MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 1, pp. 90–100 (2025)

Automatic lipreading using convolutional neural networks and orthogonal moments

Ait Khayi Y.[1], El Ogri O.[1,2], El-Mekkaoui J.[1], Benslimane M.[1], Hjouji A.[3]

[1] *TI, Laboratory, EST, Sidi Mohamed Ben Abdellah University, Fez, Morocco*
[2] *CED-ST, STIC, Laboratory of Information, Signals, Automation and Cognitivism LISAC, Dhar El Mahrez Faculty of Science, Sidi Mohamed Ben Abdellah-Fez University, Fez, Morocco*
[3] *Sidi Mohamed Ben Abdellah University, Fez, Morocco*

(Received 7 January 2024; Revised 10 January 2025; Accepted 13 January 2025)

Recently, understanding speech from a speaker's mouth using only visual interpretation of the lips movement has become one of the most complex computer vision tasks. In the present paper, we suggest a new approach named Optimized Quaternion Meixner Moments Convolutional Neural Networks (OQMMCNN) in order to develop a lipreading system based only on video images. This approach is based on Quaternion Meixner Moments (QMMs) that we use as a filter in the Convolutional Neural Networks (CNN) architecture. In addition, we use the Grey Wolf optimization algorithm (GWO) with the aim of ensuring high accuracy of classification through the optimization of the Quaternion Meixner Moments (QMMs) filter local parameters. We show that this method is an effective solution to decrease the high dimensionality of the video images and the training time. This approach is tested on a public dataset and compared to different methods that use complex models and deep architecture in the literature.

Keywords: *lipreading; GWO algorithm; Meixner polynomials; Meixner moments; quaternion representations.*

2010 MSC: 68T10, 62H30, 49M05 **DOI:** 10.23939/mmc2025.01.090

1. Introduction

Lipreading, or visual speech recognition (VSR), is the ability to detect text information from a speaker's mouth movement. It plays a vital role in speech understanding and communication for people experiencing struggles interacting with society. In the lipreading task, tongue and tooth movements are interpreted by only using visual information during speech transcription. Recently, it has notoriously attracted much attention because of the necessity for its application across many various fields, especially the medical field. People with hearing impairments (hearing issues) or deaf people, especially those affected by laryngeal cancer, rely on visual information to perceive and understand spoken words. Consequently, the lipreading is an inevitable task to help people recuperate their communication to engage in social activities, especially when therapy can take a long time. Thanks to advancements in image processing and computer vision research, it has become achievable to decode spoken words from lip movements to imitate human lipreading capability [1–6]. However, using visual information alone to comprehend oral languages can be a harder task, particularly in the absence of context. Therefore, lip reading demands specific abilities to follow lip motions, teeth, and articulation of the tongue and to face the similarity between phonemes that Fisher has explained in [7]. Furthermore, other issues that make this task more challenging are the differences between the mouth shapes of each speaker and the distinction between words that have the same pronunciation [2]. Therefore, all these variations can cause confusion and ambiguity at the level of the word. As an example, certain words like "right" and "write" or the words "pack", "back", and "mac" produce almost identical lip movements but differ in their meanings and sounds. Extracting features and recognition require a robust lip reading system that can approach all these variations. Hilder et al. in 2009 [8] conducted a comparison between automated lipreading systems and human lipreading to see who performed better, and the experiment demonstrated that automatic lipreading systems performed better than human lipreading. Thus, an automatic speech recognition system is needed to address these challenges.

Progressing towards building a visual speech recognition system (VSRS) that can face these challenges, several traditional approaches were proposed and evaluated on various databases, AVLetters [9], AVLetters2 [10], AVICAR [11], AVTIMIT [12], CUAVE [13], Grid [14], such as the work of Petridis and Pantic in 2016 [15]. They introduced a method as a framework based on auto-encoder to directly extract deep bottleneck characteristics from pixels for training a model using a Long-Short Term Memory Recurrent Neural Network (LSTM-RNN), the accuracy of this method was 58.1%. A paper by Matthews et al. [9], presents the approach of Active Shape Models (ASM), and Active Appearance Models (AAM) for extracting features and training a model using Hidden Markov model (HMM) for classifications, the results showed that the model obtained 44.6% accuracy. An approach named CFI-based CNN was suggested by Saitoh et al. [16]. They used a novel sequence image representation that is simple and contains spatial-temporal information for tackling the VSR task. It was tested on the OuluVS2 dataset, and the performance by using the AlexNet model without data augmentation DA and by using the GoogLeNet model with DA on OuluVS2 digits 90B° (profile view) were 59.3% and 87.5% respectively. A method called Hahn Convolutional Neural Network (HCNN) based on Hahn moments and CNNs was introduced by A. Mesbah et al. In 2019 [17], the authors have developed a model employing the moments of Hahn as descriptors in the architecture of CNN, the accuracy of this method was 59.23% on the AVLetters dataset. A method called Multi-Head Visual-Audio Memory (MVM) was introduced by M. Kim et al. in 2022 [18] that consists of one value memory to save representative audio features and multi-head key memories to preserve visual characteristics. The accuracy of this method was 88.5% LRW dataset.

