九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Hyper Column Model vs. Fast DCT for Feature Extraction in Visual Arabic Speech Recognition

Sagheer, Alaa
Department of Intelligent Systems, Kyushu University

Tsuruta, Naoyuki
Department of Electronics Engineering and Computer Science, Fukuoka University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

Maeda, Sakashi
Department of Electronics Engineering and Computer Science, Fukuoka University

https://hdl.handle.net/2324/5860

出版情報：Proceedings of 5th International IEEE Symposium on Signal Processing and Information Technology, pp.761-766, 2005-12
バージョン：
権利関係：

Hyper-Column Model vs. Fast DCT for Feature Extraction in Visual Arabic Speech Recognition

Alaa Sagheer[1], Naoyuki Tsuruta [2], Rin-Ichiro Taniguchi[1], Sakashi Maeda[2]

[1] Department of Intelligent Systems, Kyushu University,
6-1, Kasuga-Koen, Kasuga, Fukuoka 816-8580, Japan

[2] Department of Electronics Engineering and Computer Science, Fukuoka University,
8-9-1, Nanakuma, Jonan-ku, Fukuoka 814-0180, Japan

E-mail: {alaa, rin}@limu.is.kyushu-u.ac.jp, {tsuruta, maeda}@tl.media.fukuoka-u.ac.jp

Abstract- Recently, the multimedia signal processing community has shown increasing interest for research development on visual speech recognition domain. In this paper we present a novel visual speech recognition approach based on our model Hyper-Column Model (HCM). HCM is used for feature extraction task. The extracted features are modeled by Gaussian distributions through using Hidden Markov Model (HMM). The proposed system, HCM and HMM, can be used for any visual recognition task. We use it here to comprise a complete lip-reading system and evaluate its performance using Arabic database set. According to our knowledge, this is the first time that visual speech recognition is applied for Arabic language. Toward fair evaluation we compare our accuracy results with those using Fast Discrete Cosine Transform (FDCT) approach, in a separate experiment and using same data set and conditions of HCM experiment. Comparison turns out that HCM shows higher recognition accuracy than FDCT for Arabic sentences and words. HCM does not provide higher accuracy only but also it capable to achieve shift invariant recognition whereas FDCT can not.

Keywords- *Visual speech recognition, feature extraction, self organizing map, hyper-column model, discrete cosine transform.*

1. INTRODUCTION

An enormous amount of progress in visual speech recognition has been undergone over the past decade regarding to speaker independence, large vocabulary sets and fast and real time recognition. In visual speech recognition domain, the technique of retrieving speech content from visual clues such as the movement of the lips, tongue, and teeth is commonly known as automatic lip reading. Indeed, the need of automatic lip reading becomes high for hearing impaired and deaf people who used to understand speech without acoustic information but only by reading the speaker's lips. Further, automatic lip reading systems tend to be applied to areas such as speaker verification, multimedia telephony for the hearing impaired and home health care systems for elderly [1].

Automatic lip reading may be carried out at two stages. In the first stage the visual speech features such as phonemes are extracted. In the second stage, the extracted features are modeled by Gaussian distributions through using HMM [2]. In General, there are three main categories have been reported in literature capable to extract visual features:

1. Geometric-feature based;
2. Image transform based; and
3. Appearance based.

Geometric-feature based approach obtains information from geometric features such as lip's height or width or color or shape or all of them [3] [10]. In the image based approach the original gray level image containing the lips is used, after processing by any image transform technique, as a vector of features [4]. Whereas the appearance based approach learns the decision boundary between different articulations from training data without any extraction of geometric features.

The approach presented in this paper falls into the third category and it uses our system presented in [19], in which we use our model HCM as a feature extraction tool. Mainly, HCM [8] is a self organizing neural network model and its structure is based on Self Organizing Map (SOM) [9]. In contrast, FDCT [11] falls into the second category. As it is reported by variety of benchmark contributions that the performance of appearance based approach is better than that of geometric based approach in performing the feature extraction task [5-7]. Also in this paper we show, experimentally, that the performance of the appearance based approach is better than that of the image transform based approach. In other words, we show the superiority of HCM than FDCT for feature extraction task.

To the end of objective evaluation we achieved separate experiments using same database set and same experimental conditions. In each experiment we use a different feature extractor. Fair comparison demonstrates that HCM is capable to extract features in a better manner than FDCT. Furthermore, we achieved a third experiment using SOM as a feature extractor. Again HCM shows better performance than its ancestor; SOM. Final advantage is that HCM can support shift invariant recognition whereas both FDCT and SOM can not [3], [12].

