IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 28, 2020, accepted March 11, 2020, date of publication March 20, 2020, date of current version March 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982359

A Silent Password Recognition Framework Based on Lip Analysis

MOHAMED EZZ[1,2], (Member, IEEE), AYMAN MOHAMED MOSTAFA[1,3], (Member, IEEE), AND ABDURRAHMAN A. NASR[2]

[1]College of Computer and Information Sciences, Jouf University, Sakaka 72314, Saudi Arabia
[2]Faculty of Engineering, Al-Azhar University, Cairo 11651, Egypt
[3]Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt

Corresponding author: Ayman Mohamed Mostafa (amhassane@ju.edu.sa)

ABSTRACT Securing passwords either written or spoken is considered one of the challenging authentication issues faced by individuals and organizations. Written passwords could be easily stolen by look over, or man-behind, while spoken passwords could be recorded and replayed by attackers. A proposed silent password which is based on a dual security model for lip movement analysis will be a promoting solution to these attacks. The goal of the current research is to propose a hybrid voting framework for silent passwords recognition using lip movement analysis. The proposed framework is built for Arabic language by extracting figures from predefined Arabic lexicon. The predefined lexicon mainly contains the Arabic figures from zero to nine with different shapes. The framework takes a video or sequence of images as an input, and outputs the corresponding silent password for the frames extracted from the input video. In this paper, three techniques will be employed to extract effective visual features from mouth-lip movement. Such techniques are SURF, HoG and Haar feature extractor. The resultant features in each technique are fed separately into a classification model, namely, the hidden Markov model (HMM). The HMM identifies corresponding Arabic figure from a predefined lexicon based on input features. The final classification models that are produced from the three techniques have been grouped in a voting scheme to produce the final classification result. The proposed model will be tested on handcrafted data set of lip movement, and it has shown a promising result with improved accuracy of Arabic figures recognition.

INDEX TERMS Hidden Markov model, lip analysis, silent password, voting scheme.

I. INTRODUCTION

One of the most common methods of authentication and confidentiality used by intrusion detection systems is based on the exploitation of private passwords. Choosing a strong password can prevent security breaches from penetrating confidential information. Many organizations rely on the use of the voice signature to identify the password in order to access system resources. Even this password is hard to be predicted, it can be identified by other users who are able to record the password during user login. Using silent password protection engine, based on lip movement recognition, is a promoting solution to this problem.

Lip analysis is considered an appealing methodology for different researchers over the last few decades because it has the ability to understand audio and visual information. The lip analysis has advanced a new technology trend for cases where audio is not allowed or must be secured in professional way [1]. Indeed, this field of research has been tackled by many researches to solve for audio speech recognition, by visual speech information. Such visual speech has proven to enhance the robustness and accuracy of automatic speech recognition [2]. This is because of visual information is invariant to acoustic noise troubles.

The proposed methodology exploits visual information by means of feature detector and descriptor techniques. The visual information is encoded as a set of feature descriptors for the selected visual source (e.g. image or video). These descriptors are fed into classification model in order to identify corresponding Arabic word from a predefined lexicon based on input features. The Arabic lexicon only includes the Arabic figures from zero to nine, that is, Sefr (zero), Waahed

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed M. Elmisery.

(one), Ethnan (two), Thlatha (three), Arbaa (four), Khamsa (five), Setta (six), Sabaa (seven), Thamania (eight) and Tesaa (nine).

Here, three techniques have been employed for visual feature detection and description. Such techniques are SURF [3] (speeded up robust features), HoG [4] (histogram of oriented gradient) and Haar [5] for extracting lip contour features from visual source. SURF technique works by transforming the source image into coordinates using the integral image algorithm on ECG signals [6], which rapidly calculates summations of pixels over image sub-regions in constant time. SURF next applies the detection and description procedures for extracting visual features.

In order to detect the regions of interest in input image, SURF utilizes the Hessian matrix detector. The detector is applied on variant-size box filters on the integral image, so that it's able to work on different scales and locations. The resultant determinant of the Hessian matrix is used as a measure of local changes around the point [3]. For point description, SURF uses Wavelet responses in horizontal and vertical direction with the aid of integral image for fast calculations. A neighborhood of size 20 × 20 is taken around the key point and is divided into 4 × 4 sub-regions. For each sub-region, horizontal and vertical wavelet responses are taken and a vector is formed. The final SURF feature descriptor for the 4 × 4 sub-regions is 64 dimensions length for describing point of interest.

The second technique employed in this paper is the histogram of oriented gradient (HoG). HoG detects local edge gradient direction around the region of interest (e.g. lips) by dividing the region into four 5 × 3 cells and the gradient for each pixel within the cell is discretized into one of 9 direction boxes. Each pixel contributes to the local histogram of the cell with a "vote" equals to the gradient magnitude. The combination of the normalized histograms represents the feature descriptor, which is a vector of the components of the cell histograms from all of the block regions.

The third technique for visual figure recognition is the Haar-like features extractor. The technique encodes the oriented contrast between regions of interest in input image. Haar detects and extracts the region of interest in input image by training a decision stump classifier on set of positive and negative examples with set of different sizes kernel images [6]. The feature value is obtained by subtracting sum of pixels under white rectangle from sum of pixels under black rectangle in an integral image. The accuracy of Haar classifier is enhanced by using classifier cascade technique (several stages of decision stump that are applied to a region of interest) and AdaBoost algorithm.

In order to recognize the Arabic word from visual sources, the feature vectors produced from the aforementioned techniques, are fed separately into a statistical machine learning algorithm, namely, the hidden Markov model (HMM). Hidden Markov model aims at modeling a sequence of events with different stages, and speech is a sequence of voice with different parts. As such, HMM precisely matches our area of research. Moreover, HMM training algorithms are very popular, simple, and computationally feasible to use. In this paper, A sequence of words or phonemes is applied by merging the speaker trained hidden Markov models for the separate words and phonemes. HMM creates stochastic models from known visual word utterances training data, and compares the probability that the unknown utterance or test data was generated by each model. Visual speech recognition system represents words with hidden Markov models (HMMs) with each state corresponding to a phoneme. The probability for each state is modeled by a mixture of Gaussians, trained with the expectation-maximization algorithm (EM) [7].

The proposed silent password framework is divided into two directions; the first direction is used for extracting features from captured lip image that relies on using image recognition technologies and the second direction is used for modeling sequence of lip movements to capture a sequence of words or phonemes that relies on time-series machine learning technologies.

There are mainly two directions now for image recognition. The first direction uses traditional image recognition algorithms such as SURF, Hog, and Haar. The second direction uses deep learning algorithms such as CNN (convolution neural network) architecture that requires huge dataset and more resources for training and real time prediction. Here, traditional image recognition techniques are selected due to limited Arabic lip dataset and for providing quasi-optimal authentication mechanism suitable for devices with limited capabilities (smart devices, e.g. mobiles, tablets, etc.). So the SURF, HoG and Haar techniques have been chosen based on their remarkable efficacy and reported accuracy in image recognition as shown in [8], [9] and [10]. The Hidden Markov Model (HMM) is used on the proposed silent lip recognition framework instead of deep learning LSTM because deep learning requires large amount of data while the Arabic dataset used in the silent lip recognition is limited.

To improve the final result and accuracy of HMM classifiers that are produced from training separate HMM on each technique, a voting model has been setup to yield the final result by amalgamating the result of the three classifiers.

The contribution of this paper is as follows:

- Proposing a dual authentication framework for verifying users using access control mechanism.
- Proposing an authentication framework based on silent passwords for protecting personal password confidentiality.
- The proposed authentication framework can be adapted for deaf and dump users and any camera based devices such as smart phones, tablets, IOT devices.
- The developed recognition system will be tested through different visual features extraction techniques from mouth-lip movement

The remaining part of this paper is structured as follows: Section 2 presents the related works. Section 3 presents the dual authentication framework. Section 4 details the proposed silent password recognition framework based on lip movements Analysis. Section 5 explains the data preprocessing stage. Section 6 applies the face and lip detection process. Section 7 explains the lip feature extraction mechanism. Section 8 explores the threat model for recent security mechanisms. Section 9 explores the experimentation carried out to evaluate the proposed framework. Section 10 proposes a comparative study for evaluating the performance of recent lip analysis researches.

II. RELATED WORKS

Biometric authentication devices are used for verifying users based on two basic modules: closed set module and open set module. The closed set module is used for identifying users who are stored in database while the users of the open set module are not stored in database. The open set module requires a wide range of parameters and variables for verifying users due to the high uncertainty. The closed set module is used as a low risk mechanism for securing locations using biometric features such as voice, face, fingerprint, electrocardiogram (ECG), and iris authentications [11].

Biometric authentication devices used in closed set module are also called access control devices. Different biometric authentication schemes and mechanisms have been proposed as security measures from spoofing attacks. Recent researches proposed efficient biometric authentication schemes based on RSA algorithm and smart cards. This scheme is vulnerable to smart key losing [12].

As a result, a biometric authentication scheme based on single and multi-server environment was presented for authenticating the key exchange process [13]. Secret key mechanisms are also implemented based on biometric authentication systems [14]. In that research, the biometric signal of the user is stored and converted to a secret key and a message. During the authentication process, the user signal is matched to the stored the biometric signal. The system retrieves the message and produces and estimation of the secret key.

A biometric authentication features for smartphone users were proposed in [15]. These features are based hand movement, orientation, and grasp (HMOG) to authenticate users continuously. The objective of this research is to detect both body movements by walking and hand motions. This biometric feature is vulnerable to man in behind attack that can monitor these movements and latter perform a spoofing attack. Existing biometric authentication schemes are based on maintaining face, fingerprint, and iris biometrics on server-side that can be compromised by external attackers or service providers. One of the recent researches for dealing with this issue was presented in [16]. This issue was handled using user-centric authentication that can allow users to encrypt their signatures using encryption schemes.

Another recent methodology that deals with the untrusted server issue was presented in [17]. In that research, a biometric-based authentication framework was proposed to secure biometric features during transmissions and at untrusted servers. Based on the proposed framework, the database administrator will be unable to extract user signatures due to the security coprocessor protocol that protects biometric data in both client and server side.

Previous research papers handle biometric authentication parameters especially using face recognition, fingerprints, and body movements. An advanced workflow for acquiring iris biometrics was proposed in [18]. This workflow explains different key factors that may affect the iris image quality. The workflow was divided into three main stages: acquisition process, iris analysis, and iris biometric authentication.

Recent biometric authentication researches tended to focus on building multi-model authentication systems and mechanisms to integrate more than one biometric feature for increasing the efficiency and robustness to spoofing attacks. As presented in [19], a multimodal biometric system was developed based on two authentication systems and two different fusion algorithms. The proposed multimodal system captures authentication features from both fingerprint and ECG data for increasing the security of biometric data.

Authentication of biometric data is also used in the Internet of Things (IoT) field. Different devices in smart buildings use portable sensors to track fitness and health and unlock smart vehicles and machines. Modern home appliances and smart vehicles are mainly adopted on biometric authentication data to allow users controlling the devices. One of the recent researches for applying Internet of Things (IoT) based biometric authentication was proposed in [20]. The proposed research introduced a user authentication algorithm based on three different biometric data for locking and unlocking smart devices. These biometric data are: step count behavioral pattern, heart rate psychological pattern, and hybrid calorie burn pattern.

