JAUES

Journal Of Al Azhar University Engineering Sector

Vol. 12, No. 42, January, 2017, 155-165

AN ARABIC FIGURES RECOGNITION MODEL BASED ON AUTOMATIC LEARNING OF LIP MOVEMENT

Alzahraa H. Reda, Abdurrahman A. Nasr, Mohamed M. Ezz, Hany M. Harb
System and Computer Engineering Dept., Faculty of Engineering, Al-Azhar Univ.
Cairo, Egypt

ABSTRACT
The need for an automatic speech to text conversion is continuously increasing, especially for people with special needs. Thus, automatic speech recognition techniques have been proposed to tackle such needs. The automatic recognition allows computers to identify words from lip movement, regardless of the visual source. It is known that visual speech recognition techniques improve the accuracy of word identification and shield the recognition system against acoustic noise. In this paper, we propose a hybrid voting model for automatic Arabic figures recognition based on visual perception of mouth-lip movement. The proposed model has been built for Arabic language, such that, it is able to extract Arabic figures from predefined Arabic lexicon. The predefined lexicon mainly contains the Arabic figures from zero to nine with different shapes. The model takes a video (or sequence of images) as an input, and outputs the corresponding Arabic figure for the frames extracted from the input video.
Here, three techniques have been employed to extract effective visual features from mouth-lip movement. Such techniques are SURF (speeded up robust features), HoG (histogram of oriented gradient) 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 (i.e. classification by voting). The proposed model in this paper has been tested on handcrafted data set of lip movement, and it has shown a promising result with improved accuracy of Arabic figures recognition.

KEYWORDS: Automatic speech recognition, hidden Markov model, histogram of oriented gradients, speeded up robust feature, Haar feature extractor, voting.

1. INTRODUCTION
Lip movement information has been widely utilized in the state of the art techniques for audio-visual automatic speech recognition (AV-ASR)[1]. The study of visual features has emerged as attractive solution to speech recognition for people with special needs. 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), Wahed (one), Ethnan (two), Thlatha (three), Arbaa (four), Khamsa (five), Setta (six), Sabaa (seven), Thamania (eight) and Tesaa (nine).

In this paper, 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. These techniques have been chosen based on their remarkable efficacy and reported accuracy in different domains, for example in [6], [7] and [8]. The first technique (i.e. the SURF) detects the regions of interest (e.g. lips, faces, etc...) in an image and produces a description of these points in the form of feature vectors. The technique works by transforming the source image into coordinates using the integral image algorithm[9], 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 20X20 is taken around the key point and is divided into 4x4 sub-regions. For each sub-region, horizontal and vertical wavelet responses are taken and a vector is formed in the form $v = [\sum dx, \sum dy, \sum |dx|, \sum |dx|]$. The final SURF feature descriptor for the 4x4 sub-regions is 64 dimensions length for describing point of interest. Fig. 1 illustrates the SURF detection and description process.

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 5x3 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 the gradient magnitude. The combination of these normalized histograms represents the feature descriptor, which is a vector of the components of the cell histograms from all of the block regions. Fig. 2 illustrates the HoG detection and description process.

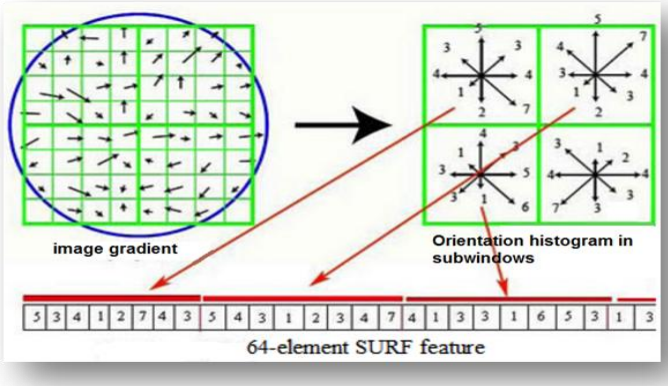


Figure 1. SURF feature vector construction from image gradients[10]