Regarding to the feature space for each model, here we show the best HCM feature space is in 2 dimensions, whereas it has 3 dimensions in case of SOM. On the other hand, regarding to FDCT, we performed its experiment at different numbers of dimensions; namely from 2 to 30. We found that the best FDCT feature space should be in 9 dimensions to express about the input image features. Of course, feature space with 9 dimensions is more complex than this with 2 or 3 dimensions. Other advantages of HCM will point out in the subsequent sections.

Last but not least, we introduce all approaches to same database set of Arabic language. We guess that this is the first

time that Arabic language is introduced for visual speech recognition applications. The database is performed using multiple full sentences gathered from 9 different native subjects with a total number of images is 6480 gray scale image. All images were captured in a natural environment without using a special lighting conditions or lip markers or coloring. The following section addresses previous work and our model advantages comparing with DCT based systems and other previous systems as well.

2. PREVIOUS WORK

The early evidence that vision can improve speech recognition was presented by *Petajan* in 1984 [13]. In Petajan's system, binary mouth images are analyzed to calculate the distance of geometric measures among different mouth shapes in order to identify the visual representations of word units. *Yuhas* [14] developed a lip-reading system where the whole gray scale image containing mouth area was used as the feature vector. He used neural networks to fuse information from the visual channel. *Stork* [15] used the time-delayed neural networks (TDNNs) for recognition.

However, with the beginning of the 90's decade, the development of hidden Markov models (HMM) [2] improved the speech recognition accuracy and made possible large-vocabulary recognition. HMM was first applied to visual speech recognition by *Goldschen* [16] who modified the earlier Petajan's system by using discrete HMMs. *Dupont* [5] used HMM based active shape models to extract active features set that includes derivative information, and compared it's performance with the performance of a static feature set. *Matthews* [7] compared between three methods to parameterize lip image sequences for recognition using HMM.

Also, it is reported that discrete orthogonal transforms can achieve the visual speech feature extraction task by a good manner [17]. *Potamianos* [4], [18] used a group of image transforms algorithms to combine the visual features of the mouth images either geometrically (lip's height and width) or nongeometrically. Potamianos concluded that DCT is the best among other algorithms as a feature extractor tool. Encouraged by Potamianos results *Heckmann* [12] investigated different strategies to choose DCT's coefficients to enhance feature extraction.

Regardless which technique is used, in our thinking, the visual feature extraction module is required to have the following two criteria:

1. It should make a parametric feature space with low dimensionality.
2. Distributions of each phoneme should be simple and approximated by the normal distributions.

Through the subsequent sections we will show that HCM satisfy these criteria.

2.1 Advantages of the Proposed Approach

In image processing domain, visual speech feature recognition is a very difficult task due to the large appearance variability during lip movements and appearance differences across subjects. Therefore, a considerable number of the reported systems include some defects. Here, we summarize the main advantages of our system over other systems. First advantage of our lip-reading (or visual speech) system is that it is capable to extract all the relevant speech features without feature reduction. It is reported in [12] and [18] that to achieve feature extraction using DCT you need to reduce its coefficients which may reduce the number of extracted features as well; see Table 2 later in this paper. In contrast, HCM based system depends on the intensity of image pixels that include information without reduction. Also, HCM can perform feature extraction without needing to lips model as the systems in [5] and [7] or lips marker or lips coloring as the systems in [3] and [12]. Most important advantage is that our system can work under shifted or rotated lip positions [8], [19], which do not available in some other approaches such as DCT-based systems in [12] and [17].

3. SELF ORGANIZING MAP (SOM)

SOM is one of the most widely used artificial neural network algorithm which uses unsupervised competitive learning [9]. Basically, it is a biological model of the Hypercolumn in human brain. SOM is usually represented as a two dimensional neural network sheet (or map) whose units, neurons, become tuned to different input vectors I. Each neuron u has a weight vector W_u. In each training step, one input vector I from the input data set is shown and a similarity measure, usually taken as Euclidian distance, is calculated between it and all neurons of the sheet. The best matching neuron c, denoted as *Winner*, is the neuron whose weight vector W_c has the greatest similarity with the input sample I. In other word, the *Winner c* is the neuron for which

$$\|I - W_c\| = \min_u(\|I - W_u\|) \quad (1)$$

After finding the *Winner* neuron, the weight vectors of SOM sheet are updated according to the rule

$$W_u(t+1) = W_u(t) + h_{cu}(t)[I(t) - W_u(t)] \quad (2)$$

$$h_{cu}(t) = \alpha(t).\exp\left(\frac{\|r_c - r_u\|}{2\sigma^2(t)}\right) \quad (3)$$

where $h_{cu}(t)$ is the neighborhood kernel around the *Winner c* at time t, $\alpha(t)$ is the learning rate and is decreased gradually toward zero and $\sigma^2(t)$ is a factor used to control the neighborhood kernel. Finally after finishing the training stage, the neurons sheet is automatically organized into a meaningful two-dimensional order denoted by feature map (or codebooks).