Although authentication schemes provide provable security and reliability, all biometric authentication devices that are based on fingerprint, voice, face, iris, electrocardiogram (ECG), secret key, and tokens are vulnerable to several brute force and spoofing attacks. Smart cards and tokens are vulnerable smart card lost attacks [12]. Secret keys are vulnerable to offline guessing [21], [22] while smart cards are vulnerable to user impersonation attacks [23], [24], and [25]. Most remote authentication schemes (RAS) such as face recognition, fingerprint and iris authentication are vulnerable to user quantum attacks [26], [27], and [28] that can cause potential security threats.

Biometric authentication based on fingerprints can cause some subtle issues. As presented in [29], the authors have addressed an important hypothesis in that the fingerprint of the user can change due to the age and the time difference. A unique database of over 400K fingerprints has been created for fingers between 0 and 25 years and between 65 and 98 years with a time interval between samples of the same fingerprint from 0 to 7 years. The hypothesis proves that quality and the matching between the same fingerprint samples varies according to the user age. Fingerprint spoofing attacks are the most common type of attacks on biometric authentication ranging from molding to using 2D and 3D printing techniques to spoof authentic fingerprint [30]. Face detection mechanisms also suffer from spoofing attacks into which the attacker tries to perform illegal access to the system by presenting artificial biometric trait of an authorized user [31].

Many techniques have been proposed by researchers to tackle the automatic speech recognition using source visual information (i.e. lip-reading). Sum *et al.* [32] have extracted the lip contour using Active Shape Model (ASM), with the aid of fuzzy clustering analysis. They achieved a real time extraction from image sequence, and their approach was insensitive to position and size of lip contour. Matthews *et al.* [33] proposed an extraction of Lip-reading for visual speech recognition using three methods to obtain a sequence of lip contour for parameterizing lip image using hidden Markov models. Two of these methods are top-down approaches for the inner and Outer lip contours and derive lip-reading features from a principal component analysis of shape The third method is the bottom-up which uses a nonlinear scale space analysis to form features directly from the pixel intensity.

Hong *et al.* proposed an approach based on discrete cosine transform (DCT) for extracting visual lip-reading. They used principal component analysis (PCA) to reduce the dimensionality of DCT coefficients. They have proven that the combination of DCT and PCA efficiently improve the recognition accuracy. Kim [34] have used SURF as a local descriptors to generate feature vectors for face description. They used support vector machines as classifier within two layers, the first layer checks feature vectors image source (e.g. face or not) and the second layer localizes face components classifier of eye and mouth. The advantage of their approach is operating time, because there is no need for windows scanning procedure. Faubel *et al.* [35] improved the speech recognition performance by combining the audio-visual activity detection with microphone array processing techniques. They used robust face tracking system to provide possibility positions for each features by a bank of Kalman filters, and integrate this features with a Bayesian filtering. Siatras *et al.* [36] proposed a model based on variation of the intensity values of the mouth region by increased values of the number of pixels with low intensities through signal detection algorithms to determine lip activity. Komai *et al.* [37] proposed a method to extract the lip area automatically in different face directions, and converting the sideways lip figure into a frontal one using Active Appearance Models (AAM). They achieved an average accuracy of 77% for visual recognition rates with normalization of face direction, and 80.7% without normalization.

The authors of [38] have improved visual information by detecting each element (edge, cell) appears four times with different normalizations, including redundant visual information. They adopted linear SVM to improve performance from 84% to 89% at 10−4 false positives per window (FPPW). The performance of the speech recognizers has been improved using different techniques for classification face detection, including support vector machine and multilayer perceptron using the Haar classifier. The mouth area is calculated and analyzed, and the information coming from that region (the level of mouth openness) is passed to several machine learning algorithms which make decisions. The software detects speech with 60 to 75% average accuracy. Morade and Patnailk [39] has improved speech recognition through lip reading with ACM algorithm for localized and HMM, and use English numeric utterance data set to achieve performance from 77.8 to79.6 with 5 HMM.

As presented in [40], an improved Arabic audio/visual recognition system is proposed based on automatic generation of fine-grained phonetic transcriptions. The proposed solution is based on developing a set of language-dependent grapheme-to-allophone rules that have the ability to predict audio/visual recognition phonemes. Another recent research paper for Arabic audio/visual recognition was presented in [41]. In that research, a mobile application for learning Arabic reading was proposed. This research is based on text-independent audio/visual recognition that allows users to record their voice while reading. The voice is then evaluated and vocal feedback is provided according the accuracy of the user reading.

One of the recent research papers based on hidden Markov model (HMM) was presented in [42]. In that research, a machine learning-based technique for activity label analysis is proposed. The technique conceptualizes activity label analysis as a tagging task based on a Hidden Markov Model. A lip analysis methodology was presented in [1] for developing a fully automated data collection from TV broadcasts. A two-stream convolutional neural network is proposed to learn the relationship between the sound and the mouth motions from unlabeled datasets. The convolutional networks are trained to effectively learn and recognize hundreds of words from this large-scale dataset. Different biometric systems are used to test lip motions and voice recognition based on different data sets. As shown in [43], the data set is collected from different video sources in order to recognize voice and lip movements.

A behavior biometric based authentication is another type of authentication based on behavior recognition. As presented in [44], the authors used a thumbprint authentication model for creating secret knocks between users. This model achieved accuracy between 85% and 91% with 10 repetitions for each thumbprint. The authors of [45] proposed a model for applying a pressure on screen with a predefined duration between each pressing process. This method achieved accuracy of 95.9% for authorizing users correctly with 20 sample repetitions. In [46], proposed a model for unlocking mobiles using finger velocity and pressing time with an accuracy of 95.23% with training samples from 15 to 25. As presented in [47], a model was built for authenticating users using their unique walking variations. The accuracy of authenticating users achieved 90.5% with 40 repetition times. In [48], a silent key system using mouth motions was proposed.

The silent key system achieved accuracy between 70% and 83.1% with training samples between 5 and 9 times. In our proposed lip model, a unique feature for authenticating

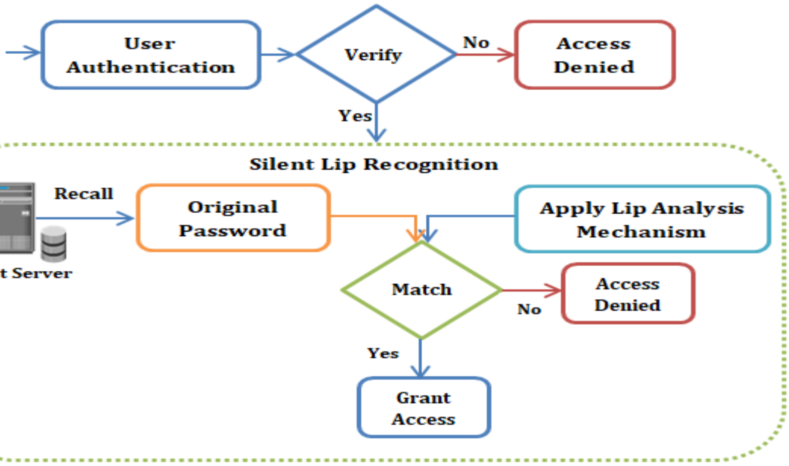


FIGURE 1. Dual authentication framework.

users using a dual authentication process is proposed. All predefined proposed models are developed with a user-dependent model for training users' samples with different repetition times. In our proposed model, a user independent model is applied by capturing face features as a first authenticating layer and then extracts the lip movements at the same time.

By comparing our proposed silent lip password with recent lip models [48], [49], and [50], our proposed lip model presents better performance. The authors of [48] proposed a silent key using ultrasonic with pre-stored secret words. The accuracy varied between 70% and 83.1% with multiple training times. As presented in [49], lip reading is captured using acoustic sensing with different phonemes from 1 to 10. This model achieved 83.7% accuracy for identifying mouth states while our proposed model achieved 96.2% accuracy in detecting lip features. As presented in [50], a user authentication process is applied for detecting lip readings.

This method is speaker dependent. Each time a new user is added to the system, the model should train detected lip features. The equal rate error for word recognition for this model was 26.9%. Our proposed model is speaker independent as face features are captured in the first authentication layer and then the model extracts lip movements at the same time.

The motivation behind this work is three folds; the first of which is to enhance the results obtained from the previous works illustrated in this section; the second is to apply the proposed methodology on Arabic language, and the third is to introduce a visual dataset for Arabic digits collected from real persons.

III. DUAL AUTHENTICATION FRAMEWORK

Authentication process is mainly based on three major factors: something you know (such as password), something you have (such as smartcard or security token), and something you are (such as biometric face recognition or fingerprint authentication).

The proposed dual authentication framework combines two authentication factors: something you are and something you know. The first authentication will use something you are such as face and fingerprint while the second authentication will use something you know by providing a silent password that overcomes most of written password attacks.

As presented in Figure 1, the face authentication method will be used as a first authentication layer in the dual authentication framework. The proposed silent password approach depends mainly on capturing face features and then extracts the lip movements at the same time.

This is more efficient, usable, and secure as the input of the user authentication process is performed once during lip movement with silent password. The captured frames for the face are used for extracting lip movements for improving the authentication process by providing faster and less resource consuming mechanism. This is suitable for limited resources devices such as mobile devices and tablet.

The original stored user lip password is recalled from the host server and is matched with the entered lip password. If both lip passwords are matched, the user is granted access to the system; otherwise an access denies alarm is raised.

The dual authentication framework strengthens the biometric face authentication due to different limitations in face recognition such as spoofing attacks [51]. The efforts for solving spoofing attacks especially on fingerprint and face recognition are still limited [12], and [30] due to the availability of user's image on public websites. The intruder can download the image from social media or capturing the image by high resolution camera. The intruder can use the captured image to spoof the legitimate user image to penetrate into the system. In addition, the face recognition techniques have some restrictions like false acceptance error that can falsely identify intruders as authorized users. Therefore, it was a challenging task to detect the image whether the input face is from a live person or from a photograph.

