Vision – ACCV 2016 Workshops*, vol. 10117, C.-S. Chen, J. Lu, et K.-K. Ma, Éd., in Lecture Notes in Computer Science, vol. 10117. , Cham: Springer International Publishing, 2017, p. 251-263. doi: 10.1007/978-3-319-54427-4_19.
[22] T. Saitoh, Z. Zhou, G. Zhao, M. Pietikäinen. Concatenated Frame Image Based CNN for Visual Speech Recognition. in *Computer Vision – ACCV 2016 Workshops*, vol. 10117, C.-S. Chen, J. Lu, et K.-K. Ma, Éd., in Lecture Notes in Computer Science, vol. 10117. , Cham: Springer International Publishing, 2017, p. 277-289. doi: 10.1007/978-3-319-54427-4_21.
[23] Mesbah A, Berrahou A, Hammouchi H, Berbia H, Qjidaa H, M.. Daoudi. Lip reading with Hahn convolutional neural networks. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: Image Vis Comput 2019;88:76–83. https://www.sciencedirect.com/science/article/pii/S0262885619300605.
[24] Lu Y, Li H. Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Consulté le: 20 avril 2024. [En ligne]. Disponible sur: Appl Sci 2019;9(8):1599. https://www.mdpi.com/2076-3417/9/8/1599.
[25] Kim M, Yeo JH, Ro YM. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In: In *Proceedings of the AAAI Conference on Artificial Intelligence*; 2022. p. 1174–82. Consulté le: 20 avril 2024. [En ligne]. Disponible sur:.
[26] Baaloul A, Benblidia N, Reguieg FZ, Bouakkaz M, Felouat H. An arabic visual speech recognition framework with CNN and vision transformers for lipreading. Multimed Tools Appl, Févr 2024. https://doi.org/10.1007/s11042-024-18237-5.
[27] Zhu H, Liu M, Shu H, Zhang H, Luo L. General form for obtaining discrete orthogonal moments. IET Image Proc 2010;4(5):335. https://doi.org/10.1049/iet-ipr.2009.0195.
[28] H. Zhu, M. Liu, Y. Li, H. Shu, H. Zhang. Image description with nonseparable two-dimensional charlier and meixner moments. Int J Pattern Recognit Artif Intell 25 (01), p. 37-55, févr. 2011, doi: 10.1142/S0218001411008506.
[29] W. R. Hamilton, *Elements of quaternions*. London: Longmans, Green, & Company, 1866. Consulté le: 11 août 2024. [En ligne]. Disponible sur: https://books.google.com/books?hl=fr&lr=&id=fIRAAAAAIAAJ&oi=fnd&pg=PR1&dq=Hamilton,+W.+R.+(1866).+Elements+of+quaternions.+London:+Longmans,+Green,+%26+Company&ots=DHcH0V5hRK&sig=S0tSi44fz_ttHVSpdieQtYvnnQI.
[30] J. Kennedy et R. Eberhart. Particle swarm optimization. in *Proceedings of ICNN'95-international conference on neural networks*, ieee, 1995, p. 1942-1948. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/488968/.
[31] Y. Shi, R. C. Eberhart. Empirical study of particle swarm optimization. in *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, IEEE, 1999, p. 1945-1950. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/785511/.
[32] Garg H. An efficient biogeography based optimization algorithm for solving reliability optimization problems. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: Swarm Evol Comput 2015;24:1–10. https://www.sciencedirect.com/science/article/pii/S2210650215000395.
[33] Dorigo M, Birattari M, Stutzle T. Ant colony optimization. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: IEEE Comput Intell Mag 2006;1(4):28–39. https://ieeexplore.ieee.org/abstract/document/4129846/.
[34] B. Xing, W.-J. Gao. Imperialist Competitive Algorithm. in *Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms*, vol. 62, in Intelligent Systems Reference Library, vol. 62. , Cham: Springer International Publishing, 2014, p. 203-209. doi: 10.1007/978-3-319-03404-1_15.
[35] B. Xing, W.-J. Gao. Invasive Weed Optimization Algorithm. in *Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms*, vol. 62, in Intelligent Systems Reference Library, vol. 62. , Cham: Springer International Publishing, 2014, p. 177-181. doi: 10.1007/978-3-319-03404-1_13.
[36] M. M. Eusuff, K. E. Lansey. Optimization of Water Distribution Network Design Using the Shuffled Frog Leaping Algorithm. J Water Resour Plan Manag 129(3), p. 210-225, mai 2003, doi: 10.1061/(ASCE)0733-9496(2003)129:3(210).
[37] Muhamediyeva DT. Fuzzy cultural algorithm for solving optimization problems. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: J Phys: Conf Ser, IOP Publish 2020:012152. https://iopscience.iop.org/article/10.1088/1742-6596/1441/1/012152/meta.
[38] Storn R, Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 1997;11(4):341–59. https://doi.org/10.1023/A:1008202821328.
[39] X.-S. Yang. Firefly Algorithms for Multimodal Optimization. in *Stochastic Algorithms: Foundations and Applications*, vol. 5792, O. Watanabe et T. Zeugmann, Éd., in Lecture Notes in Computer Science, vol. 5792. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 169-178. doi: 10.1007/978-3-642-04944-6_14.