In order to address all these variations, this paper provides an approach called Optimized Quaternion Meixner Moments Convolutional Neural Networks (OQMMCNN). It consists of three main parts: firstly, quaternion algebra with its ability to treat a color image holistically as a vector field. Second, Meixner moments, which are able to identify, hold, and extract the most useful information from images with less redundancy and effectiveness. Third, we use the convolutional neural network (CNN) because of its performance in learning patterns and image classification. Additionally, the Grey Wolf Optimization algorithm (GWO), [19] is utilized to enhance classification accuracy through the optimization of the local parameters α and β of the Quaternion Meixner Moments (QMMs).

The remainder of this paper is structured as follows: in Section 2, we will present Quaternion Meixner moments and their theoretical foundation. In Section 3, we will concentrate on the GWO algorithm and the steps for selecting the optimal parameters of QMMs. In Section 4, we will introduce the architecture of the proposed method. The description of the dataset, the experiments, and the results are carried out in Section 5. Finally, Section 6 will conclude this paper.

2. Quaternion Meixner moments

2.1. Quaternion representation

The quaternion numbers are an expansion of complex numbers that were first introduced by Hamilton in 1843 [20]. In general, a number of quaternion is formed by one real part and three imaginary parts represented in the form:

$$q = q_0 + q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k},$$

where $q_0, q_1, q_2, q_3 \in \mathbf{R}$, and $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are three imaginary units that obey the following equation:

$$\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1, \quad \boldsymbol{ij} = -\boldsymbol{ji} = \boldsymbol{k}, \boldsymbol{jk} = -\boldsymbol{kj} = \boldsymbol{i}, \quad \boldsymbol{ki} = -\boldsymbol{ik} = \boldsymbol{j}.$$

The previous equations show that quaternion multiplication is not commutative. In the case of $q_0 = 0$, the number q is named a pure quaternion. The following expressions are the definitions of a quaternion's conjugate and modulus, respectively,

$$q^* = q_0 - q_1\boldsymbol{i} - q_2\boldsymbol{j} - q_3\boldsymbol{k},$$
$$\|q\| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}.$$

Let $f(x,y)$ be the intensity function of a color image with three RGB channels defined in Cartesian coordinates. The quaternion representation can be used to represent each pixel with the pure quaternion using:

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

while the blue, red, and green pixel components (x,y) are denoted, respectively, by $f_B(x,y)$, $f_R(x,y)$, and $f_G(x,y)$.

2.2. Quaternion Meixner moments

Let $f(x,y)$ be the intensity function of a color image with three channels RGB defined in Cartesian coordinates. As the quaternion number is not commutative, we define the right side of QMMs using the following formula:

$$\mathrm{QMM}_{nm}^{R}(f) = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} f(x,y)\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\mu \tag{1}$$

(R is a reference to right-side quaternion) for $n = 0\colon N-1$, $m = 0\colon M-1$, and μ represents a pure quaternion unit that was selected for this work as $\mu = -(i+j+k)/\sqrt{3}$. $\tilde{M}_n^{(\alpha,\beta)}(x)$ represents the discrete orthogonal Meixner polynomials of the n-th order, given by

$$\tilde{M}_n^{(\alpha,\beta)}(x) = M_n^{(\alpha,\beta)}(x)\sqrt{\frac{\omega(x)}{\rho(n)}},$$

where

$$\omega(x) = \frac{\beta^x(\alpha)_x}{x!}, \quad \rho(n) = \frac{n!(\alpha)_n}{\beta^n(1-\beta)^\alpha}$$

and

$$M_n^{(\alpha,\beta)}(x) = (\alpha)_n \sum_{k=0}^{n} \frac{(-n)_k(-x)_k}{(\alpha)_k k!}\left(1-\frac{1}{\beta}\right)^k$$

with $(\gamma)_k$ is the symbol of Pochhammer. Additionally, the Eq. (1) can be explained by

$$\mathrm{QMM}_{nm}^{R}(f) = Q_0^R + Q_1^R\boldsymbol{i} + Q_2^R\boldsymbol{j} + Q_3^R\boldsymbol{k},$$

where

$$Q_0^R = \frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_0\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right],$$

$$Q_1^R = -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_1\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right],$$

$$Q_2^R = -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_2\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right],$$

$$Q_3^R = -\frac{1}{\sqrt{3}}\left[\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} W_3\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right].$$

With

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

More explicitly, the coefficients Q_0^R, Q_1^R, Q_2^R, and Q_3^R can also be expressed using the moments of Meixner of an intensity function $f(x,y)$ with order $n \times m$ defined over $[0; N-1]$ and $[0; M-1]$ given

as

$$\mathrm{MM}_{nm}(f) = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} \tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)f(x,y)$$

by

$$\begin{aligned}
Q_0^R &= \frac{1}{\sqrt{3}}\left[\mathrm{MM}_{nm}(f_R) + \mathrm{MM}_{nm}(f_G) + \mathrm{MM}_{nm}(f_B)\right],\\
Q_1^R &= -\frac{1}{\sqrt{3}}\left[\mathrm{MM}_{nm}(f_G) - \mathrm{MM}_{nm}(f_B)\right],\\
Q_2^R &= -\frac{1}{\sqrt{3}}\left[\mathrm{MM}_{nm}(f_B) - \mathrm{MM}_{nm}(f_R)\right],\\
Q_3^R &= -\frac{1}{\sqrt{3}}\left[\mathrm{MM}_{nm}(f_R) - \mathrm{MM}_{nm}(f_G)\right].
\end{aligned}$$