By combining face authentication process with a silent password, the robustness of the authentication framework

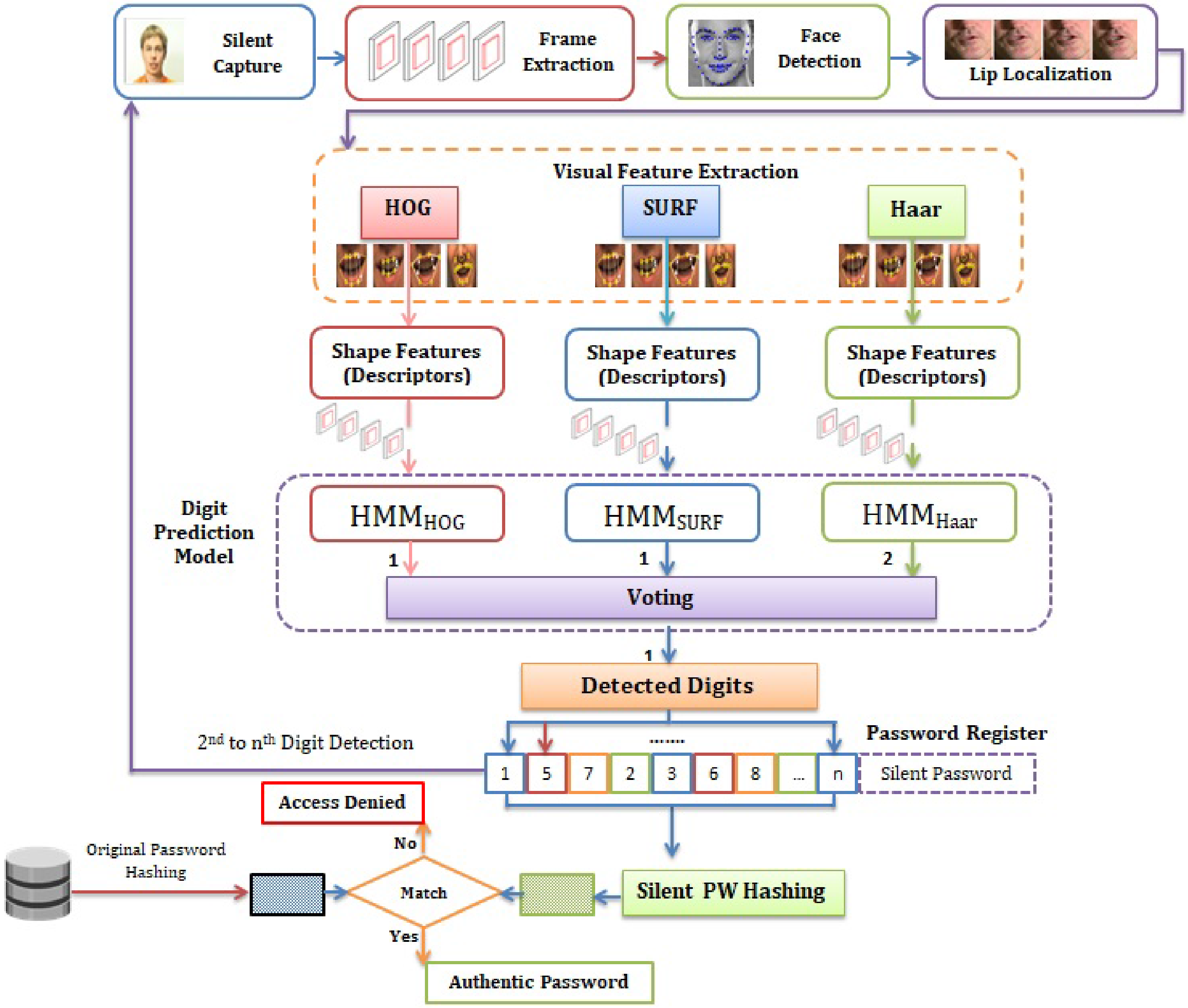


FIGURE 2. Silent password recognition framework.

will increase. The silent password is the replacement for normal written password that always comes as second stage after user identification and verification. So, a unique feature of face recognition technique is used as a first authentication step for users. The second authentication layer uses a silent lip password to improve the security and efficiency of identifying users.

IV. PROPOSED SILENT PASSWORD RECOGNITION FRAMEWORK

As shown in Figure 2, a methodology for silent password using automatic face recognition based on visual perception is proposed. The model consists basically of 5 layers, namely, input layer, preprocessing, visual feature extraction, digit prediction model, and finally the password matching layer.

- Layer 1: user in front of video camera start speaking first number such as ''one'' using silent password by moving lip without voice
- Layer 2: face detection and lip localization are applied on extracted sequence of frames.

In this work, we have utilized Viola-Jones algorithm [52] for both face detection and lip localization. Initially, input video is converted to 4 equal-sized frames that localize the face region, and then face detection algorithm is applied to extract and isolate the face from the background.

Finally, the lip localization is applied based on specific threshold to extract the region around the mouth. The output of this layer is 4 equal-sized frames per video, which only localizes the mouth region.

- Layer 3: feature extraction from mouth lip movement using three visual features extraction modules SURF, HoG, and Haar. As presented in Figure 3, valid points of lip features were detected using SURF, Haar and HoG features respectively.
- Layer 4: each features set extracted from three visual features extraction modules SURF, HoG, and Haar are

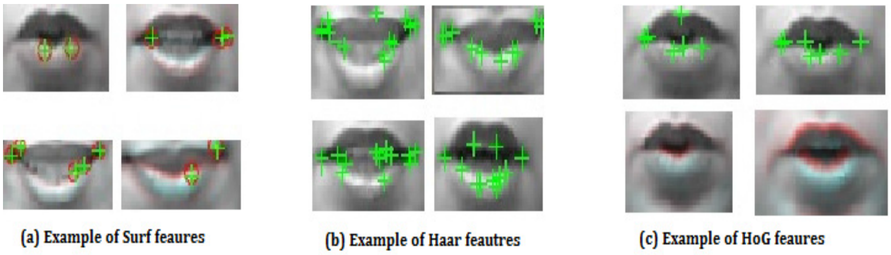

(a) Example of Surf feaures
(b) Example of Haar feautres
(c) Example of HoG feaures

FIGURE 3. Different lip features detection examples.

passed to different HMM model for independently predict the spoken digits, then output from different HMM model voted to output the predicted digits which is the "One" digit in this example.

The algorithm computes the likelihood of a particular observation given the first observation in the sequence by summing over the probabilities of all possible hidden state paths that could generate such observation [39], and finally multiplies the whole conditional probabilities altogether as presented in formula (1):

$$\mathrm{P(O|Q)} = \pi_{\mathrm{i=1}}^{\mathrm{T}}\mathrm{P(o_i|q_i)} \tag{1}$$

where $\mathrm{O} = \mathrm{O_0, O_1, \ldots\ldots., O_T}$, represents a set of possible observation sequence, $\mathrm{Q} = \mathrm{q_0, q_1, \ldots\ldots., q_T}$ represents a set of hidden state sequence, and P (O | Q) is the likelihood of the observation given some sequence at time T.

The digit prediction model layer produces three different learning models, with each one corresponds respectively to Surf, Haar, and HoG. The final classification (digit detection) result is obtained by feeding the generated models to a voting scheme to vote over the odd number of the inputs, then the predicted digit is fed to password matching layer to keep track of predicted digits and compose whole password.

- Layer 5: predicted digit is added to password register and the recognition process repeats over (layers 1 - 4) until password digits are totally consumed as shown in formula (2). After which the detected password is compared with stored hashed password.

$$C_i = b_{i1} \otimes b_{i2} \otimes \ldots . \otimes b_{im} \tag{2}$$

where:

$\boldsymbol{C_i}$ presents the hash code.

$\boldsymbol{b_i}$ is the i^{th} block of the password.

$\boldsymbol{m}$ is the number of n blocks in the input password.

The hash code components are formulated using XOR operation ⊗. The resulting hash code $\boldsymbol{C_i}$ of the silent password recognition framework is matched with the hash code $\boldsymbol{C_j}$ produced by the database server. If both hash codes are matched, the password will be authentic and the user will be verified, otherwise an access denied alarm will be raised.

V. DATA PREPROCESSING

In this section, the implementation of the case study is illustrated on the proposed model. The implementation targets the Arabic visual figures recognition, with the aid of Matlab image processing toolbox.

Initially, real video dataset has been generated for frontal visual face that consists of 20 samples for training (13 males and 7 females).

Each sample is expanded to 10 Arabic numerical words, and each numeric word utterance is repeated 10 different times for the same speaker. This yields a total dataset of 2000 training records for the ten Arabic digits, another 2000 records with different distribution of Arabic digits generated as test dataset as shown in Table 1 based on different number of samples for Arabic words from 0 to 9.

Moreover, the dataset was generated with random noise and different background for each speaker to simulate real life situation of word recognition. A laptop camera was used for the recording mission, and the generated video format was "avi" with resolution of (640 × 480) at 30 frames/second and average video length of 30 second.

VI. FACE AND LIP DETECTION PROCESS

For face detection, images are captured every 5 frames from different videos in the dataset, and Matlab image processing tool box was used with the implementation of Viola-jones algorithms. Region of interest (ROI) mask was used to crop faces from image, so as to reduce the errors of lip detection by focusing on the face only. For lip detection, the Haar corner detection technique is used to extract 8 points from the lips boundary. The contour extraction of the lip is obtained by finding the optimum partition of a given RGB image into lip and non-lip regions based on intensity and color.

As presented in Figure 4 a geometric model for extracting lip shapes is based on two curves y_1 and y_2 with three points that represent the maximum and minimum points of the lips corner at x and y directions. The first point indicates the vertical distance h_2 from the upper lip contour to the horizontal axis x. The second point indicates the vertical distance h_1 from the lower lip contour to the horizontal axis x. The third point indicates the horizontal distance w from the corner of the lip to the center of the lip.

The lip contour extraction of curves is based on two formulas (3) and (4). Formula (3) presents the lip curve y_1 while formula (4) presents the lip curve y_2.

$$y_1 = h_1\left(\left(\frac{x - ly_1}{w}\right)^2\right)^{1+\partial^2} - h_1 \tag{3}$$

TABLE 1. Arabic LIP samples.

English Number	0	1	2	3	4	5	6	7	8	9
Arabic Word	Sefr	Waahed	Ethnan	Thlatha	Arbaa	Khamsa	Setah	Saabaa	Thamania	Tesaa
No. of Samples	100	100	100	300	200	400	200	200	200	200

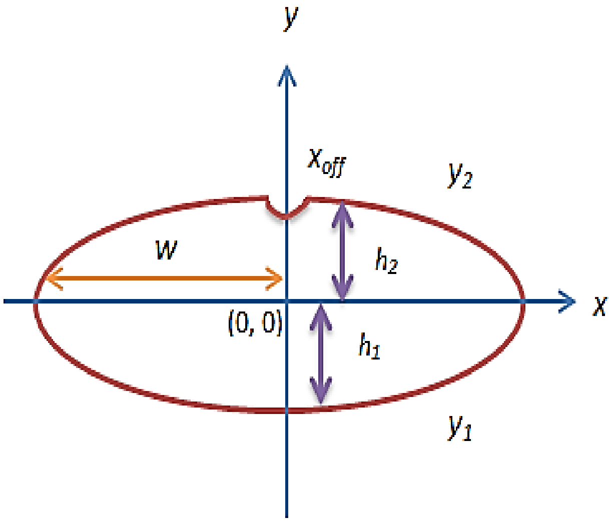


FIGURE 4. Lip model parameters.

where:

$\boldsymbol{x} \in [-\boldsymbol{w}, \boldsymbol{w}]$, at the origin point of the lip (0, 0)

$\boldsymbol{ly_1}$, refers to the skewness of the lip shape

$\boldsymbol{\partial}$, refers to the deviation of the lip curve y_1

$$y_2 = \frac{-h_2}{\left(w - x_{off}\right)^2}\left(|x - ly_2| - x_{off}\right)^2 + h_2 \tag{4}$$

where:

$\boldsymbol{ly_2}$, refers to the skewness of the lip shape

$\boldsymbol{x_{off}}$, refers to the deflection of the upper lip contour with vertical axis y

Figure 5 illustrates a visual dataset for lips with different shapes. These shapes have different movements whether the spoken words were silent or not.

VII. LIP FEATURE EXTRACTION

Lip feature extraction in the proposed model is based on 3 feature extraction and description techniques, namely, Surf, Haar and Hog. In the three methods, the feature parameters are given in an array of vectors with variable length (based on the employed technique), that represents the descriptors of lip contour. The Surf feature vector is obtained by dividing the lip rectangle into 4 main sub-regions that represent the key points of the lip in the x and y directions. For each sub-region, Surf calculates the gradient of the 4 corner in 4 directions to produce a total of 16 dimensions' feature vector. For all 4 sub-regions, a total of 64 feature vector is produced whole the whole lip rectangle.