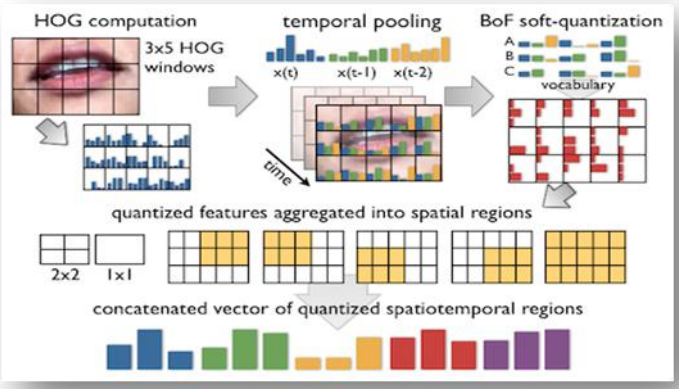


Figure 2.HoG feature vector construction from image gradients[11].

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[9]. Fig. 3 Represents Haar-like features used in training the classifier. 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.

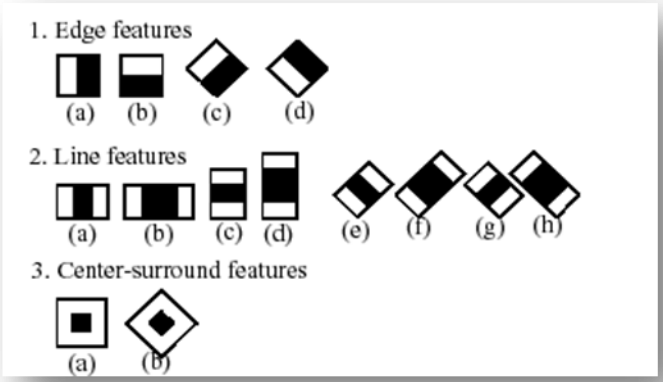


Figure 3.Haar-like features[12].

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, so HMM matches our area of research very well. Moreover, HMM training algorithms are very popular, simple, and computationally feasible to use. In this paper, A sequence of words or phonemes is made by concatenating 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 emission probability of each state is modeled by a mixture of Gaussians, trained with the expectation-maximization algorithm (EM)[13].

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 rest of this paper is organized as follows; Section 2 introduces a literature review for related work, section 3 introduces the proposed model, section 4 illustrate a case study on the proposed model, Section 5 shows the experimental result, and section 6 summarizes the proposed model.

Related work

Many techniques have been proposed by researchers to tackle the automatic speech recognition using source visual information (i.e. lip-reading).Sum et al. [14] 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 el al. [15] 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.[16] 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 et al.[17] 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. [18] 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. [19] 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.[20] 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. Dalal et al. [21] has 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). Usman et al.[5] Has improved the performance of the speech recognizers by 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. Sanual et al. [22] 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.
The motivation behind this work is three folds; the first of which is to enhance the results obtained from the previous work 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.

Proposed Model

Fig. 4 Illustrates the proposed methodology for automatic figure recognition based on visual perception. The model consists basically of 5 layers, namely, input, preprocessing, visual feature extraction, learning and finally the voting scheme.

The input layer represents a dataset of short length videos that is focused on the facial region. Those videos are fed to preprocessing layer, in which faces detection and lip localization algorithms are applied. In this work, we have utilized Viola-Jones algorithm 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.