The SOM codebooks have two remarkable characteristics:

1. The Probability Distribution Function PDF of the codebooks is a good approximation for the PDF of the training data.
2. The topographical order of the training data is preserved in the codebooks, even if the dimensionality of the SOM is smaller than that of the training data.

The last characteristic means that each two similar samples near to each other in the input field are correspondingly located near each other in the feature map. This ordering takes place automatically without external supervision based on only the internal relations in the structure of the input patterns and on the coordination of the neurons activities through the lateral connections between the neurons. In our thinking, this topological nature of SOM as a mapping is the key point of the applicability of SOM to speech recognition domain. In other

words, similar speech features are mapped to nearby positions in the feature map.

4. HYPER-COLUMN MODEL (HCM)

Hyper-Column Model is also an unsupervised neural network model. Mainly it designed as a pyramidal piling up hierarchical layers derived from the Neocognitron neural network (NC) model [20] by replacing each NC unit cell with a Hierarchical Self Organizing Map (HSOM) [21] cell. For the present experiments, we chose HCM with three layers, two as intra-layers and the third one as output layer, as it shown in Figure 1.

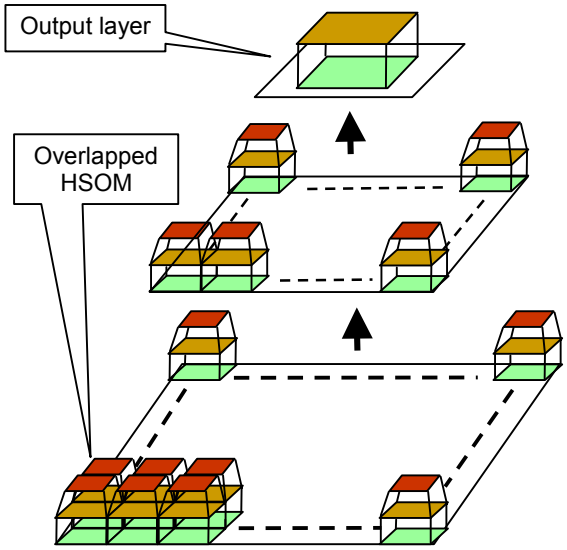


Fig. 1. Hyper-Column Model (HCM)

Each HCM intra-layer has number of overlapped HSOM cells to cover all lips positions in case of lip's shift or rotate. Each HSOM cell consists of input layer and two hierarchical SOM layers as it is shown in Figure 2. The lower SOM layer is called feature extraction layer, and performs features extraction by quantizing the input space and mapping it into a neuron array of low dimensional.

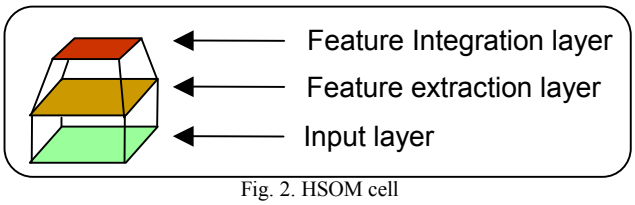


Fig. 2. HSOM cell

The upper SOM layer is called feature integration layer and inputs the winner neuron index from the first SOM layer to perform optimum feature integration and then choose the winner neuron. Moreover, this layer enables shifted, rotated, scaled and distorted invariant recognition [8] by a similar way as it is in the Neocognitron. Each HSOM cell works by selecting one winner for each input position, the winner of all winners is considered as the cell's output. Then, the array of all winner position indexes from all HSOM cells in each intra-layer forms the final output of that HCM intra-layer and, in the same time, the input of the following HCM layer.