For HoG feature descriptor, the technique works by dividing the image into small connected regions called cells, and for each cell compute a histogram of gradient directions or edge orientations for the pixels within the cell and discretizing each cell into angular bins according to the gradient orientation. Moreover, adjacent cells are grouped together in spatial regions to form blocks. The grouping of cells into a block is the basis for grouping and normalization of histograms. To obtain feature description of lip region, HoG normalizes the group of histograms represented in the block histogram. The set of these block histograms represents the descriptor.

Haar-like feature extract and describe lip region by considering adjacent rectangular regions at a specific location in a detection window, then sums up the pixel intensities in each region using integral image, and calculates the feature value, which is the difference between these sums. The feature value is then compared to a learned threshold that separates non-objects from objects (e.g. lips).

The lip feature extraction process is based on three layers: machine learning classifier, hidden Marcov model, and cumulative voting formula as presented in the following subsections.

A. MACHINE LEARNING CLASSIFIER

In order to learn from different Arabic figures utterance, a classifier is needed to differentiate between different figures based on the features descriptors produced from the aforementioned techniques.

In this paper, HMM was used as a machine learner classifier which was reported as an efficient classification algorithm in the field of automatic speech recognition. Such HMM is a statistical model of a process consisting of two random variables O and Y, which change their state sequentially.

The variable Y with states $(Y_1, Y_2, \ldots.., Y_n)$ called the "hidden variable", since its state is not directly observable. The state of Y changes sequentially based on its current state and does not change in time. The variable O with states $(O_1, O_2, \ldots.., O_n)$ is called the "observable variable", since its state can be directly observed. O does not have a Markov Property, but its state probability depends statically on the current state of Y.

It's worthy to mention that there are different parameters that control the efficiency of the classification results of HMM, one of which is the hidden states. Based on such classifier, three HHM models have been produced for each feature descriptor. The three models have been grouped in a voting scheme to enhance the final result of the classification.

FIGURE 5. Sample real dataset for different lips.

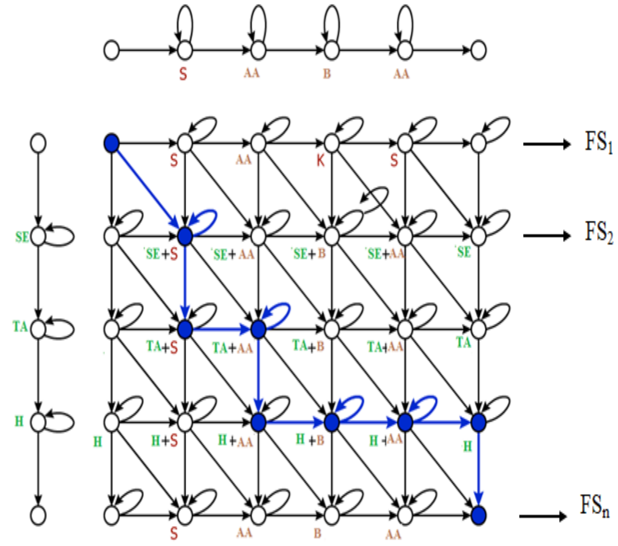


FIGURE 6. Search graph for silent passwords.

B. HIDDEN MARKOV MODELS

The process of applying Hidden Markov Model (HMM) for predicting lip movements and generating silent passwords is based on identifying the current state of lip movement that will affect future states based on the following formula:

$$\forall t \in T \text{ such that } \pi_t \subset \pi \tag{5}$$

where:

$\boldsymbol{t}$: is the time frame

$\boldsymbol{T}$: is the overall state time

$\boldsymbol{\pi_t}$: is the current state of the parse

$\boldsymbol{\pi}$: are the parsing states

As presented in Figure 6, for a silent password of "Setah" and "Saabaa" which mean 6 and 7 respectively, the frame sequence FS_2 is based on the current frame sequence FS_1 until the final lip movement ends.

To explain the previous figure, given the frame sequence $FS = fs_1||fs_2||\ldots fs_n$ and the parsing time $\pi = \pi_1||\pi_2||\ldots||\pi_n$, the likelihood for each frame sequence fs_i with the following frame sequence fs_k at each parsing time π_t based on state k with a fixed position i is identified based on the following formula:

$$\sum_{i=1}^{\mathbf{n}} P(fs_i, \pi_t) \quad \forall k \in K \tag{6}$$

where P is the probability for each frame sequence fs_i at parsing time π_t is calculated by summing all possible ways from position i to position n at each k state. The initial and likelihood frame states are defined based on the following definitions:

Definition 1 (Initial Frame State): for each state k at position i; the initial frame state is defined based on the following formula:

$$k\ (i) = max_{(\pi 1 \ldots \pi i-1)} P(fs_1 \ldots fs_{i-1}, \pi 1 \ldots \pi_{i-1}, fs_i, \pi_i) \tag{7}$$

where, all subsequent frames are calculated to determine the probability for each frame.

Definition 2 (Likelihood Frame State): the most likelihood frame is determined by calculating the maximum value for frame sequence fs_i based on the following formula:

$$k\ (i+1) = max_{(\pi 1 \ldots \pi i-1)} P(fs_1 \ldots fs_i, \pi 1 \ldots \pi_i, fs_{i+1}, \pi_{i+1}) \tag{8}$$

Based on the previous definitions, all possible frame sequences are calculated and the maximum likelihood frame is selected based on the previous or current frame.

C. CUMULATIVE VOTING FORMULAS

In this paper, a cumulative voting algorithm has been exploited. Such algorithm is a mathematical method for computing optimal result for recognition system to maximize the accuracy. This voting has been used by grouping result of recognition through different types of feature extraction as a method to find the best result for speech recognition.

The mathematical equation that represents the cumulative voting algorithm to elect a majority of Figures is presented in formula (9):

$$X = \frac{SN}{D+1} + 1 \tag{9}$$

where

X: number of methods of feature extraction needed to elect a given number of figures

S: Total Number of feature extraction methods to Vote at model

D: Number of times votes want to elect

N: number of figures needed

VIII. THREAT MODEL

The objective of the attacker in a secure system is to penetrate the security layers to disclose information, modify data, or to disrupt services. The following are a set of assumptions for the capabilities of an attacker that may perform to compromise any security system.

- Key losing: This is an attack in which the authorized user may lose his secret or private key and as a result the attacker can compromise the system with the authority of a normal user.
- Man-In-Behind attack: In this type of attack, the attacker stands behind the authorized user. This attacker can see the security key of the authorized user or eavesdropping on the secret password if the system uses voice recognition techniques.
- Brute Force attack: In this attack, the attacker tries every possible key until he obtains an intelligible secret key.
- Spoofing attack: This type of attack depends on counterfeiting data by attackers to gain unauthorized access to the system
- Offline Guessing attack: This type of attack is effective if the guessing of passwords is occurred automatically until the secret key is verified.
- User Impersonation attack: In this attack, the adversary uses the credentials of a legitimate user during user authentication process.
- User Quantum Attack: This type of attacks tries to break through cryptographic algorithms using quantum computers.

Table 2 summarizes the threats and vulnerabilities for the proposed authentication framework and related works. Each one of the explained papers is checked whether it can defend against the presented threats and attacks or not or the proposed mechanism is not applicable to the security feature.

Based on the related works presented in Table 2, the proposed silent lip authentication framework is effective against the aforementioned attacks as indicated below:

- ✓ Key losing: the proposed model does not depend on a predefined secret key or security token that can be lost.
- ✓ Man-in-Behind attack: the proposed model is effective due to the inability of the intruder to detect and predict silent lip password.
- ✓ Brute force attacks: the proposed model is effective due to the complexity of predicting silent lip password. Moreover, the proposed silent lip framework is designed to a limited number of failed attempts. So, the user will not be able to perform a brute force attack.
- ✓ Spoofing attacks: the proposed model is effective as the intruder will not be able to counterfeit the silent lip of the authorized user.
- ✓ The proposed model is also effective against user impersonation attacks as the intruder cannot compromise user credentials due to the using of dual authentication mechanism of face and silent lip authentication.
- ✓ Offline guessing attacks: the proposed model is effective as the model combine silent password with biometric face authentication that prevents offline guessing attacks due to the need of person's face and the dynamic lip during speaking of silent password.
- ✓ For quantum attacks, it is not applicable in our proposed silent lip framework as the framework does not support post quantum cryptography such as public key algorithm that can defend against attacks by quantum computers. In our proposed silent lip authentication, there is no need to add a time complexity overhead over the developed framework for adding a third layer of security as the dual authentication framework achieved better accuracy in authenticating users.

IX. EXPERIMENTAL RESULTS

In this section, the evaluation of Haar, HoG, Surf, and voting mechanism with a comparison of their performance is conducted.

As presented in Figure 7, the visual speech recognition accuracy for different Hidden Markov Model states is stated. The value of N refers to the number of features or figures needed to detect the silent lip password. As shown, there is a direct correlation between increasing the number of N features and the measured accuracy of HMM states.

For N=2, the HoG, SURF, Haar, and voting mechanism achieved 88.5%, 82.2%, 84.8%, and 91.3% respectively.

For N=4, the HoG, SURF, Haar, and voting mechanism achieved 91.3%, 85.2%, 89.9%, and 93.9% respectively.

For N=6, the HoG, SURF, Haar, and voting mechanism achieved 92.9%, 89.4%, 91.4% and 95.8% respectively.

For N=8, the HoG, SURF, Haar, and voting mechanism achieved 93.9%, 92.2%, 92.9%, and 95.6% respectively.

Finally, for N=10, the HoG, SURF, Haar, and voting mechanism achieved 94.9%, 92.2%, 93.2%, and 96.2% respectively.

As noticed the voting mechanism achieved the highest accuracy for N=10 with 96.2% when compared to the HoG, SURF, and Haar algorithms due to the digit prediction model of voting process that was applied on the silent password recognition framework presented in Figure 2. The second recorded accuracy was the HoG algorithm with 94.9% for N=10.

The proposed model is tested using test dataset for Arabic silent passwords from 0 to 9 with different evaluation parameters such as: accuracy, true positive rate (TPR), false negative rate (FNR), true negative rate (TNR), precision, recall, and F-measure.

As presented in Figure 8, a confusion matrix of 10∗10 Arabic samples is applied for recording the accuracy of the voting mechanism, where the row presents the actual sample values while the column presents the predicted values.

As shown in formula (10), the overall accuracy recorded is based on the correct classified records over all records.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{10}$$

TABLE 2. Security gap analysis.