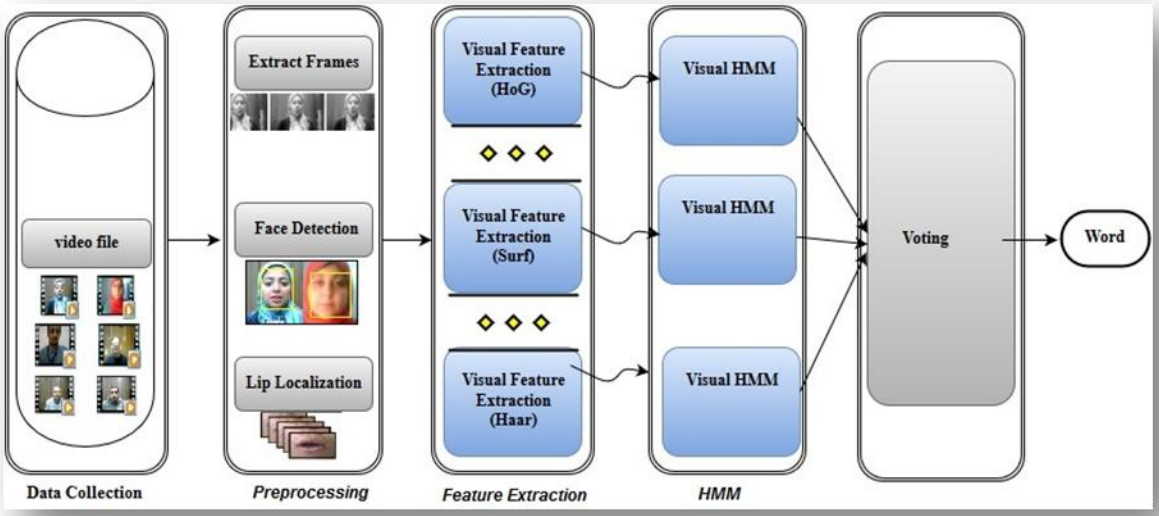


Figure 4. Proposed system model.

The feature extraction layer takes the output of preprocessing layer and applies the corresponding technique to extract the features that represent the mouth. This will yield three types of feature representation based on the feature detection and extraction algorithms mentioned earlier in this work. Each type of resultant features is fed separately to the learning layer, namely the HMM, to apply the training step on mouth regions. Figs 5a, 5b and 5c illustrate valid points of mouth features for SURF, Haar and HoG features respectively.

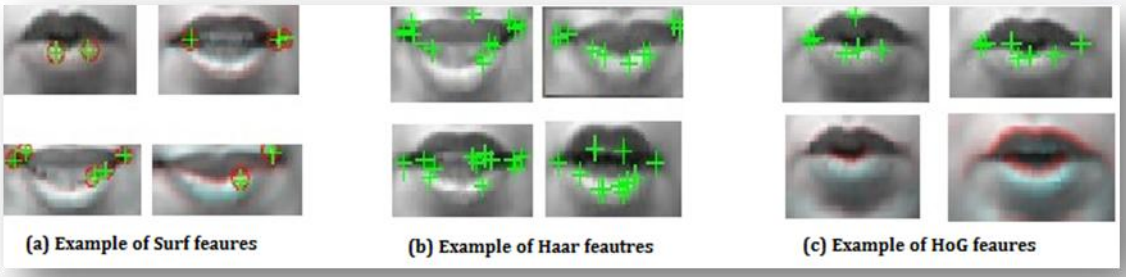


Figure 5. Example of different lip features detection

The learning layer applies forward-backward learning algorithm for learning the HMM parameters. 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 [22], and finally multiplies the whole conditional probabilities altogether. More formally:

$$P(O \mid Q) = \Pi^{T}_{i=1} \; P(o_i \mid q_i)$$

Where $O = o_0, o_1 \ldots o_T$, represents a set of possible observation sequence,
$Q = q_0, q_1 \ldots q_T$ represents a set of hidden state sequence, and P (O | Q) is the likelihood of the observation given some sequence at time T.
The learning layer produces three different learning models, with each one corresponds respectively to Surf, Haar, and HoG. The final classification (word recognition) result is obtained by feeding the generated models to a voting scheme to vote over the odd number of the inputs.

Case Study
In this section, we illustrate the implementation of a specific case study on the model proposed in this paper. 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 (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 records for the ten Arabic digits. 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.