As a result of the orthogonality property exhibited by Meixner polynomials, the inverse transformation of the right side of QMMs can be easily calculated according to the following equation:

$$\hat{f}(x,y) = \sum_{n=0}^{\check{N}-1}\sum_{m=0}^{\check{M}-1} \tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\,\mathrm{QMM}_{nm}^R(f)\mu, \tag{2}$$

where $0 \leqslant \check{N} \leqslant N$, $0 \leqslant \check{M} \leqslant M$. More explicitly, Eq. (2) can also be given by the form:

$$\hat{f} = \hat{f}_0 + \hat{f}_1\boldsymbol{i} + \hat{f}_2\boldsymbol{j} + \hat{f}_3\boldsymbol{k},$$

where

$$\begin{aligned}
\hat{f}_0 &= \frac{1}{\sqrt{3}}\left[\sum_{n=0}^{\check{N}-1}\sum_{m=0}^{\check{M}-1}(Q_1+Q_2+Q_3)\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right],\\
\hat{f}_1 &= -\frac{1}{\sqrt{3}}\left[\sum_{n=0}^{\check{N}-1}\sum_{m=0}^{\check{M}-1}(Q_0+Q_2-Q_3)\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right],\\
\hat{f}_2 &= -\frac{1}{\sqrt{3}}\left[\sum_{n=0}^{\check{N}-1}\sum_{m=0}^{\check{M}-1}(Q_0-Q_1+Q_3)\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right],\\
\hat{f}_3 &= -\frac{1}{\sqrt{3}}\left[\sum_{n=0}^{\check{N}-1}\sum_{m=0}^{\check{M}-1}(Q_0+Q_1-Q_2)\tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\right].
\end{aligned}$$

The representation of images $\bar{f}$ using QMMs with the order $n\times m$ results in reconstruction errors. To calculate the error of reconstruction between the initial image (f) and its reconstructed representation ($\bar{f}$), the following mean-square error (MSE) criterion can be used:

$$\mathrm{MSE} = \frac{1}{N\times M}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left[f(x,y) - \bar{f}(x,y)\right]^2, \tag{3}$$

where $\bar{f}$ is the image reconstructed using Meixner moments, which can be calculated as follows:

$$\bar{f}(x,y) = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} \tilde{M}_n^{(\alpha,\beta)}(x)\tilde{M}_m^{(\alpha,\beta)}(y)\mathrm{MM}_{nm}(f).$$

To accelerate the numerical computation, we have chosen to use the following matricial form:

$$\bar{f} = \tilde{M}_1\mathrm{MM}\tilde{M}_2^T, \tag{4}$$

while $(\cdot)^T$ denotes the transposed matrix, and $\bar{f}$, $\tilde{M}_1$, $\tilde{M}_2^T$, and MM are respectively the matrix forms of $\bar{f}(x,y)$, $\tilde{M}_n^{(\alpha,\beta)}(x)$, $\tilde{M}_m^{(\alpha,\beta)}(y)$, and MM_{nm}.

When the MSE value approaches zero, it indicates a high degree of similarity between the original and its representation reconstructed image.

3. Optimization of QMMs using the GWO

To maximize the utility of the Meixner polynomials in practical applications like classification, it is essential to carefully choose the appropriate parameters. This objective becomes progressively more complicated as the number of parameters increases. However, in order to attain this objective, the parameters must satisfy the following conditions: $\left\{\beta \in \left[\frac{N}{2}, N\right], \alpha = 0.5\right\}$ [21]. Inspired by the desire to address this issue, we will employ a meta-heuristic algorithm [22,23], to determine the parameters of the QMMs from hundreds of choices. Meta-heuristic algorithms for optimization have demonstrated their utility in addressing challenges across diverse fields [24–27]. One notable example of those algorithms is the Grey Wolf Optimization algorithm (GWO), that has a particular place in swarm intelligence methods. The GWO algorithm is inspired by the hierarchical framework observed in the social gray wolves organizations. This organization is divided into four categories: α solution (representing the fittest one), β solution (signifying the second fittest one), δ solution (corresponding to the third fittest one), and the rest of solutions ω [19]. It is crucial to point out that the GWO algorithm progresses through three primary hunting phases: chasing, encircling, and attack.