Reference	Key Losing	Man-in-behind-Attack	Brute Force Attack	Spoofing Attacks	Offline Guessing	User Impersonation Attacks	User Quantum Attacks
[12]	×	×	√	√	√	√	NA
[13]	√	×	×	×	√	√	√
[14]	×	×	×	×	√	√	NA
[15]	NA	×	×	√	√	×	NA
[16]	×	NA	√	×	√	×	NA
[17]	√	NA	√	×	√	√	×
[18]	NA	√	√	×	NA	×	NA
[19]	NA	NA	×	×	√	×	NA
[20]	NA	NA	×	×	√	×	NA
[21]	×	NA	√	√	×	√	√
[22]	×	NA	×	×	×	√	×
[23]	×	NA	√	×	×	√	√
[24]	√	NA	×	×	√	√	√
[25]	√	×	√	×	√	√	√
[26]	√	NA	×	×	×	×	√
[27]	√	NA	×	×	×	×	√
[28]	√	NA	×	×	√	×	√
[44]	NA	×	×	×	×	×	NA
[45]	NA	×	×	×	×	×	NA
[46]	NA	×	×	×	×	×	NA
[47]	NA	×	×	×	×	×	NA
[48]	NA	√	×	√	√	√	NA
[49]	NA	√	√	√	√	√	NA
[50]	NA	√	√	√	√	√	NA
Proposed Model	NA	√	√	√	√	√	NA

Figure 9 presents the accuracy of Arabic silent passwords from 0 “Sefr” to 9 “Tesaa”. As shown, the highest accuracy recorded was the word “Sefr” which is 0 with 99.8% while the lowest accuracy was the word “Thlatha” and “Setah” which are 3 and 6 respectively.

The precision and recall of the confusion matrix were conducted using formula (11) and (12) as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{11}$$

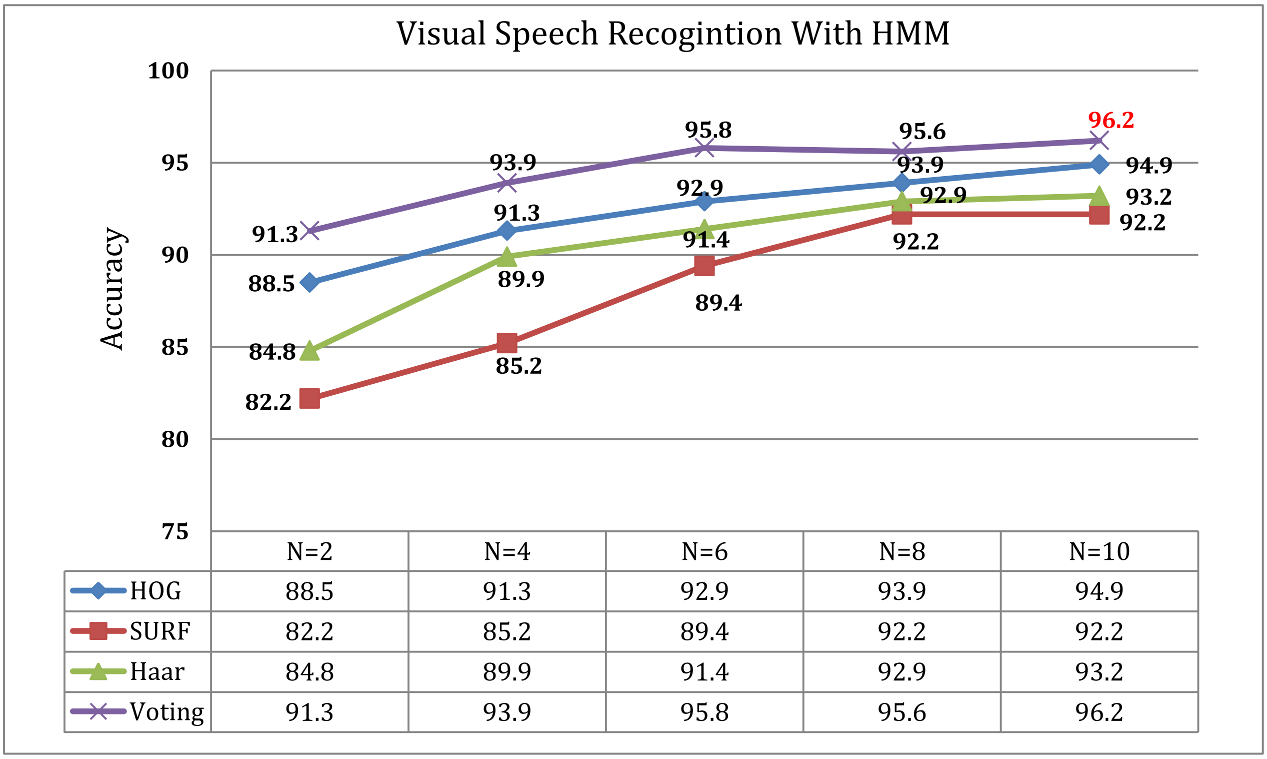
$$Recall = \frac{TP}{(TP + FN)} \tag{12}$$

As presented in Figure 10, the precision and recall for the numbers from 0 to 9 showed that highest precision and recall were 98.5% for the number 5 which is “Khamsa”. The second highest precision recorded 98% for the numbers 0 and 1 which are “Sefr” and “Waahed” respectively; while the second highest recall recorded 98% for the number 0 “Sefr” only. The lowest precision and recall were 93% for the number 8 “Thamania”.

The F1 Measure of the voting confusion matrix was conducted using formula (13) as follows:

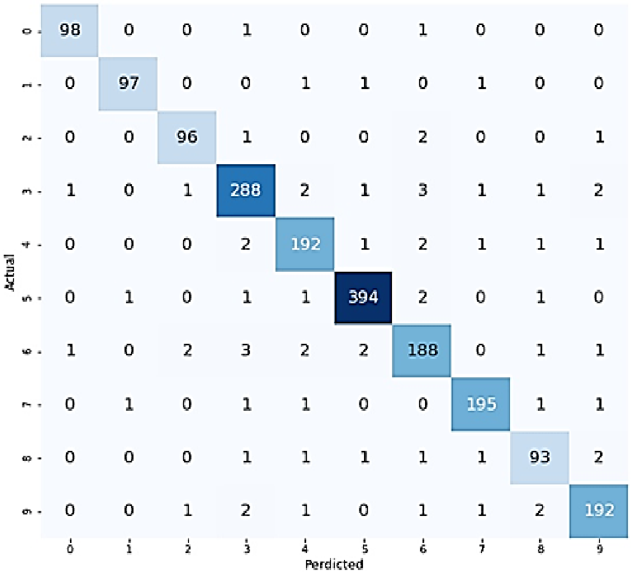
$$F1Measure = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{13}$$

As presented in Figure 11, the highest F1 measure recorded 98.5% for the number 5 “Khamsa” while the lowest F1 measure recorded 93% for the number 8 “Thamania”.



	N=2	N=4	N=6	N=8	N=10
HOG	88.5	91.3	92.9	93.9	94.9
SURF	82.2	85.2	89.4	92.2	92.2
Haar	84.8	89.9	91.4	92.9	93.2
Voting	91.3	93.9	95.8	95.6	96.2

FIGURE 7. Hidden Markov model for HoG, SURF, Haar, and voting mechanism.



	0	1	2	3	4	5	6	7	8	9
0	98	0	0	1	0	0	1	0	0	0
1	0	97	0	0	1	1	0	1	0	0
2	0	0	96	1	0	0	2	0	0	1
3	1	0	1	288	2	1	3	1	1	2
4	0	0	0	2	192	1	2	1	1	1
5	0	1	0	1	1	394	2	0	1	0
6	1	0	2	3	2	2	188	0	1	1
7	0	1	0	1	1	0	0	195	1	1
8	0	0	0	1	1	1	1	1	93	2
9	0	0	1	2	1	0	1	1	2	192

FIGURE 8. Confusion matrix for voting model.

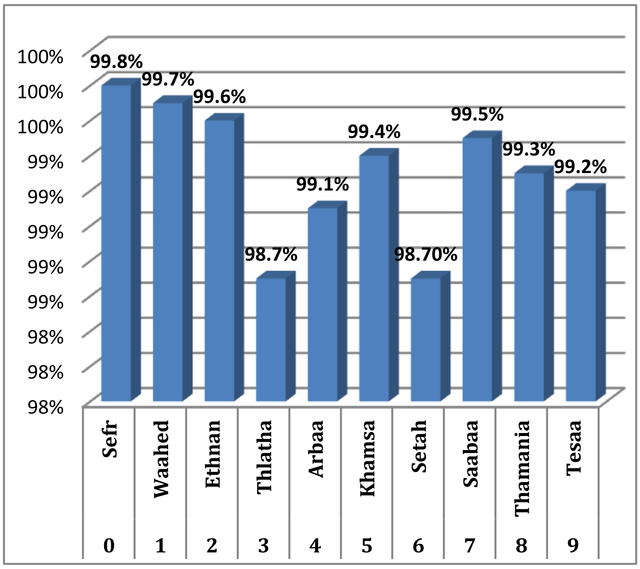


FIGURE 9. Overall word accuracy.

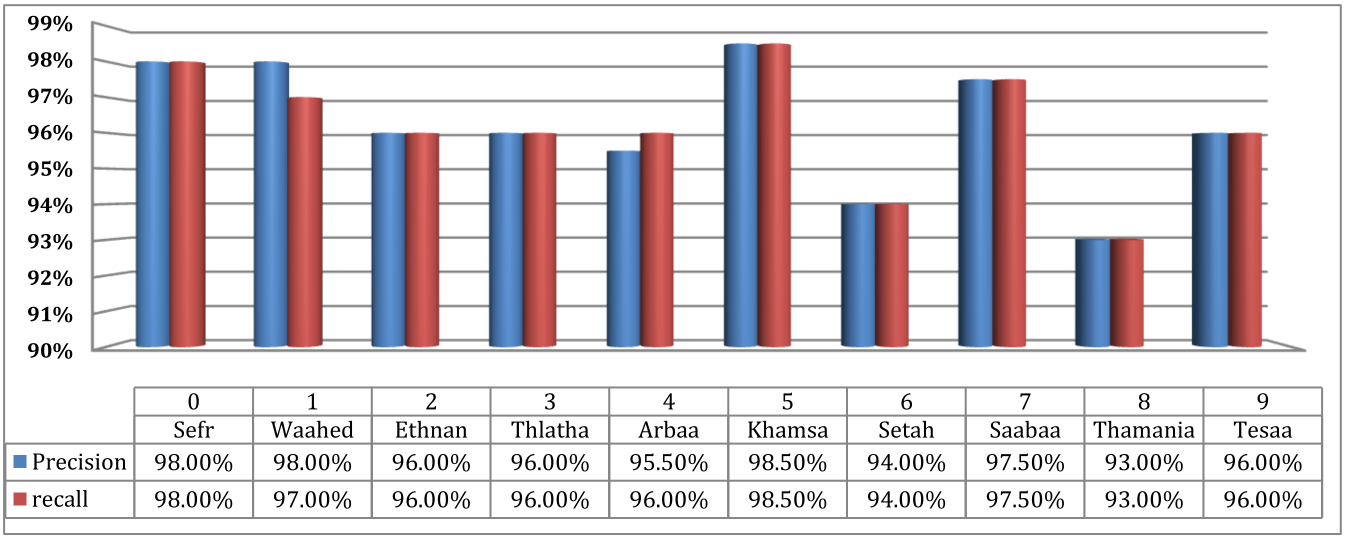
The false positive rate (FPR) of the voting confusion matrix refers to the percent of incorrectly identified Arabic figures. The FPR is calculated using formula (14) as follows:

$$FPR = \frac{FP}{(FP + TN)} \tag{14}$$

As presented in Figure 12, the highest false positive rate recorded 0.8% for the number 3 which is "Thlatha" while the lowest false positive rate recorded 0.1% for both 0 and 1 which are "Sefr" and "Waahed" respectively.