Face and lip Detection
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. Fig. 6 depicts a simple geometric lip model, in which four points that represents the maximum and minimum points of the lips corner at x and y direction were selected. The four points include (1) the vertical distance from the nose to the corner of the mouth; (2) horizontal opening of inner lip contour; (3) vertical distances from horizontal axis, which is the line joining mouth corners, to the inner upper contour; and (4) inner lower contour at four equally spaced points between the center and the right corner of the mouth. These points help in detection and representation of lip shape, such that it can identify the utterance of a word with its different vowels. Fig. 7 illustrates a visual dataset for lips.

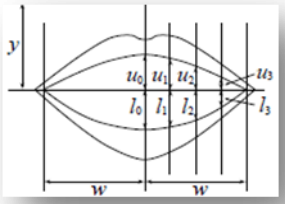


Figure 6. Lip model parameters [23]

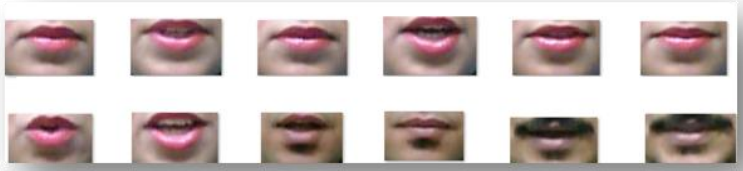


Figure 7. Sample (Real) dataset for different lips

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 utilized 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 dimension 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).

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 $\{y_1, y_2, \ldots, y_n\}$ is 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 $\{o_1, o_2, \ldots, o_m\}$ 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.

Cumulative Voting Formulas

In this paper, 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 as follows:

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

Where

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

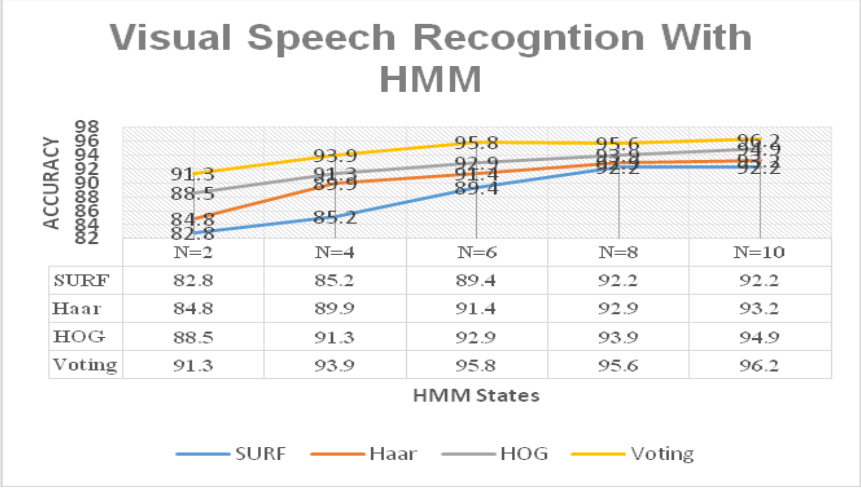
S: Total Number of feature extraction methods to Vote at model

D: Number of times votes want to elect

N: number of Figures needed

Experimental result

In this section, the evaluation of Haar, HoG and Surf, and a comparison of their performance are given. The proposed model is tested using cross validation and the evaluation is based on: (i) Accuracy, which is the correct classified records, over all records, (ii) true positive rate (TPR), which measures the proportion of Arabic figures that are correctly identified as such, and (ii) false alarm rare (a.k.a. false positive rate, FPR), which measure the percent of incorrectly identified Arabic figures. Fig. 8 illustrates the accuracy of different models against variation of hidden Markov states.



	N=2	N=4	N=6	N=8	N=10
SURF	82.8	85.2	89.4	92.2	92.2
Haar	84.8	89.9	91.4	92.9	93.2
HOG	88.5	91.3	92.9	93.9	94.9
Voting	91.3	93.9	95.8	95.6	96.2

Figure 8. Accuracy versus number of hidden states

From the figture, it's clear that voting scheme enhances the result by about 2%. Table 1 shows 10x10 confusion matrix that represents the 10 Arabic figures. Table 2 summarizes the DR, FPR and accuracy of the voting scheme for each Arabic figure.