The hunting actions of wolf packs can be defined by the following equations:

$$\boldsymbol{X}(t+1) = \boldsymbol{X}_p(t) - \boldsymbol{A} \times \boldsymbol{D}, \tag{5}$$

$$\boldsymbol{D} = |\boldsymbol{C} \times \boldsymbol{X_p}(t) - \boldsymbol{X}(t)|, \tag{6}$$

such that t represents the present iteration, $\boldsymbol{D}$ is a vector defining the novel location of the gray wolves, $\boldsymbol{X}(t)$ corresponds to the current locations of the gray wolf, and the vector $\boldsymbol{X}_p(t)$ represents the prey location. $\boldsymbol{A}$ and $\boldsymbol{C}$ represent the coefficient vectors computed using the following formulas [19]:

$$\boldsymbol{A} = 2\boldsymbol{a}\boldsymbol{r}_1 - \boldsymbol{a},$$
$$\boldsymbol{C} = 2\boldsymbol{r}_2,$$

$\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are random numbers ranging from $[0, 1]$, while $\boldsymbol{a}$ represent a fit parameter which is linearly changes from 2 to 0. With the use of both Eqs. (3) and (5), the GWO algorithm leverages the value of alpha to determine the optimal solution. However, it employs beta and delta to update the position of the other grey wolves using the following equations [19]:

$$\boldsymbol{D_\alpha} = |\boldsymbol{C}_1 \times \boldsymbol{X}_\alpha(t) - \boldsymbol{X}(t)|,$$
$$\boldsymbol{D_\beta} = |\boldsymbol{C}_2 \times \boldsymbol{X}_\beta(t) - \boldsymbol{X}(t)|,$$
$$\boldsymbol{D_\delta} = |\boldsymbol{C}_3 \times \boldsymbol{X}_\delta(t) - \boldsymbol{X}(t)|,$$

$$\boldsymbol{X_1} = \boldsymbol{X}_\alpha - \boldsymbol{A}_1 \times \boldsymbol{D}_\alpha,$$
$$\boldsymbol{X_2} = \boldsymbol{X}_\beta - \boldsymbol{A}_2 \times \boldsymbol{D}_\beta,$$
$$\boldsymbol{X_3} = \boldsymbol{X}_\delta - \boldsymbol{A}_3 \times \boldsymbol{D}_\delta,$$

$$\boldsymbol{X}_p(t+1) = \frac{\boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_3}{3}. \tag{7}$$

After locating the prey, the attack process is modeled as a linear reduction in $\boldsymbol{a}$ from two to zero. $\boldsymbol{A}$ is a value that is chosen randomly from $[-2\boldsymbol{a}, 2\boldsymbol{a}]$. According to Eq. (6), the decrease in $\boldsymbol{a}$ also results in a corresponding decrease in $\boldsymbol{A}$.

— If $|A| < 1$, the search agents follow the best solution α to lunch the attack. This stage is known as global optimum convergence.
— Otherwise, the search agents are obligated to diverge from the prey to leave from the local optimum and find the optimal solution.

Specifically, the suggested procedure for choosing the best parameter values for QMMs $V_{\text{opt}} = \left[\alpha_1^{\text{opt}}, \alpha_2^{\text{opt}}, \beta_1^{\text{opt}}, \beta_2^{\text{opt}}\right]$ by the GWO algorithm is provided in Algorithm 1.

Algorithm 1 QMMs Optimization with GWO.

Input: Intensity $f(x, y)$ of image, order of moment $N_{\max}$, size of image $N \times M$.
Output: Optimized parameters for QMMs $V_{\text{opt}} = \left[\alpha_1^{\text{opt}}, \alpha_2^{\text{opt}}, \beta_1^{\text{opt}}, \beta_2^{\text{opt}}\right]$.
Data: $SA = 50$ (population size), $T = 100$ (maximum number of iterations), $D = 4$ (problem dimension), $U = [1, 1, N, N]$ (maximum values of parameter), $L = [0, 0, 1, 1]$ (minimum values of parameter).
Initialize $\boldsymbol{a}$, $\boldsymbol{A}$, and $\boldsymbol{C}$.
Evaluation of the fitness function MSE provided by equation Eq. (3) for each grey wolf and sorting them using their genre: $\boldsymbol{X_\alpha}$, $\boldsymbol{X_\beta}$, $\boldsymbol{X_\delta}$.
$t = 0$.
while $(t < T)$
    for For every search agent
        Randomly initialise $\boldsymbol{r}_1$, $\boldsymbol{r}_2$.
        Update the location of the current search agent by using Eq. (7).
    Update $\boldsymbol{a}$, $\boldsymbol{A}$, and $\boldsymbol{C}$.
    Evaluate the objective function values for all search agents and sort them.
    Update the location of $\boldsymbol{X}_\alpha$, $\boldsymbol{X}_\beta$, $\boldsymbol{X}_\delta$.
    $t = t + 1$.
return $\boldsymbol{X}_\alpha$.

Figure 1 demonstrates that the parameter values obtained through the algorithm of GWO exhibit excellent performance in the representation of the color image when compared with the classical choice of parameter $\left\{\beta \in \left[\frac{N}{2}, N\right], \alpha = 0.5\right\}$.