The HCM output layer is selected as only one SOM with 64 neurons distributed in two dimensions. Due to the competitive learning nature of HCM, only one neuron from this layer will activate in correspondence to the category of the input pattern, other neurons respond to other categories. Finally, the learning process line of HCM is carried out beginning from bottom to top layer by layer where the standard learning algorithm of SOM, equations 1-3, is used to train each SOM's neurons.

4.1 HCM's Advantages

Indeed, the fact of combining structure of both NC and SOM approaches lets HCM inherits the advantages of both of them too. First of all, we demonstrated, experimentally, that HCM is capable to reduce high dimensional input space and generating ordered mappings of the input data onto some low-dimensional feature map (space) [8]; which satisfies the first criterion of best feature extractor tool mentioned in section 2. Another challenge is inherited from SOM that the HCM's ability to partitioning the input data into clusters in such a way that gathering the similar input data items into one cluster and keeping them close to each other in the output space [19]. Of course this capability is very useful in analyzing high-dimensional data. Other HCM advantages can be summarized as follows [8]:

1. HCM can recognize distort and invariant objects with variations in position and size.
2. Accept random initialization for network weights.
3. No preprocessing for input images is needed.

5. HMM-BASED VISUAL SPEECH RECOGNITION

Needless to say that HMM holds the greatest promise among the various techniques used for visual speech recognition studied so far due to its capabilities in handling either the variability or the sequence of speech features. Visual speech features, extracted by HCM, is recognized using HMM where each HMM is trained using the HTK toolkit [22]. One HMM is constructed for each phoneme and continuous speech is recognized by joining the phonemes together to make any required word or sentence using pronunciation dictionary. Each HMM has five states from left to right and allows self-loops and sequential transitions between current state and next state; see Figure 3.

Recognition process using HMM is divided into two phases: training and testing. In training phase, a training set of features and their associated transcriptions for each sentence are used to estimate the HMM parameters of that sentence. In testing phase, unknown features are transcribed and then the probability of each model generating that sentence is calculated. Finally, most likely model identifies the target sentence.

5.1 Visual Speech Features Modeling

For modeling visual speech features, consider a visual observation O of uttered sentence is represented by the following sequence of features vectors: $O = o_1, o_2, o_3, \ldots, o_T$, where o_t is the feature vector extracted at time t. Each HMM is initialized using a uniform segmentation, followed by iterative segmentation using Viterbi alignment approach. Each model parameters are further re-estimated using Baum–Welch procedure.

It is demonstrated in HMM based approach that each HMM representing a particular utterance is defined by the parameter set: ω_i = (A, B, π) where A={a_{ij}} is a matrix of state transition probabilities from state *i* to state *j*, B is vector of observation (or output) probabilities $b_j(o)$ for state *j*, and π is the vector with probabilities π_i of entering the model at state *i* [2].

HMM-based recognition is performed using Viterbi algorithm which calculates the most likely state sequence for each HMM of having generated the observed sequence. In other words, sentence recognition is performed by estimating the following maximum a posteriori probability:

$$\arg\max_i \{P(\omega_i|O)\} \quad (4)$$

The probability included in 4 can be obtained using Bayes rule:

$$P(\omega_i|O) = \frac{P(O|\omega_i)P(\omega_i)}{P(O)} \quad (5)$$

where P(ω_i) represents the prior probability of a category i and it is assumed to be equal for all categories, and $P(O)$ is assumed to be constant for all categories. So, for simplicity, both $P(\omega_i)$ and P(O) are ignored. For a state sequence has the form of x(1), x(2),…,x(T) of any model, the most probable spoken sentence depends only on the likelihood $P(O \mid \omega_i)$ and can be estimated as the product of state transition probabilities ax and output probabilities $b_x(o)$ of the most likely state sequence. In other words,

$$P(O|\omega_i) = \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} = \prod_{t=1}^{T} b_{x(t)}(o_t) \prod_{t=1}^{T} a_{x(t)x(t+1)} \quad (6)$$

5.2 Requirements for Phoneme Features Space

According to the above representation of each sentence, it is clear that our system is concerned with the output probability

$$\prod_{t=1}^{T} b_{x(t)}(o_t)$$

In the current project we approximate such probability using Gaussian distribution. This means that the feature space of HCM satisfies the condition of simple distribution. From this point of view, second criteria of best features extractor tool, which mentioned in section 2, is satisfied.