Table 1. Confusion matrix with voting system

Word	Sefr	Wahed	Ethnan	Thlatha	Arbaa	Khamsa	Setta	Sabaa	Thamania	Tesaa
	100	100	100	300	200	400	200	200	200	200
Sefr(zero)	98	0	0	1	0	0	1	0	0	0
Wahed (one)	0	97	0	0	1	1	0	1	0	0
Ethnan (two)	0	0	96	1	0	0	2	0	0	1
Thlatha (three)	1	0	1	288	2	1	3	1	1	2
Arbaa (four)	0	0	0	2	191	1	2	1	1	1
Khamsa (five)	0	1	0	1	1	394	2	0	1	0
Setta (six)	1	0	2	3	2	2	188	0	1	1
Sabaa (seven)	0	1	0	1	1	0	0	195	1	1
Thamania (eight)	0	0	0	1	1	1	1	1	93	2
Tesaa (nine)	0	0	1	2	1	0	1	1	2	192

Table 2. Proposed model Accuracy, TPR and FPR

Word	Accuracy	TPR	FPR
Sefr	98.0%	95.80%	1%
Wahed	97.9%	95.80%	1%
Ethnan	96.0%	96.90%	3%
Thlatha	96.0%	94.90%	7%
Arbaa	95.5%	96.90%	9%
Khamsa	98.5%	96.80%	2%
Setta	94.0%	96.90%	1%
Sabaa	97.5%	96.20%	2%
Thamania	93.0%	96.60%	2%
Tesaa	96.0%	94.60%	2%

Comparative study

Table 3 compares the proposed model to similar researches in [20], [22], [23], and [24]

Table 3. Comparison with another models

Model	Language	Dataset	classifier	ACCURACY
ACM Model[22]	**English**	**English figure(0-9)**	**HMM,5 hidden states**	**77.8 to 79.6**
AAM Model [20]	**Japanese**	**216 words**	**HMM,5 hidden states**	**77 to 80.7**
Proposed Model	**Arabic**	**Arabic figures(0-9)**	**HMM,10 hidden states**	**96.2**
Fuzzy K-NN Model [24]	**Arabic**	**Arabic figures(1-10)**	**K-NN**	**55.8**
Hyper column Neural Network &HMM Model [23]	**Arabic**	**Arabic sentences [9 sentences]**	**HNM+HMM,5 hidden states**	**62.9**

We propose automatic Arabic figures recognition based on visual lip movement. The final classification models that are produced from the three techniques (Haar, Surf and Hog) have been grouped in a voting scheme to produce the final classification result.

CONCLUSION

In this paper, a hybrid model has been introduced for Arabic visual speech recognition based on three automatic feature extraction and description techniques, viz. Haar, HoG and Surf. Such features are introduced to a machine learning classifier, namely the HMM to learn different utterance of Arabic figures. Three classification models for each feature descriptor have been built up and grouped in a voting scheme. The voting scheme has remarkably improved the classification result with 2%. The model has proven to achieve high detection rate and accuracy, while keeping low false positive rate. The future work may include expanding the proposed model by employing other classification techniques, and experimenting 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.