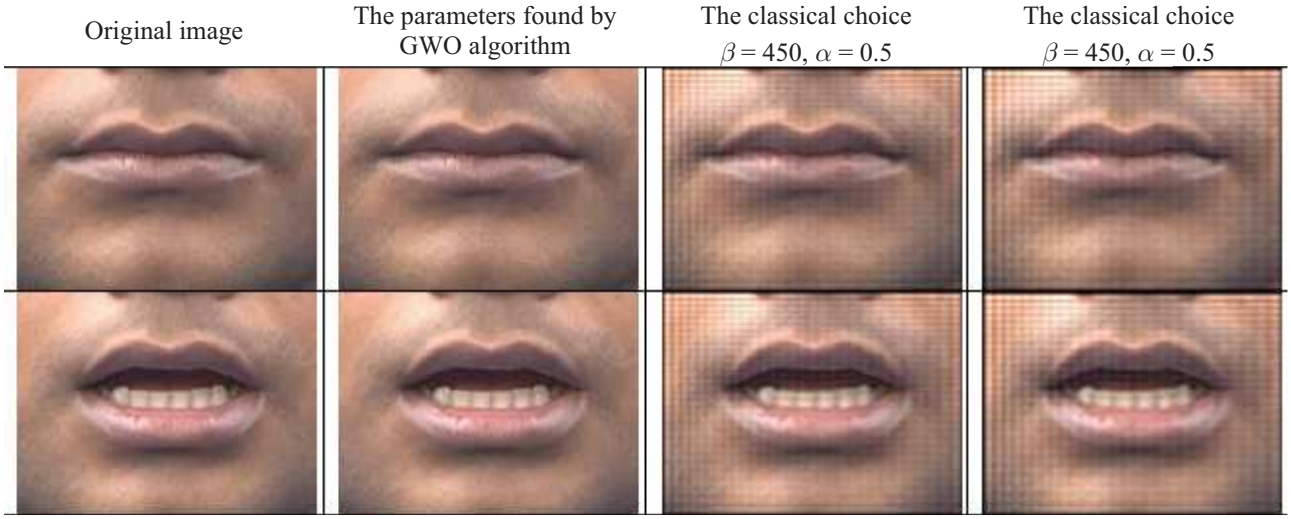

Fig. 1. The images reconstructed through various parameter choices.

4. The proposed OQMMCNN architecture

The OQMMCNN architecture proposed in this paper aims to address the VSR challenges by providing powerful classification and processing lip images rapidly. It is a combination of the Quaternion Meixner Moments approach, the meta-heuristic GWO algorithm, and the convolutional neural network model (CNN).

The OQMMCNN method offers a solution to overcome the high computational costs and complicated hardware requirements. In addition, using QMMs as a filter enables the representation of global properties in the images and the description of their significant features. Furthermore, the GWO algorithm aims to optimize the objective function MSE towards a global optimum. Consequently, achieving a reconstructed image with a minimal MSE value indicates the efficiency of the method of parameter optimization employed. The parameter's description and the model used in this paper are introduced as shown in Figure 2.

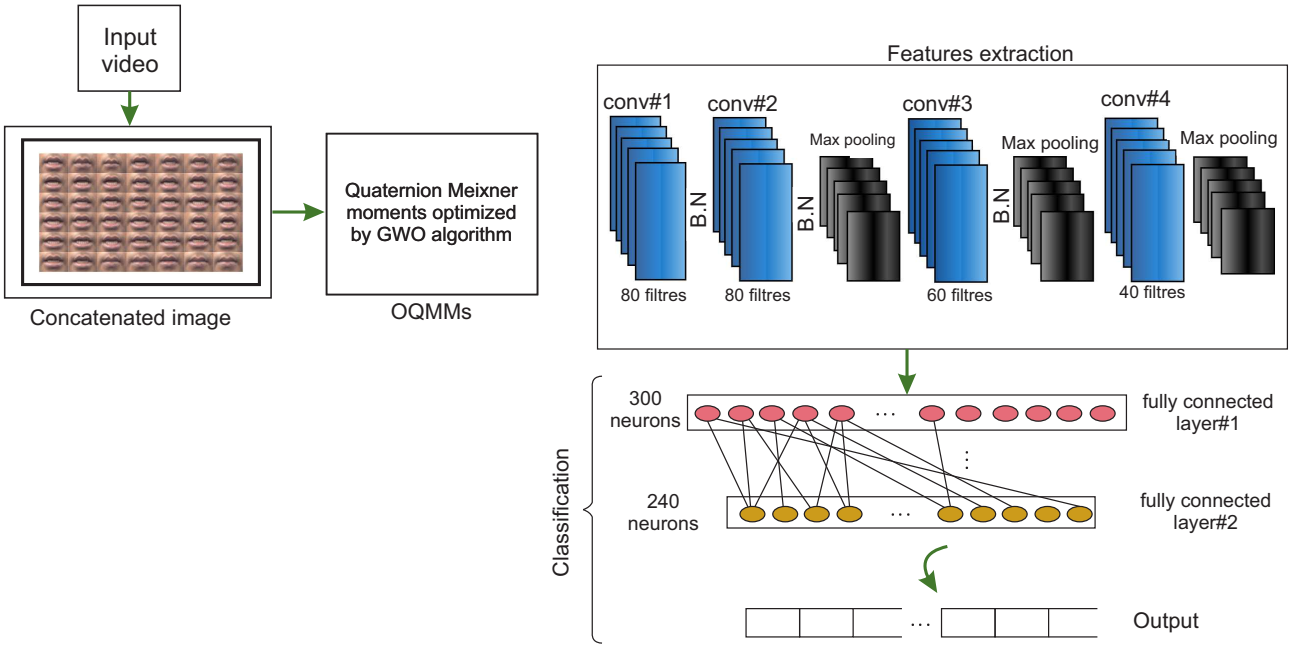

Fig. 2. OQMMCNN parameters: filter of the optimized quaternion Meixner moments with the desired order, convolution 1 (kernel 3×3 and 80 filters), convolution 2 (kernel 3×3 and 80 filters), max pooling 1 (pool size 3×3), convolution 3 (kernel 3×3 and 60 filters), max pooling 2 (pool size 3×3), convolution 4 (kernel 3×3 and 40 filters), max pooling 3 (pool size 3×3). Fully connected layer 1 (300 neurons), fully connected layer 2 (240 neurons). Finally, there is an output layer of 26 classes.