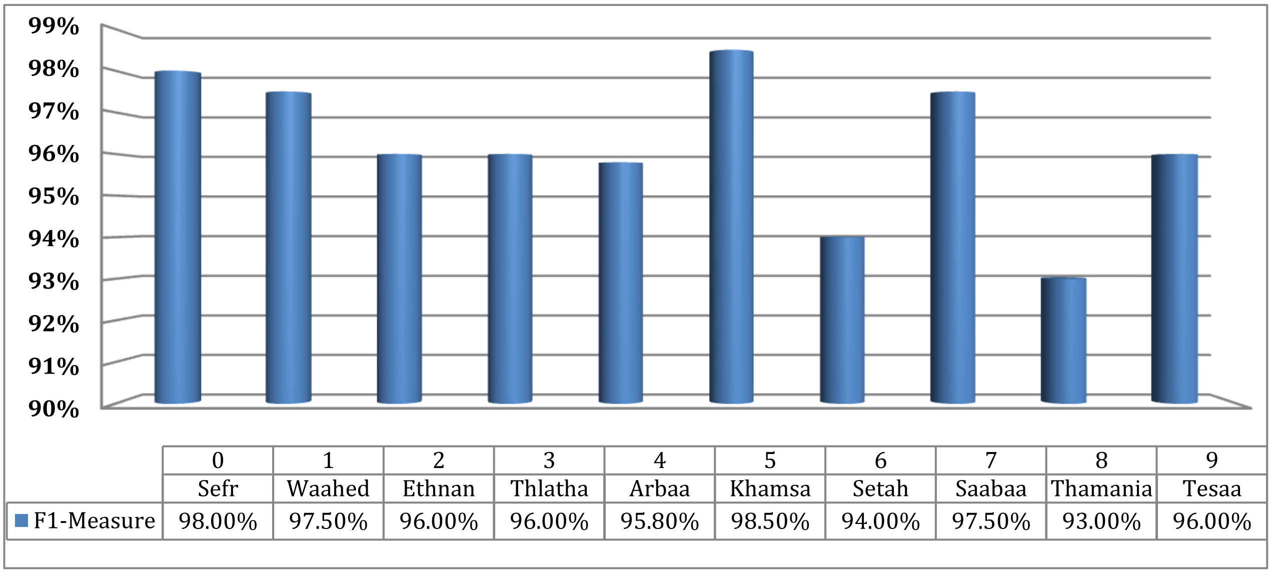
The false negative rate (FNR) of the voting confusion matrix refers to the proportion of positive words that have been detected as negative. The FNR is calculated using formula (15) as follows:

$$FNR = \frac{FN}{(TP + FN)} \tag{15}$$



	0 Sefr	1 Waahed	2 Ethnan	3 Thlatha	4 Arbaa	5 Khamsa	6 Setah	7 Saabaa	8 Thamania	9 Tesaa
🟦 Precision	98.00%	98.00%	96.00%	96.00%	95.50%	98.50%	94.00%	97.50%	93.00%	96.00%
🟥 recall	98.00%	97.00%	96.00%	96.00%	96.00%	98.50%	94.00%	97.50%	93.00%	96.00%

FIGURE 10. Precision and recall for voting model.



	0 Sefr	1 Waahed	2 Ethnan	3 Thlatha	4 Arbaa	5 Khamsa	6 Setah	7 Saabaa	8 Thamania	9 Tesaa
🟦 F1-Measure	98.00%	97.50%	96.00%	96.00%	95.80%	98.50%	94.00%	97.50%	93.00%	96.00%

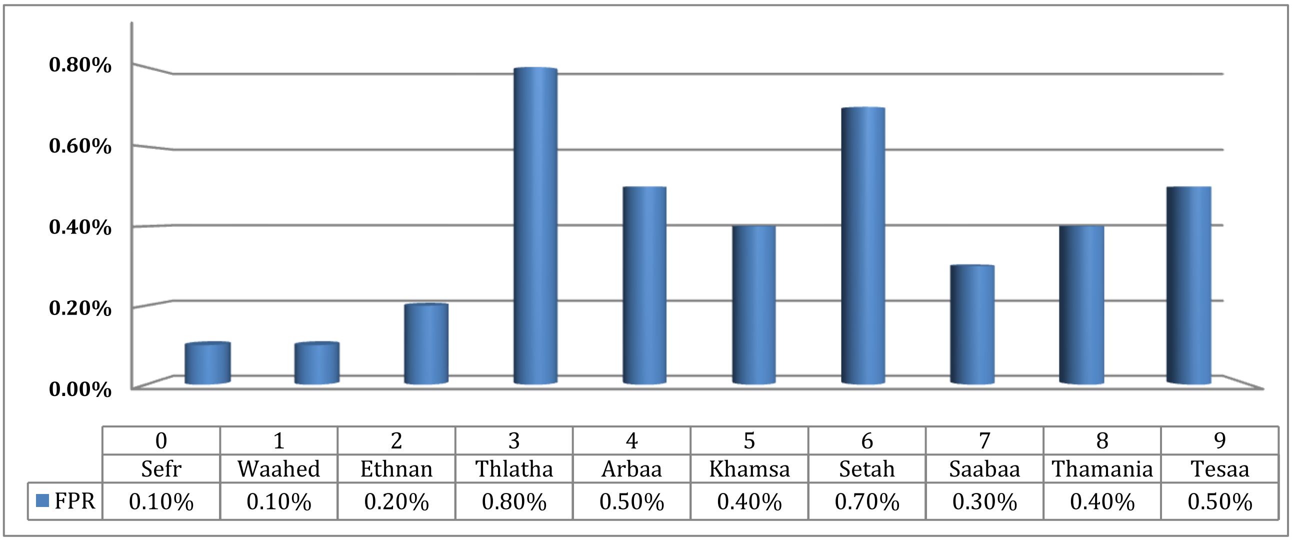
FIGURE 11. F1 measure for voting model.

As presented in Figure 13, the highest false negative rate recorded 7% for the number 8 which is "Thamania" while the lowest false negative rate recorded 2% for the number 0 that is "Sefr".

X. COMPARATIVE MODELS

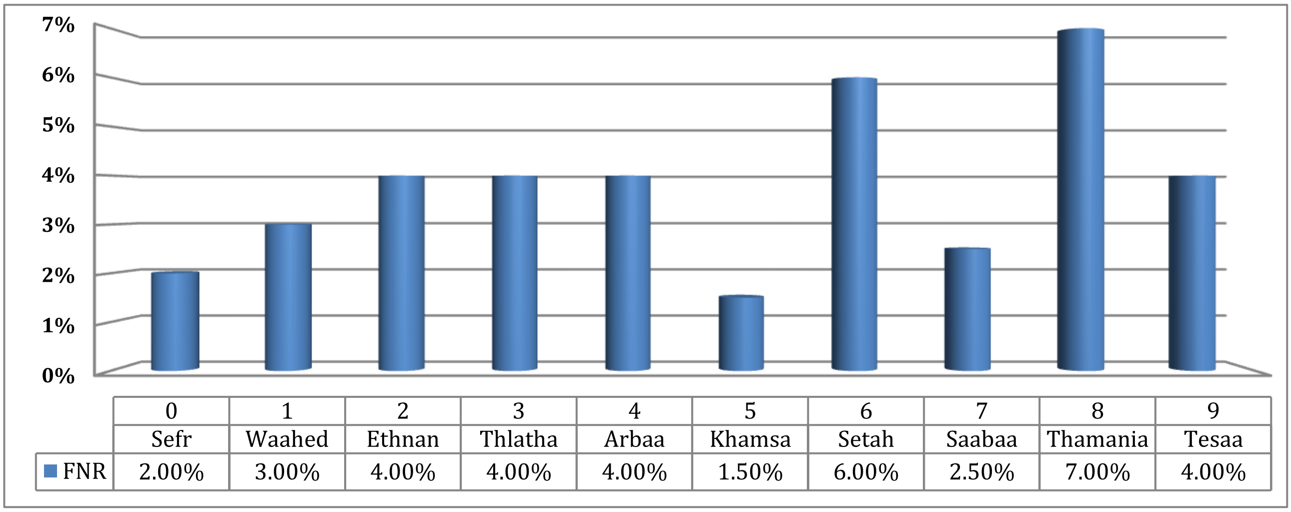
In the proposed approach, the HMM model is used as it is much simpler than the LSTM. The proposed model relies on the assumption that the state transitions depend mainly on calculating all possible frame sequences and the maximum likelihood frame is selected based on the previous or current frame, as presented in equations (7) and (8). So like always, these assumptions are valid. As such, the proposed model shows better performance.

The LSTM may perform better if a very large dataset is available, because the inherent LSTM mechanism for learning patterns can make a better use of big data. But LSTM can be difficult to get working, especially when the dataset is small as in our situation for Arabic dataset. One of the advantages of HMM over LSTM is that it can be deployed on limited memory and processing resource devices such as mobile devices for authentication and mobile screen-unlock. This is a much simpler model for authorizing users compared to the LSTM that might consume most of memory and processing power, as mobile unlock happen many time per day. Also the HMM training algorithms are very popular, simple, and computationally feasible to use.



	0 Sefr	1 Waahed	2 Ethnan	3 Thlatha	4 Arbaa	5 Khamsa	6 Setah	7 Saabaa	8 Thamania	9 Tesaa
■ FPR	0.10%	0.10%	0.20%	0.80%	0.50%	0.40%	0.70%	0.30%	0.40%	0.50%

FIGURE 12. FPR for voting model.



	0 Sefr	1 Waahed	2 Ethnan	3 Thlatha	4 Arbaa	5 Khamsa	6 Setah	7 Saabaa	8 Thamania	9 Tesaa
■ FNR	2.00%	3.00%	4.00%	4.00%	4.00%	1.50%	6.00%	2.50%	7.00%	4.00%

FIGURE 13. FNR for voting model.

The Hidden Markov Model (HMM) is used on the proposed silent lip recognition framework instead of deep learning LSTM because deep learning requires large amount of data while the Arabic dataset used in the silent lip recognition is limited.

As presented in [53], and [54], the HMM achieved better performance than LSTM for few training sequences per class. In addition, the HMM can perform better performance on early learning behavior.

By applying HMM on our proposed silent lip recognition framework, better prediction of action hypothesis will lead to better prediction of lip frames that can be applicable on human computer interaction applications.

The average recognition rate for extracting voice and image features are conducted by mixing the dictionary or codebook image/phonemes. As presented in Table 3, the experimental results of [40] were conducted in English language using Hidden Markov Model (HMM) with 3 and 5 hidden shapes on two different databases.

The recognition rate using 3 hidden shapes achieved 66.3% and 64.7% on Cuave and In-house database respectively while the 5 hidden shapes achieved 78.33% and 76.6% on Cuave and In-house database respectively. As presented in [37], the experimental results were conducted in Japanese language for 216 words using single and multiple regression and achieved 78.67% and 79.56% respectively.

As shown in [55], a lip reading mechanism is presented for Japanese and Arabic language with nine Arabic sentences but the performance for Arabic dataset recorded 79.5%.

The authors of the paper [56] perform their experimental results based on two languages: Arabic and Japanese, where the Arabic language achieved 79.5% accuracy while

TABLE 3. Comparison models.