REFERENCES

[1] G. Potamianos, C. Neti, and Y. Heights, “Audio-Visual Speech Recognition in Challenging Environments,” pp. 1293–1296, 2003.
[2] V. Estellers and J. Thiran, “Multi-Pose Lipreading and Audio-Visual Speech Recognition,” pp. 1–23, 2012.
[3] S. Lin, B. Liu, and J. Lin, “Combining Speeded-Up Robust Features with Principal Component Analysis in Face Recognition System,” *J. Innov. Comput. Inf.*, vol. 8, no. 12, pp. 8545–8556, 2012.
[4] C. Georgakis and S. Petridis, “Visual-Only Discrimination Between Native and Non-Native Speech Department of Computing , Imperial College London , London,” pp. 4861–4865, 2014.
[5] M. Usman, G. Khan, S. Mahmood, M. Ahmed, and Y. Gotoh, “Visual Speech Detection Using Opencv,” pp. 24–29.
[6] X. Lu, X. Lu, and E. Lansing, “Image Analysis for Face Recognition,” *Pers. Notes, May*, vol. 5, pp. 1–37, 2003.
[7] L. El Shafey, E. Khoury, and S. Marcel, “Audio-Visual Gender Recognition in Uncontrolled Environment Using Variability Modeling Techniques,” *IEEE Int. Jt. Conf. Biometrics*, 2014.
[8] C. Gottschlich, E. Marasco, A. Y. Yang, and B. Cukic, “Fingerprint Liveness Detection Based on Histograms of Invariant Gradients,” *Proc. IJCB*, pp. 1–7, 2014.
[9] P. Viola and M. Jones, “Rapid Object Detection Using A boosted Cascade of Simple Features,” *Comput. Vis. Pattern Recognit.*, vol. 1, pp. I–511–I–518, 2001.
[10] D. Schmitt and N. McCoy, “Object Classification and Localization Using SURF Descriptors,” *Object Classif. Localization Using SURF Descriptors*, pp. 1–5, 2011.
[11] “AT&T Labs Research - Enhanced Indexing and Representation with Vision-Based Biometrics.” .
[12] W. Najwa and W. Ismail, “Object Detection System Using Haar-Classifier,” Faculty of Electrical & Electronic Engineering, University Malaysia Pahang, 2009.

[13] M. Gurban and J. P. Thiran, "Audio-Visual Speech Recognition with A hybrid SVM-HMM System," *13th Eur. Signal Process. Conf. EUSIPCO 2005*, pp. 728–731, 2005.

[14] 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," *2001 IEEE Int. Conf. Acoust. Speech, Signal Process. Proc. (Cat. No.01CH37221)*, vol. 3, pp. 1485–1488.

[15] 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, 2002.

[16] X. P. Hong, H. X. Yao, Y. Q. Wan, and R. Chen, "A PCA Based Visual DCT Feature Extraction Method for Lip-Reading," *IIH-MSP 2006 Int. Conf. Intell. Inf. Hiding Multimed. Signal Process. Proc.*, pp. 321–324, 2006.

[17] D. Kim, "Face Components Detection using SURF Descriptors and SVMs," " in Proceedings of the 2008 International Machine Vision and Image Processing Conference, ser. IMVIP '08

[18] 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," *Jt. Work. Hands-free Speech Commun. Microphone Arrays*, no. Mccc, pp. 70–75, 2011.

[19] S. Siatras, N. Nikolaidis, and I. Pitas, "Visual Speech Detection Using Mouth Region Intensities,"In IEEE Transactions on Circuits and Systems for Video Technology, Volume 19 Issue 1, January 2009 , Pages 133-137 .

[20] Y. Komai, N. Yang, T. Takiguchi, and Y. Ariki, "Robust AAM-Based Audio-Visual Speech Recognition Against Face Direction Changes," *Proc. 20th ACM Int. Conf. Multimed. - MM '12*, p. 1161, 2012.

[21] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[22] S.Morade and S Patnailk , "Anovel Lip Reading Algorithm by Using Localized ACM and HMM :Tested for Digit Recogntion ," Optik –International Journal for Light and Electronic Optics ,Sept 2014 ,Vol.125(18).

[23] A. El Sagheer and N. Tsuruta, "Combination of Hypercolumn Neural Networks Model with Hidden Markov Model Based Lip-reading for Arabic Language." Artificial Intelligence and Soft Computing ASC, 9.1 - 9.3, Marbella, Spain, 2004.

[24] I.Eldirawy, W. Ashour, "Visual Speech Recognition," LAP Lambert Academic Publishing, P 116, no. May, 2011.