5. Experimental results

In this part, the results of the proposed model in comparison to some works in the literature are carried out. The dataset is introduced first, and following that, the experimental results are displayed and presented.

5.1. Dataset

The AVletters2 dataset is an extension of AVLetters. It is one of the biggest datasets for VSR. that was established by Cox et al. in 2008 [10]. The AVLetters2 dataset consists of the letters 'A' to 'Z' uttered by five people, seven times for each letter. Every speaker commences and finish with a closed mouth, and the frame count varies in every video. This dataset was taken in color with high-definition cameras 1920×1080 in RGB. Figure 3 illustrates a sample of images from the AVletters2 dataset.

5.2. Preprocessing

To prepare the chosen database, our first step involves extracting frames from the provided videos in AVLetters2. The number of frames extracted varies depending on the video length, which can be challenging when it comes to handling them. To address this issue, we add an extra frame to the first of each video to ensure a consistent total frame count across each video. We opt to include the first frame in every video, as we anticipate that it will not impact the video's content. The first frame typically consists of repetition and does not contribute any additional information. We have fixed the total frame number to 42 for every video. Therefore, for videos with fewer than 42 frames, we replicate the first frame until we get the desired count of 42 frames. Additionally, we employ the Concatenated Frame Image (CFI) method, as introduced by Saitoh et al. in their work [16], in order to represent all frames as a single image. Finally, we organize the 42 frames into a single image with a layout of 7 frames in each row and 6 frames in each column, as illustrated in Figure 4.

Fig. 3. Example of extracted frames for the AVLetters2 dataset.

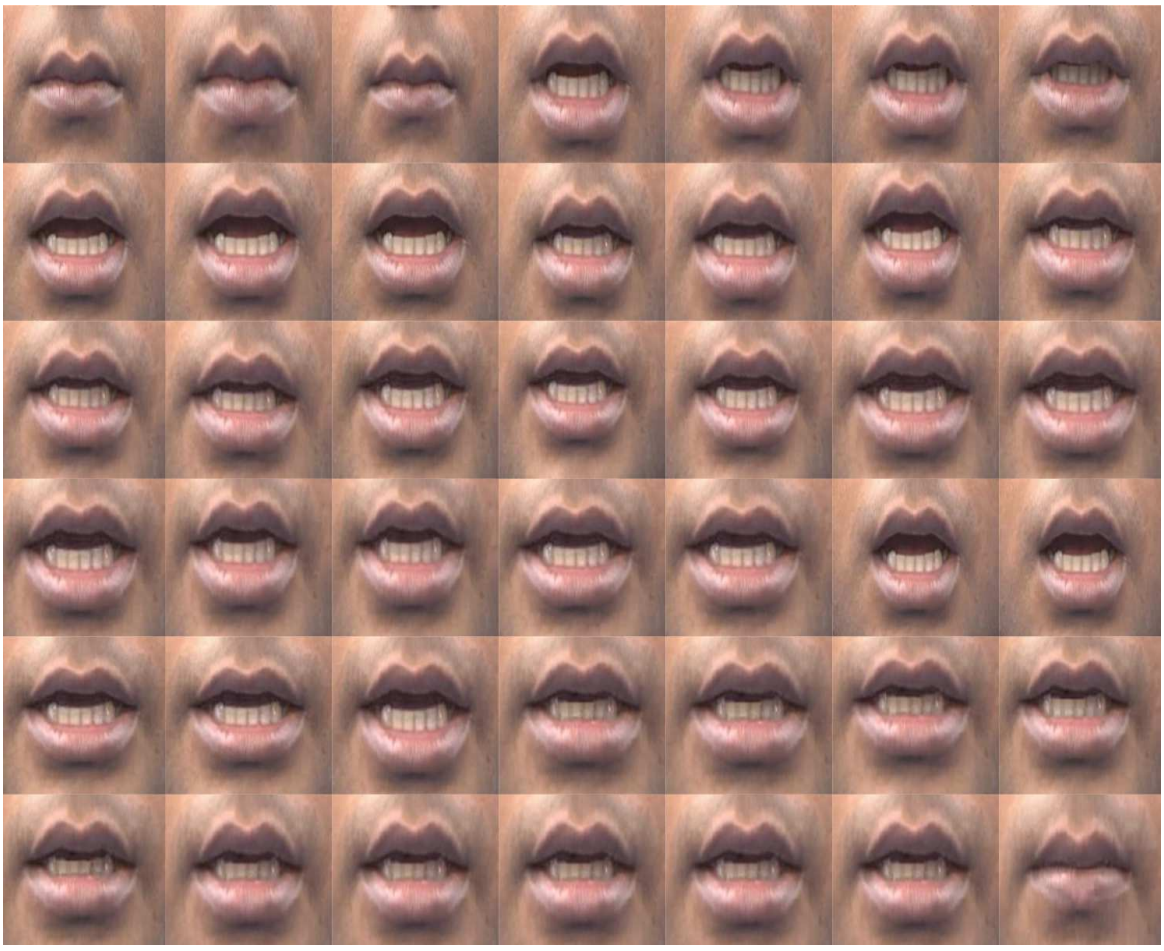

Fig. 4. Example of concatenated frame image for the AVLetters2 dataset.