<table>
<tr><th>Model</th><th>Language</th><th>Input</th><th>Dataset</th><th>Classifier</th><th colspan="3">Performance</th></tr>
<tr><td rowspan="2">ACM Model [39]</td><td rowspan="2">English</td><td rowspan="2">Capture visual lib movement using Camera</td><td rowspan="2">English Figure (0-9)</td><td rowspan="2">HMM
3 Hidden States</td><td colspan="3">66.3% Cuave DB</td></tr>
<tr><td colspan="3">64.7% In-house DB</td></tr>
<tr><td rowspan="2">ACM Model [39]</td><td rowspan="2">English</td><td rowspan="2">Capture visual lib movement using Camera</td><td rowspan="2">English Figure (0-9)</td><td rowspan="2">HMM
5 Hidden States</td><td colspan="3">78.33% Cuave DB</td></tr>
<tr><td colspan="3">76.6% In-house DB</td></tr>
<tr><td rowspan="2">AAM Model [37]</td><td rowspan="2">Japanese</td><td rowspan="2">Capture visual lib movement using Camera</td><td rowspan="2">216 Words</td><td rowspan="2">HMM
5 Hidden States</td><td colspan="3">78.67% Single Regression</td></tr>
<tr><td colspan="3">79.56% Multiple Regression</td></tr>
<tr><td rowspan="2">Hyper Column Neural Network &amp; HMM Model [55]</td><td rowspan="2">Arabic
Japanese</td><td rowspan="2">Capture visual lib movement using Camera</td><td rowspan="2">9 Arabic Sentences containing 26 Words</td><td rowspan="2">HCM + HMM
5 Hidden States</td><td colspan="3">79.5% for Arabic Language</td></tr>
<tr><td colspan="3">83.3% for Japanese Language</td></tr>
<tr><td>Arabic Sign Language (ArSL) [56]</td><td>Arabic</td><td>Capture visual lib movement using Camera</td><td>30 Samples</td><td>HMM
5 Hidden States</td><td colspan="3">94.5% Classification Accuracy</td></tr>
<tr><td>[57] Biometric Authentication</td><td>English</td><td>Capture visual lib movement using Camera</td><td>Alphanumeric Samples</td><td>Hamming Distance Classification</td><td colspan="3">95% Accuracy</td></tr>
<tr><td rowspan="3">[58] Lip-password hereinafter Model</td><td rowspan="3">English</td><td rowspan="3">Capture visual lib movement using Camera</td><td rowspan="3">English Figure (0-9)</td><td rowspan="3">M-boosted HMM + RSM</td><td>F</td><td>EER</td><td>Accuracy</td></tr>
<tr><td>1st Feature Set</td><td>9.78%</td><td>90.22%</td></tr>
<tr><td>Last Feature Set</td><td>4.06%</td><td>95.94%</td></tr>
<tr><td>EF-25 [1]</td><td>English</td><td>Capture visual lib movement using Camera</td><td>English Figure (0-9)</td><td>Deep Learning</td><td colspan="3">92.6% Performance</td></tr>
<tr><td>MT-1 [1]</td><td>English</td><td>Capture visual lib movement using Camera</td><td>English Figure (0-9)</td><td>Deep Learning</td><td colspan="3">94.2% Performance</td></tr>
<tr><td>MT-5 [1]</td><td>English</td><td>Capture visual lib movement using Camera</td><td>English Figure (0-9)</td><td>Deep Learning</td><td colspan="3">95.6% Performance</td></tr>
<tr><td>LF-5 [1]</td><td>English</td><td>Capture visual lib movement using Camera</td><td>English Figure (0-9)</td><td>Deep Learning</td><td colspan="3">94.8% Performance</td></tr>
<tr><td>LSTM-5 [1]</td><td>English</td><td>Capture visual lib movement using Camera</td><td>English Figure (0-9)</td><td>Deep Learning</td><td colspan="3">96.4% Performance</td></tr>
<tr><td>Silent Key [48]</td><td>English</td><td>Capture lib behavior using ultrasonic</td><td>Pre-stored Secret Words</td><td>SVM</td><td colspan="3">70% to 83.1%</td></tr>
<tr><td>Lip Acoustic Sensing [49]</td><td>English</td><td>Capture lib behavior using Acoustic Sensing</td><td>Phonemes from 1 to 10</td><td>Deep Learning</td><td colspan="3">83.7% for Registered Users
90.2% User Identification</td></tr>
<tr><td>Visual Password [50]</td><td>English</td><td>Capture visual lib movement using Camera</td><td>12 distinct Words</td><td>HMM</td><td colspan="3">73.1% Accuracy
Equal Error Rate 26.9%</td></tr>
<tr><td><b>Proposed Model</b></td><td><b>Arabic</b></td><td><b>Capture visual lib movement using Camera</b></td><td><b>Arabic Figures (0-9)</b></td><td><b>HMM
10 Hidden States</b></td><td colspan="3"><b>96.2% Performance</b></td></tr>
</table>

the Japanese language achieved 83.3% accuracy using Hyper Column Neural Network (HCM) model and hidden Markov model (HMM). As presented in [57], an Arabic sign language was tested with 30 samples using HMM with 5 hidden states and achieved 94.5% accuracy, while the authors of [58] performed biometric authentication using hamming distance classification on English alphanumeric samples and achieved 95% accuracy. The experimental results were conducted on figures from 0 to 9 using multiple boosted Hidden Markov Model and random subspace method (RSM). The contour-based features achieved an equal error rate (EER) ranging from 9.78% for the first feature set and 4.06% for the last feature set while the accuracy for the first and last feature set achieved 90.22% and 95.94% respectively.

XI. CONCLUSION

In this paper, a hybrid voting framework has been introduced for silent passwords recognition using lip movement analysis. The proposed framework is based on three techniques for automatic visual features extraction, namely SURF, HoG and Haar. The resultant features in each technique are fed separately in a digit prediction model, which is the hidden Markov model (HMM), to learn different utterance of Arabic figures. The final classification models that are produced from the three techniques have been grouped in a voting scheme to produce the final classification result. The proposed framework has been developed and experimented on handcrafted data set of 2000 test records for the ten Arabic digits. The dataset was generated with random noise and different background for each speaker to simulate real life situation of word recognition. The voting scheme has remarkably improved the classification result with 2%. Furthermore, we compared our model with recent similar researches. It has proven to achieve high detection rate and accuracy, while keeping low false positive rate. Indeed, we reach an accuracy percentage of 96.2%, which is higher than the accuracy percentages provided by the most recent five lip reading architectures used in the comparison, namely early fusion (EF-25), multiple towers (MT-1), multiple towers (MT-5), and late fusion (LF-5), while it lower than long short-term memory (LSMT-5) by 0.2% which is one of heavily deep learning architecture and is applied on English language not for Arabic. As short-term future work, we plan to expand the proposed framework by employing other classification techniques, and experiment the model on whole face, instead of using only the mouth region. Moreover, scalability can be targeted by addressing huge dataset from Arabic lexicon in different domains.

Moreover, scalability can be targeted by addressing huge dataset from Arabic lexicon in different domains. Over a medium-term research perspective, we propose to adapt the developed framework for deaf and dump users.

REFERENCES

[1] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Comput. Vis. Image Understand.*, vol. 173, pp. 76–85, Aug. 2018, doi: 10.1016/j.cviu.2018.02.001.

[2] V. Estellers and J.-P. Thiran, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–23, Dec. 2012.

[3] S. Lin, B. Liu, and J. Lin, "Combining speeded-up robust features with principal component analysis in face recognition system," *Int. J. Innov. Comput., Inf. Control*, vol. 8, no. 12, pp. 8545–8556, 2012.

[4] C. Georgakis, S. Petridis, and M. Pantic, "Visual-only discrimination between native and non-native speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4861–4865, doi: 10.1109/ICASSP.2014.6854519.

[5] W.-W. Hu, R.-G. Zhou, A. El-Rafei, and S.-X. Jiang, "Quantum image watermarking algorithm based on Haar wavelet transform," *IEEE Access*, vol. 7, pp. 121303–121320, 2019, doi: 10.1109/ACCESS.2019.2937390.

[6] M. Hooshmand, D. Zordan, T. Melodia, and M. Rossi, "SURF: Subject-adaptive unsupervised ECG signal compression for wearable fitness monitors," *IEEE Access*, vol. 5, pp. 19517–19535, 2017, doi: 10.1109/ACCESS.2017.2749758.

[7] M. Gurban and J. P. Thiran, "Audio-visual speech recognition with a hybrid SVM-HMM system," in *Proc. IEEE 13th Eur. Signal Process. Conf.*, Sep. 2005, pp. 728–731.

[8] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature matching for partial face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 791–802, Feb. 2019, doi: 10.1109/TIP.2018.2870946.

[9] C. Bai, M. Li, T. Zhao, and W. Wang, "Learning binary descriptors for fingerprint indexing," *IEEE Access*, vol. 6, pp. 1583–1594, 2018, doi: 10.1109/ACCESS.2017.2779562.

[10] F. Saeed, M. Hussain, and H. A. Aboalsamh, "Classification of live scanned fingerprints using histogram of gradient descriptor," in *Proc. 21st Saudi Comput. Soc. Nat. Comput. Conf. (NCC)*, Apr. 2018, pp. 1–5, doi: 10.1109/NCG.2018.8592949.

[11] S. C. Eastwood, V. P. Shmerko, S. N. Yanushkevich, M. Drahansky, and D. O. Gorodnichy, "Biometric-enabled authentication machines: A survey of open-set real-world applications," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 2, pp. 231–242, Apr. 2016, doi: 10.1109/THMS.2015.2412944.

[12] N. M. R. Lwamo, L. Zhu, C. Xu, K. Sharif, X. Liu, and C. Zhang, "SUAA: A secure user authentication scheme with anonymity for the single & multi-server environments," *Inf. Sci.*, vol. 477, pp. 369–385, Mar. 2019, doi: 10.1016/j.ins.2018.10.037.

[13] H. Yao, C. Wang, X. Fu, C. Liu, B. Wu, and F. Li, "A privacy-preserving RLWE-based remote biometric authentication scheme for single and multi-server environments," *IEEE Access*, vol. 7, pp. 109597–109611, 2019, doi: 10.1109/ACCESS.2019.2933576.

[14] N. Merhav, "Ensemble performance of biometric authentication systems based on secret key generation," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2477–2491, Apr. 2019, doi: 10.1109/TIT.2018.2873132.

[15] Z. Sitova, J. Sedenka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K. S. Balagani, "HMOG: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 877–892, May 2016, doi: 10.1109/TIFS.2015.2506542.

[16] K. Zhou and J. Ren, "PassBio: Privacy-preserving user-centric biometric authentication," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 12, pp. 3050–3063, Dec. 2018, doi: 10.1109/TIFS.2018.2838540.

[17] T. A. T. Nguyen and T. K. Dang, "Privacy preserving biometric-based remote authentication with secure processing unit on untrusted server," *IET Biometrics*, vol. 8, no. 1, pp. 79–91, Jan. 2019, doi: 10.1049/iet-bmt.2018.5101.

[18] S. Thavalengal, P. Bigioi, and P. Corcoran, "Iris authentication in handheld devices–considerations for constraint-free acquisition," *IEEE Trans. Consum. Electron.*, vol. 61, no. 2, pp. 245–253, May 2015, doi: 10.1109/TCE.2015.7150600.