5.3 HMM-Based Phoneme Recognition: Example

Here, we show an example of speech recognition process using our system. For instance, consider the word RASI of the last sentence in Table 1; ALM FI RASI (A pain in my head). Specifically, consider the phoneme S of this word. In the structure of HMM in Figure 3, first and fifth node represents start and end node, respectively. The other three nodes, from node 2 to node 4, are representing phoneme S. Node 3 represents the stable state of the phoneme, while node 2 and node 4 are representing changing status from the previous phoneme A to the current and from the current to the following phoneme I, respectively. Since images have large dimensionality, then feature extraction process using HCM model is applied to draw a low-dimensional continuous space. In the recognition process, each node from 2 to 4 outputs one frame image at the transition space according to the output probability $b_x(o)$, as it is shown in Figure 3. These output probabilities are approximated by single Gaussian distribution. Finally, the most likely state sequence of each HMM is calculated using Viterbi algorithm to represent the phoneme S.

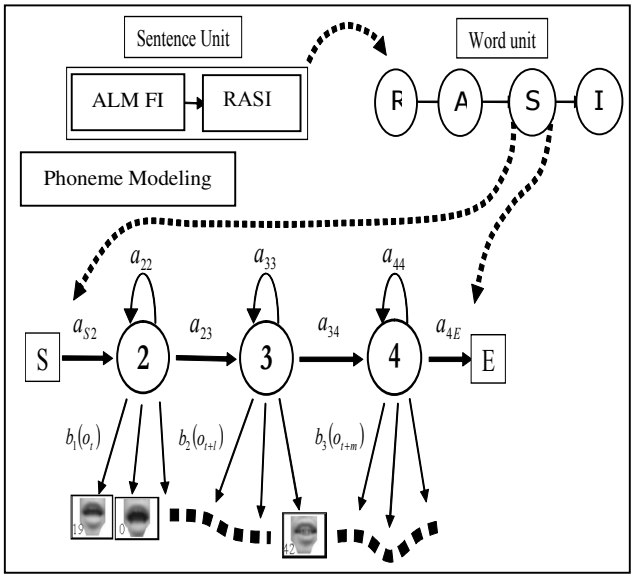


Fig. 3. HMM-Based Lip-Reading for ALM FI RASI

6. EXPERIMENTS

6.1 Arabic Database

Arabic language is considered as the sixth most widely spoken language in the world. The estimated number of Arabic native speakers is more than 300 millions in the area of Middle East only. Despite this fact there is no remarkable research in visual speech recognition domain based on Arabic language. According to our knowledge, our system presented here is the first contribution for visual Arabic speech recognition.

Due to rare contributions of Arabic language in visual speech field we captured our database by ourselves. Namely, it includes 9 full sentences uttered by 9 different Arabian subjects male and female, such that each subject uttered all sentences only one time. Most sentences consist of 3 words. Table 1 below shows the uttered sentences and their English meaning.

It has been remarked that a considerable part of the existing work in visual speech recognition domain has been explored only by small data sets such as digits from 1 to 10 or isolated words or words repeated more than one time by same speaker as the systems in [5], [6], [7] and [12]. Of course handling full grammatical sentences with a complete meaning has a higher challenge than isolated words.

Table 1. Arabic sentences database

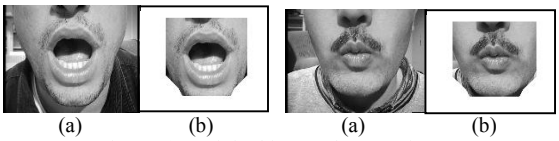
English Meaning	Arabic sentence
1- A pain in my teeth	١- الم فى ضروسى

2- A headache in head	- صداع فى راسى
3- A swelling in my back	- ورم فى ظهرى
4- A pain in my gum	- الم فى لثتى
5- The Arabic Salutation	- سلام عليكم
6- A swelling in my leg	- ورم فى قدمى
7- A pain in my back	- الم فى ظهرى
8- A swelling in my tooth	- ورم فى ضرسى
9- A pain in my head	- الم فى راسى

6.2 Image database

We designed our database according to speaker-independent-based rule using different speakers for both training and testing phases. Using this rule enables us to investigate how well the proposed system generalizes to new speakers; see Table 2 later. Speaker is not restricted to fix his head; just he or she should keep his or her mouth area in the supposed frame. The image database set includes 6480 gray scale and divided into two subgroups: First group, training group, has 4320 image gathered from 6 adults (male and female). The second group, test group, has 2160 images gathered from 3 speakers one of them belongs to the first group. Each uttered sentence represented by 80 visual frames.