5.3. Results and discussion

The simulation results in terms of classification rates for the AVLetters2 dataset with different moments' order are listed in Table 1. It is obvious that the greatest results can be achieved at order 300.

Table 1. Achieved results for AVLetters2 using different orders of optimized quaternion Meixner moments.

Order	16	32	64	100	150	200	300
Accuracy	82.55%	93.25%	95.84%	96.67%	97.23%	98.96%	99.75%

Table 2. The results achieved for AVLetters2 compared to different methods in the literature.

Method	HMM [9]	LBM-SVM [28]	LSTM-RNN [15]	HCNN [17]	Proposed OQMMCNN
Accuracy	44.64%	58.85%	58.10%	59.23%	98.65%

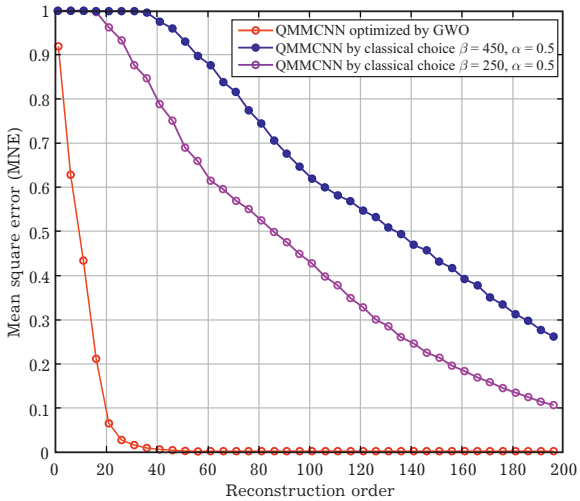

Fig. 5. The MSE for QMMCNN.

Table 2 provides a comparison between our method and previous works conducted on the AVLetters2 dataset.

Our method clearly outperforms the compared methods, illustrating the effectiveness of employing the optimized quaternion Meixner moments filter. This highlights the superiority of our approach in terms of classification accuracy and robustness. Figure 5 shows the MSE curve for our proposed method. It can be observed that the MSE approach to zero in small orders, which shows the efficiency of our optimization algorithm.

6. Conclusion

In this work, we have proposed a novel approach named Optimized Quaternion Meixner Moments Convolutional Neural Network for VSR. The proposed approach offers an effective solution for mitigating the extensive computational requirements of deep learning. The performance of the suggested model was assessed on the publicly dataset, AVLetters2. The results illustrate the effectiveness of the proposed method in accurately classifying letters given as video images.

[1] Fernandez-Lopez A., Sukno F. M. Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing. **78**, 53–72 (2018).

[2] Hao M., Mamut M., Yadikar N., Aysa A., Ubul K. A survey of research on lipreading technology. IEEE Access. **8**, 204518–204544 (2020).

[3] Chen X., Jixiang D., Hongbo Z. Lipreading with DenseNet and resBi-LSTM. Signal, Image and Video Processing. **14**, 981–989 (2020).

[4] Fenghour S., Chen D., Guo K., Xiao P. Lip reading sentences using deep learning with only visual cues. IEEE Access. **8**, 215516–215530 (2020).

[5] Rashkevych Yu., Peleshko D., Pelekh I., Izonin I. V. Speech signal marking on the base of local magnitude and invariant segmentation. Mathematical Modeling and Computing. **1** (2), 234–244 (2014).

[6] Ma S., Wang S., Lin X. A transformer-based model for sentence-level Chinese Mandarin lipreading. 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC). 78–81 (2020).

[7] Fisher C. G. Confusions among visually perceived consonants. Journal of Speech and Hearing Research. **11** (4), 796–804 (1968).

[8] Hilder S., Harvey R., Theobald B.-J. Comparison of human and machine-based lip-reading. AVSP 2009 – International Conference on Audio-Visual Speech Processing University of East Anglia (2009).

[9] Matthews I., Cootes T. F., Bangham J. A., Cox S., Harvey R. Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence. **24** (2), 198–213 (2002).

[10] Cox S., Harvey R., Lan Y., Newman J., Theobald B.-J. The challenge of multispeaker lip-reading. International Conference on Auditory-Visual Speech Processing (AVSP). (2008).

[11] Lee B., Hasegawa-Johnson M., Goudeseune C., Kamdar S., Borys S., Liu M., Huang T. AVICAR: Audio-visual speech corpus in a car environment. Eighth International Conference on Spoken Language Processing. 2489–2492 (2004).