[40] Wang C-M, Huang Y-F. Self-adaptive harmony search algorithm for optimization. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: Expert Syst Appl 2010;37(4): 2826–37. https://www.sciencedirect.com/science/article/pii/S0957417409007891.

[41] The GRID audiovisual sentence corpus ». Consulté le: 21 avril 2024. [En ligne]. Disponible sur: https://spandh.dcs.shef.ac.uk/gridcorpus/.

[42] Lip Reading in the Wild (LRW) dataset. Consulté le: 21 avril 2024. [En ligne]. Disponible sur: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html.

[43] D. Hu, X. Li. Temporal multimodal learning in audiovisual speech recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, p. 3574-3582. Consulté le: 11 août 2024. [En ligne]. Disponible sur: http://openaccess.thecvf.com/content_cvpr_2016/html/Hu_Temporal_Multimodal_Learning_CVPR_2016_paper.html.

[44] G. Schwiebert, C. Weber, L. Qu, H. Siqueira, S. Wermter. GLips - German Lipreading Dataset. 1 mars 2022. doi: 10.25592/uhhfdm.10048.

[45] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25, 2012, Consulté le: 21 avril 2024. [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[46] T. Saitoh, Z. Zhou, G. Zhao, M. Pietikäinen. Concatenated Frame Image Based CNN for Visual Speech Recognition. in *Computer Vision – ACCV 2016 Workshops*, vol. 10117, C.-S. Chen, J. Lu, et K.-K. Ma, Éd., in Lecture Notes in Computer Science, vol. 10117. , Cham: Springer International Publishing, 2017, p. 277-289. doi: 10.1007/978-3-319-54427-4_21.

[47] T. Stafylakis, G. Tzimiropoulos. Combining Residual Networks with LSTMs for Lipreading. 8 septembre 2017, *arXiv*: arXiv:1703.04105. Consulté le: 20 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1703.04105.

[48] Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas. LipNet: End-to-End Sentence-level Lipreading. 16 décembre 2016, *arXiv*: arXiv:1611.01599. Consulté le: 20 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1611.01599.

[49] M. Wand, J. Koutník, et J. Schmidhuber. Lipreading with long short-term memory. in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, p. 6115-6119. Consulté le: 20 janvier 2024. [En ligne]. Disponible sur: https://ieeexplore.ieee.org/abstract/document/7472852/.

[50] D. Hu et X. Li. Temporal multimodal learning in audiovisual speech recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, p. 3574-3582. Consulté le: 20 janvier 2024. [En ligne]. Disponible sur: http://openaccess.thecvf.com/content_cvpr_2016/html/Hu_Temporal_Multimodal_Learning_CVPR_2016_paper.html.

[51] Yuan Y, Tian C, Lu X. Auxiliary loss multimodal GRU model in audio-visual speech recognition. Consulté le: 20 janvier 2024. [En ligne]. Disponible sur: IEEE Access 2018;6:5573–83. https://ieeexplore.ieee.org/abstract/document/8279447/.

[52] NadeemHashmi S, Gupta H, Mittal D, Kumar K, Nanda A, Gupta S. A lip reading model using CNN with batch normalization. in *2018 eleventh international conference on contemporary computing (IC3)*. Consulté le: 20 janvier 2024. [En ligne]. Disponible sur: IEEE 2018:1–6. https://ieeexplore.ieee.org/abstract/document/8530509/.

[53] Ait KY, et al. Automatic lipreading using convolutional neural networks and orthogonal moments. Mathemat Model Comput 2025;12(1):90–100.

[54] Debnath, Saswati, Pinki Roy. Audio-visual automatic speech recognition using PZM, MFCC and statistical analysis. (2021).

[55] Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Array 2022;16:100258.

[56] Shashidhar R, Patilkulkarni S, Puneeth SB. Combining audio and visual speech recognition using LSTM and deep convolutional neural network. Int J Inf Technol 2022;14(7):3425–36.

[57] Ramadan RA. RETRACTED ARTICLE: Detecting adversarial attacks on audio-visual speech recognition using deep learning method. Int J Speech Technol 2022;25(3): 625–31.

[58] Rajab MhA, Hashim KM. An automatic lip reading for short sentences using deep learning nets. Int J Adv Intellig Informat 2023;9:1.