[19] M. Hammad, Y. Liu, and K. Wang, "Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint," *IEEE Access*, vol. 7, pp. 26527–26542, 2019, doi: 10.1109/ACCESS.2018.2886573.

[20] S. Vhaduri and C. Poellabauer, "Multi-modal biometric-based implicit authentication of wearable device users," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3116–3125, Dec. 2019, doi: 10.1109/TIFS.2019.2911170.

[21] D. Wang and P. Wang, "On the anonymity of two-factor authentication schemes for wireless sensor networks: Attacks, principle and solutions," *Comput. Netw.*, vol. 73, pp. 41–57, Jul. 2014, doi: 10.1016/j.comnet.2014.07.010.

[22] M. L. Das, "Two-factor user authentication in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1086–1090, Mar. 2009, doi: 10.1109/TWC.2008.080128.

[23] Q. Feng, D. He, S. Zeadally, and H. Wang, "Anonymous biometrics-based authentication scheme with key distribution for mobile multi-server environment," *Future Gener. Comput. Syst.*, vol. 84, pp. 239–251, Jul. 2018, doi: 10.1016/j.future.2017.07.040.

[24] G. Xu, S. Qiu, H. Ahmad, G. Xu, Y. Guo, M. Zhang, and H. Xu, ‘‘A multi-server two-factor authentication scheme with un-traceability using elliptic curve cryptography,’’ *Sensors*, vol. 18, no. 7, p. 2394, Jul. 2018, doi: 10.3390/s18072394.

[25] B. Ying and A. Nayak, ‘‘Lightweight remote user authentication protocol for multi-server 5G networks using self-certified public key cryptography,’’ *J. Netw. Comput. Appl.*, vol. 131, pp. 66–74, Apr. 2019, doi: 10.1016/j.jnca.2019.01.017.

[26] E. Alkim, L. Ducas, and T. Pöppelmann, ‘‘Post-quantum key exchange—A new hope,’’ in *Proc. 25th Int. Symp. USENIX Secur.*, 2016, pp. 327–343.

[27] J. W. Bos, C. Costello, M. Naehrig, and D. Stebila, ‘‘Post-quantum key exchange for the TLS protocol from the ring learning with errors problem,’’ in *Proc. IEEE Symp. Secur. Privacy*, May 2015, pp. 553–570, doi: 10.1109/SP.2015.40.

[28] J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehle, ‘‘CRYSTALS–kyber: A CCA-secure Module-Lattice-Based KEM,’’ in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Apr. 2018, pp. 353–367, doi: 10.1109/EuroSP.2018.00032.

[29] J. Galbally, R. Haraksim, and L. Beslay, ‘‘A study of age and ageing in fingerprint biometrics,’’ *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1351–1365, May 2019, doi: 10.1109/TIFS.2018.2878160.

[30] T. Chugh, K. Cao, and A. K. Jain, ‘‘Fingerprint spoof buster: Use of minutiae-centered patches,’’ *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2190–2202, Sep. 2018, doi: 10.1109/TIFS.2018.2812193.

[31] S. R. Arashloo, J. Kittler, and W. Christmas, ‘‘An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol,’’ *IEEE Access*, vol. 5, pp. 13868–13882, 2017, doi: 10.1109/ACCESS.2017.2729161.

[32] K. L. Sum, W. H. Lau, S. H. Leung, A. W. C. Liew, and K. W. Tse, ‘‘A new optimization procedure for extracting the point-based lip contour using active shape model,’’ in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 1485–1488.

[33] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, ‘‘Extraction of visual features for lipreading,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Aug. 2002, doi: 10.1109/34.982900.

[34] D. Kim and R. Dahyot, ‘‘Face components detection using SURF descriptors and SVMs,’’ in *Proc. Int. Mach. Vis. Image Process. Conf.*, Sep. 2008, pp. 51–56, doi: 10.1109/IMVIP.2008.15.

[35] F. Faubel, M. Georges, K. Kumatani, A. Bruhn, and D. Klakow, ‘‘Improving hands-free speech recognition in a car through audio-visual voice activity detection,’’ in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, May 2011, pp. 70–75, doi: 10.1109/HSCMA.2011.5942412.

[36] S. Siatras, N. Nikolaidis, and I. Pitas, ‘‘Visual speech detection using mouth region intensities,’’ in *Proc. IEEE Conf. Signal Process.*, Sep. 2006, pp. 1–5.

[37] Y. Komai, N. Yang, T. Takiguchi, and Y. Ariki, ‘‘Robust AAM-based audio-visual speech recognition against face direction changes,’’ in *Proc. 20th ACM Int. Conf. Multimedia (ICM)*, 2012, pp. 1161–1164, doi: 10.1145/2393347.2396408.

[38] N. Dalal and B. Triggs, ‘‘Histograms of oriented gradients for human detection,’’ in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[39] S. S. Morade and S. Patnaik, ‘‘A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition,’’ *Optik*, vol. 125, no. 18, pp. 5181–5186, Sep. 2014, doi: 10.1016/j.ijleo.2014.05.011.

[40] E. Alsharhan and A. Ramsay, ‘‘Improved arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions,’’ *Inf. Process. Manage.*, vol. 56, no. 2, pp. 343–353, Mar. 2019.

[41] N. Alsunaidi, L. Alzeer, M. Alkatheiri, A. Habbabah, M. Alattas, M. Aljabri, and M. Altassan, ‘‘Abjad: Towards interactive learning approach to arabic reading based on speech recognition,’’ *Procedia Comput. Sci.*, vol. 142, pp. 198–205, 2018, doi: 10.1016/j.procs.2018.10.476.

[42] H. Leopold, H. van der Aa, J. Offenberg, and H. A. Reijers, ‘‘Using hidden Markov models for the accurate linguistic analysis of process model activity labels,’’ *Inf. Syst.*, vol. 83, pp. 30–39, Jul. 2019, doi: 10.1016/j.is.2019.02.005.

[43] M. Colasito, J. Straub, and P. Kotala, ‘‘Correlated lip motion and voice audio data,’’ *Data Brief*, vol. 21, pp. 856–860, Dec. 2018, doi: 10.1016/j.dib.2018.10.043.

[44] S. Das, G. Laput, C. Harrison, and J. I. Hong, ‘‘Thumprint: Socially-inclusive local group authentication through shared secret knocks,’’ in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 3764–3774.

[45] S. Sen and K. Muralidharan, ‘‘Putting–pressure’ on mobile authentication,’’ in *Proc. 7th Int. Conf. Mobile Comput. Ubiquitous Netw. (ICMU)*, Jan. 2014, pp. 56–61, doi: 10.1109/ICMU.2014.6799058.

[46] M. Shahzad, A. X. Liu, and A. Samuel, ‘‘Secure unlocking of mobile touch screen devices by simple gestures: You can see it but you can not do it,’’ in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 39–50, doi: 10.1145/2500423.2500434.

[47] W. Wang, A. X. Liu, and M. Shahzad, ‘‘Gait recognition using WiFi signals,’’ in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, 2016, pp. 363–373, doi: 10.1145/2971648.2971670.

[48] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, ‘‘SilentKey: A new authentication framework through ultrasonic-based lip reading,’’ in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–18, Mar. 2018, doi: 10.1145/3191768.

[49] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li, ‘‘Lip reading-based user authentication through acoustic sensing on smartphones,’’ *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 447–460, Feb. 2019, doi: 10.1109/TNET.2019.2891733.

[50] P. C. Rabaneda, *Lip-Reading Visual Passwords for User Authentication*. Barcelona, Spain: Pompeu Fabra Univ., 2018.

[51] S. Singh and S. V. A. V. Prasad, ‘‘Techniques and challenges of face recognition: A critical review,’’ *Procedia Comput. Sci.*, vol. 143, pp. 536–543, 2018, doi: 10.1016/j.procs.2018.10.427.

[52] W.-Y. Lu and M. Yang, ‘‘Face detection based on Viola-Jones algorithm applying composite features,’’ in *Proc. Int. Conf. Robots Intell. Syst. (ICRIS)*, Jun. 2019, pp. 82–85, doi: 10.1109/ICRIS.2019.00029.

[53] M. Panzner and P. Cimiano, ‘‘Comparing hidden Markov models and long short term memory neural networks for learning action representations,’’ in *Proc. Int. Workshop Mach. Learn., Optim., Big Data*, vol. 10122. Springer, 2016, pp. 94–105.

[54] E. Alp and H. Keles, ‘‘A comparative study of HMMs and LSTMs on action classification with limited training data,’’ in *Proc. Int. Conf. SAI Intell. Syst.*, vol. 868. Springer, 2018, pp. 1102–1115.

[55] A. El Sagheer and N. Tsuruta, ‘‘Arabic lip-reading system: A combination of hypercolumn neural networks model with hidden Markov model based,’’ in *Proc. Int. Conf. Artif. Intell. Soft Comput.*, 2004, pp. 311–316.

[56] M. Mohandes, M. Deriche, U. Johar, and S. Ilyas, ‘‘A signer-independent arabic sign language recognition system using face detection, geometric features, and a hidden Markov model,’’ *Comput. Electr. Eng.*, vol. 38, no. 2, pp. 422–433, Mar. 2012.

[57] R. K. Shubhangi and N. B. Balbhim, ‘‘Lip’s movements biometric authentication in electronic devices,’’ in *Proc. Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Jul. 2017, pp. 998–1001, doi: 10.1109/ICCMC.2017.8282619.

[58] X. Liu and Y.-M. Cheung, ‘‘Learning multi-boosted HMMs for lip-password based speaker verification,’’ *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 2, pp. 233–246, Feb. 2014, doi: 10.1109/TIFS.2013.2293025.

MOHAMED EZZ (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in systems and computers engineering from the Faculty of Engineering, Al Azhar University. He is currently an Associate Professor with the Faculty of Engineering, Al Azhar University, and a Visiting Professor with the College of Computer and Information Sciences, Jouf University. He has published 20 scientific articles in various national and international journals and conferences. He has contributed in more than 16 mega software projects in electronic banking EBPP, EMV, mobile banking, and e-commerce, also CBAP Certified. His research interests include pattern recognition, applied machine learning, application security, intrusion detection, and semantic web.

AYMAN MOHAMED MOSTAFA (Member, IEEE) received the M.Sc. and Ph.D. degrees in information systems from the Faculty of Computers and Informatics, Zagazig University, Egypt. He is currently an Assistant Professor with the Faculty of Computers and Informatics, Zagazig University, and the College of Computer and Information Sciences, Jouf University, Saudi Arabia. He has published more than 20 scientific articles in various national and international journals and conferences. His research interests include information security, cloud computing, e-business, e-commerce, big data, and data science. He is also an Oracle Certified Associate, an Oracle Certified Professional, and an EMC Academic Associate in cloud infrastructure and services.

ABDURRAHMAN A. NASR received the M.Sc. and Ph.D. degrees in electrical engineering from the Systems and Computers Department, Al-Azhar University, Cairo, in 2012 and 2014, respectively. He is currently a Lecturer of software engineering with the Systems and Computers Engineering Department, Faculty of Engineering, Al-Azhar University. His research interests include artificial intelligence, stochastic process, machine learning, data mining, mathematics, and operating systems.