[12] Hazen T. J., Saenko K., La C.-H., Glass J. R. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. Proceedings of the 6th international conference on Multimodal interfaces. 235–242 (2004).

[13] Patterson E. K., Gurbuz S., Tufekci Z., Gowdy J. N. CUAVE: A new audio-visual database for multimodal human-computer interface research. 2002 IEEE International conference on acoustics, speech, and signal processing. **II**, 2017–2020 (2002).

[14] Cooke M., Barker J., Cunningham S., Shao X. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America. **120** (5), 2421–2424 (2006).

[15] Petridis S., Pantic M. Deep complementary bottleneck features for visual speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2304–2308 (2016).

[16] Saitoh T., Zhou Z., Zhao G., Pietikäinen M. Concatenated frame image based CNN for visual speech recognition. Computer Vision – ACCV 2016 Workshops. 277–289 (2017).

[17] Mesbah A., Berrahou A., Hammouchi H., Berbia H., Qjidaa H., Daoudi M. Lip reading with Hahn convolutional neural networks. Image and Vision Computing. **88**, 76–83 (2019).

[18] Kim M., Yeo J. H., Ro Y. M. Distinguishing homophenes using multi-head visual-audio memory for lip reading. Proceedings of the AAAI Conference on Artificial Intelligence. **36** (1), 1174–1182 (2022).

[19] Mirjalili S., Mirjalili S. M., Lewis A. Grey Wolf Optimizer. Advances in Engineering Software. **69**, 46–61 (2014).

[20] Lewis A. C. William Rown Hamilton, Lectures on quaternions (1853). Landmark Writings in Western Mathematics 1640–1940. 460–469 (2005).

[21] Sayyouri M., Hmimid A., Qjidaa H. A fast computation of novel set of Meixner invariant moments for image analysis. Circuits, Systems, and Signal Processing. **34**, 875–900 (2015).

[22] Sadeeq H., Abdulazeez A. M. Hardware implementation of firefly optimization algorithm using FPGAs. 2018 International Conference on Advanced Science and Engineering (ICOASE). 30–35 (2018).

[23] Sadeeq H. T., Abdulazeez A. M. Giant trevally optimizer (GTO): A novel metaheuristic algorithm for global optimization and challenging engineering problems. IEEE Access. **10**, 121615–121640 (2022).

[24] Naserbegi A., Aghaie M. Exergy optimization of nuclear-solar dual proposed power plant based on GWO algorithm. Progress in Nuclear Energy. **140**, 103925 (2021).

[25] Naserbegi A., Aghaie M., Zolfaghari A. Implementation of Grey Wolf Optimization (GWO) algorithm to multi-objective loading pattern optimization of a PWR reactor. Annals of Nuclear Energy. **148**, 107703 (2020).

[26] Gachkevich M., Gachkevich O., Torskyy A., Dmytruk V. Mathematical models and methods of optimization of technological heating regimes of the piecewise homogeneous glass shell. State-of-the-art investigations. Mathematical Modeling and Computing. **2** (2), 140–153 (2015).

[27] Raskin L., Sira O., Sagaydachny D. Multi-criteria optimization in terms of fuzzy criteria definitions. Mathematical Modeling and Computing. **5** (2), 207–220 (2018).

[28] Zhao G., Barnard M., Pietikainen M. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia. **11** (7), 1254–1265 (2009).

Автоматичне читання з губ за допомогою згорткових нейронних мереж і ортогональних моментів

Айт Хайі Й.[1], Ель Огрі О.[1,2], Ель-Меккауї Дж.[1], Бенсліман М.[1], Хйоджі А.[3]

[1] *TI, Лабораторія, EST, Університет Сіді Мохамеда Бен Абделлаха, Фес, Марокко*
[2] *CED-ST, STIC, Лабораторія інформації, сигналів, автоматизації та когнітивізму LISAC, Факультет природничих наук Дхар Ель Махрез, Університет Сіді Мохамед Бен Абделлах-Фез, Фез, Марокко*
[3] *Університет Сіді Мохамед Бен Абделлах-Фез, Фез, Марокко*

Останнім часом розуміння мови з вуст оратора за допомогою лише візуальної інтерпретації рухів губ стало одним із найскладніших завдань комп'ютерного зору. У цій роботі пропонується новий підхід, названий "Оптимізовані згорткові нейронні мережі кватерніонних моментів Мейкснера" (OQMMCNN), щоб розробити систему читання з губ, засновану лише на відеозображеннях. Цей підхід базується на кватерніонних моментах Мейкснера (QMM), які використовуються як фільтр в архітектурі згорткових нейронних мереж (CNN). Крім того, використовується алгоритм оптимізації сірого вовка (GWO) з метою забезпечення високої точності класифікації за допомогою оптимізації локальних параметрів фільтра кватерніонних моментів Мейкснера (QMM). Показано, що цей метод є ефективним для зменшення розмірності відеозображень і часу навчання. Цей підхід перевірено на загальнодоступному наборі даних і порівнюється з різними відомими з літератури методами, які використовують складні моделі та глибоку архітектуру.

Ключові слова: *читання по губах; алгоритм GWO; поліноми Мейкснера; моменти Мейкснера; кватерніонні представлення.*