[59] Alameen SA, Alhothali AM. A lightweight driver drowsiness detection system using 3DCNN with LSTM. Comput Syst Sci Eng 2023;44(1).

[60] Ameer AWA, Salehpour P, Asadpour M. Deep transfer learning for lip reading based on NASNetMobile pretrained model in wild dataset. IEEE Access 2024.

[61] Li Y, Hashim AS, Lin Y, Nohuddin PN, Venkatachalam K, Ahmadian A. AI-based visual speech recognition towards realistic avatars and lip-reading applications in the metaverse. Appl Soft Comput 2024;164:111906.

[62] Vekkot S, Gupta TS, Karthik KP, Kaushik D. Enhanced lip reading using deep model feature fusion: a study on the MIRACL-VC1 dataset. Procedia Comput Sci 2025; 258:1189–98.

[63] Baloch A, Ali M, Hussain L, Sadiq T, Alkahtani BS. Urdu lip reading systems for digits in controlled and uncontrolled environment. IEEE Access 2025.

[64] Wang X. DBN-FTLSTM: an optimized deep learning framework for speech and image recognition. Informatica 2025;49(20).

[65] Rakkiyappan R, et al. Stability and synchronization analysis of inertial memristive neural networks with time delays. Cogn Neurodyn 2016;10(5):437–51.

[66] Thendral T, Marimuthu, et al. Synchronization of Markovian jump neural networks for sampled data control systems with additive delay components: Analysis of image encryption technique. Mathematical Methods in the Applied Sciences. 2022.

Omar El Ogri received the B.Eng. degree in Mathematical and Computer Sciences, the M.S. degree in engineering science and PhD degrees in Signals, Systems and Informatics from the Faculty of science, University of Sidi Mohammed Ben Abdellah, Fez, Morocco in 2012, 2018 and 2022, respectively. He is currently a Professor in Computer Science with the Higher School of Technology of Sidi bennour, Chouaib Doukkali University, El Jadida, Morocco. His research interests include applied mathematics, artificial intelligence, pattern recognition, multimedia security, image processing, and data science.

Jaouad EL-Mekkaoui received his Bachelor's, Master's, and State Doctorate degrees in Applied Mathematics from the Faculty of Science and Technology at Sidi Mohammed Ben Abdellah University, Morocco, in 2004, 2006, and 2013, respectively. He is currently a professor of Applied Mathematics and Computer Science at the Higher School of Technology, Sidi Mohammed Ben Abdellah University in Fez, Morocco. His research interests include numerical methods, applied mathematics, image processing, and artificial intelligence.

Mohamed Benslimane is a Moroccan computer science scholar born in Fez (Morocco) in 1986. He is currently a Senior Lecturer (Maître de Conférences Habilité) at the Higher School of Technology of Fez (ESTF), affiliated with Sidi Mohamed Ben Abdellah University. He serves as Head of the Computer Engineering Department and Director of the LSIG Research Laboratory. With a Ph.D. in Engineering Sciences and Techniques, his research interests include intelligent distance education, wireless sensor networks and artificial intelligence. He has completed numerous professional certifications in e-learning, networking, and data science. He is President of the Moroccan Association of Innovative Technologies (AMTI) and Treasurer of CRREP. Prof. Benslimane has been involved in academic teaching, research, and project supervision for over a decade. He is also certified by Huawei and has received training from prestigious institutions. His dynamic academic career blends scientific excellence with leadership in technological innovation.

Amal Hjouji received her Bachelor's, Master's, and State Doctorate degrees in Computer Science from the Faculty of Sciences at Sidi Mohammed Ben Abdellah University in Morocco, in 2014, 2016, and 2020, respectively. She is currently a faculty member in the Computer Science Department at the same university. Her research interests include image processing, pattern recognition, image classification, artificial intelligence, neural networks, deep learning, data science, and big data.

Abdelali Saidi received the B.Eng. degree in Mathematical and Computer Sciences, the M.S. degree in Computer Networks and Systems and PhD degrees in Computer Sciences from the Faculty of science, University of Mohammed V, Rabat, Morocco in 2009, 2011 and 2016, respectively. He is currently a Professor in Computer Science with the Higher School of Technology of Sidi bennour, Chouaib Doukkali University, El Jadida, Morocco. His research interests include intrusion detection systems, cloud computing security and artificial intelligence.

Musheer Ahmad received the B.Tech., and M.Tech. degrees from the Department of Computer Engineering, Aligarh Muslim University, India, in 2004 and 2008, respectively, and the Ph.D. degree in the area of Chaos-based Cryptography from the Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India. From 2007 to 2010, he has worked in the Department of Computer Engineering, Aligarh Muslim University, Aligarh. Since 2011, he has been working as an Assistant Professor in the Department of Computer Engineering, Jamia Millia Islamia. He has published over 150 research papers in internationally reputed refereed journals and conference proceedings of the IEEE, Springer, Elsevier. He has more than 5200 citations of his research works with an H-index of 42, i-10 index of 105, and cumulative JCR impact factor nearing 300. He has been consecutively listed among World's Top 2% researchers in studies conducted by Elsevier and Stanford University in 2021, 2022, 2023, and 2024. His research interests include multimedia security, chaos-based cryptography, cyber security, machine learning & deep learning, and optimization techniques. He has served as a reviewer and a technical program committee member of many international conferences. He is serving as associate editor of International Journal of Information Security and Privacy (IJISP), and International Journal of Artificial Intelligence in Scientific Disciplines (IJAISD) from IGI. He is Editorial Board member of Scientific Reports, and Discovery Computing, both from Springer. He has also served as referee of some renowned journals, such as Information Sciences, Signal Processing, Journal of Information Security and Applications, Expert Systems with Applications, Knowledge-based Systens, Applied Soft Computing, Engineering Applications of Artificial Intelligence, Chaos Solitons & Fractals, Physica A, Signal Processing: Image Communication, Neurocomputing, IEEE IOTJ, IEEE JSAC, IEEE JBHI, IEEE TCYB, IEEE TCSVT, IEEE TII, IEEE TPAMI, IEEE TNNLS, IEEE TITS, IEEE SMCA, IEEE TCE, IEEE TCDS, IEEE TNSE, IEEE TNB, IEEE TCAS, IEEE TBD, IEEE TR, IEEE MULTIMEDIA, IEEE ACCESS, Wireless Personal Communications, Neural Computing and Applications, Multimedia Tools & Applications, International Journal of Bifurcation and Chaos, IET Information Security, IET Image Processing, Security and Communication Networks, Optik, Complexity, Computers in Biology and Medicine, Computational and Applied Mathematics, Concurrency and Computation & so on. Recently, he is felicitated with Jamia Achievers Award by Jamia Millia Islamia, New Delhi.

Hela Elmannai an Associate professor at the department of Infromation Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia. She got her PhD degree in Information Technology from SUPCOM in Tunisia. Her research interests include artificial intelligence, networking, blockchain and engineering applications.

Ahmed A. Abd El-Latif is a Professor of Quantum Cybersecurity, with over 18 years of professional experience. He earned his Ph.D. with honors from Harbin Institute of Technology, China, in 2013, and has since led numerous successful research projects and grants across Egypt, the Russian Federation, Saudi Arabia, and Tunisia. Since 2024, he has held the position of Full Professor at Menoufia University since July 2024. His extensive research portfolio includes over 330 publications in high-impact journals and conferences, 20 books, and more than 14,000 citations. Since 2022, he has served as the Head of the MEGANET 6G Lab in the Russian Federation and is the Vice Chair of the EIAS Research Lab. Additionally, he is the Founder and Deputy Director of the Centre of Excellence in Quantum & Intelligent Computing. Prof. Abd El-Latif has held prestigious academic and research roles globally, including Head of the MEGANET6G Lab in Russia, Postdoctoral Fellow at Harbin Institute of Technology, China, and Visiting Professorships at the Polish Academy of Sciences and Warsaw University of Technology, Poland, producing over 36 high-impact publications during these engagements. His research focuses on pioneering advancements in quantum computing, chaotic dynamical systems, and artificial intelligence, with applications in cybersecurity and the Artificial Intelligence of Things (AIoT). His work in quantum computing emphasizes the development of novel algorithms for data representation, cryptography, privacy, authentication, and information hiding in quantum scenarios. He is currently leading groundbreaking research in quantum information processing and quantum-inspired algorithms to address the challenges of secure data transmission and storage in the quantum computing era. His contributions extend to applications in wireless sensor networks, IoT, healthcare, and smart cities. A recognized leader in his field, Prof. Abd El-Latif has received several prestigious awards, including the State Encouragement Award in Engineering Sciences (2016, Egypt), the Best Ph.D. Student Award from Harbin Institute of Technology (2013, China), and the Young Scientific Award from Menoufia University (2014, Egypt). He serves as Editor-in-Chief of the International Journal of Information Security and Privacy and as a series editor for Quantum Information Processing and Computing and Advances in Cybersecurity Management. He also holds editorial roles in numerous indexed journals (WoS and Scopus quartile-ranked).Prof. Abd El-Latif`s vision is to lead the transition to quantum-safe cybersecurity by developing innovative cryptographic protocols and educating the next generation of researchers and professionals in post-quantum cryptography and quantum-inspired security solutions. His work continues to shape the future of secure data transmission and storage in the digital and quantum eras.