In order to implement fast DCT, the size of input images should be as a power of 2 (i.e. 2^x). So the resolution of each input image in both data sets is 128x128 pixels. It is known in visual speech recognition field that the region of interest (ROI) is the region which contains lips and the neighboring part of mouth that undergoes movement when a subject speaks. Here, to detect the ROI, we define a constant window for all images. The part of the image that falls inside this window defines ROI and retained. As the rest part of the image which lies outside of this window is useless so it is truncated to speed up the computation as it is shown in Figure 4. Note that this operation of defining ROI with constant window is automatic and doesn't affect the recognition of vowels or the mouth shifted positions.



(a) (b) (a) (b)

Fig. 4. (a) Original image (b) Input images

7. DISCRETE COSINE TRANSFORM

It is known that the orthogonal transform has two properties:

1. It decorrelates the signal in the transform domain,
2. It contains the most variance (or energy) in the fewest number of transform coefficients.

DCT is the best orthogonal transform in comparison with the KLT (Karhunen-Loeve transform). Beside the high compaction of energy the availability to achieve fast implementation give DCT another advantage. The main disadvantage of DCT is that it is not a shift invariant recognition approach, so a precise tracking for the area around the lips have to be done before the DCT coefficients calculation [12], [18].

7.1 Accuracy and Coefficients Selection

It is known between DCT researchers that the coefficients with higher energy are including much information than those with less energy such that we can easily neglect them [11], [17]. Thus, in this paper, we select DCT coefficients according to the highest energy based rule. Consider the DCT coefficients are C_i, $1 \le i \le$ (128x128). Then the L selected coefficients are between $2 \le L \le 30$. Table 2 shows the accuracy results of some of those selected coefficients for training and test data for word and sentence units.

To judge the suitable number of coefficients which can give enough information about the object we calculated the ratio of the energy summation of the selected coefficients to the energy summation of the whole coefficients which denoted as Cumulative Proportional; namely

Cumulative **P**roportional (**CP**) = C_L/C_i, for each L.

The results shown in Table 2 indicate that **CP** and accuracy are increase as the number of selected coefficients are increase starting from C_L =2 , and they attain the maximum value at C_L =9. Later, if L increases beyond 9, **CP** will grow up slowly whereas the testing accuracy will remains the same or degrades gradually. This means that FDCT uses around 76% from the shown information using 9 coefficients only to give best accuracy for the Arabic data set. If we use more coefficients (or more information) training task becomes difficult and recognition accuracy remarkably deteriorates or not improved. Final remark on Table 2, that there is a big difference (or dwindling) between training data and test data results. In our thinking, this shows the non generalization nature of FDCT.

Table 2. Recognition accuracy and CP for DCT coefficients

No of Coef	Training Data		Testing Data		**CP** (%)
	Word	Sent	Word	Sent	
2	71.2	48.2	50.0	22.2	50.5
3	83.3	70.4	67.6	40.7	56.2
4	92.3	87	62.8	40.7	61.7
5	96.8	92.6	69.3	45.2	66.7
9	**98.1**	**94.3**	**70.5**	**45.7**	**76.4**
20	100	100	67.4	45.2	84.5
30	100	100	67.4	45	87.9

8. EXPERIMENTAL RESULTS & COMPARISON

Table 3 shows a comparison of recognition results using our proposed systems (HCM+HMM) with both (SOM+HMM) and (FDCT+HMM) in case of using a full rank covariance matrix with single Gaussian. All experiments are performed using same database, same word dictionary and same phoneme models. The results shown in Table 3 are for test data (new subjects), where the value in parentheses indicates the best number of dimensions of feature space used in each system. Fairly comparison demonstrates the superiority of HCM than FDCT, especially for sentence accuracy. This has the meaning that HCM can extract the visual features in the feature space in a better than FDCT. In addition, comparison between HCM and SOM also asserts the superiority of HCM over SOM as a feature extractor.

Table 3. Comparison of Recognition Accuracy

Target	*HCM* (2)	SOM (3)	DCT (9)
Word	**74.4**	60.3	70.5
Sentence	**55.5**	44.4	45.7

In general, we can remark in Tables 2 and 3 the dwindling of recognition results. This dwindling may due to the nature of the Arabic letters when they have been used in speech. Strictly speaking in Arabic language there are couples of letters have different sounds, but when the speaker vocalizes one letter of each couple the mouth will take the same shape (or appearance) of the other letter. These couples are (ص and ض), (ع and غ), (ك and ق), (س and ش), (ط and ت). Of course, this feature may exist in other languages, but we think that it is embedded clearly in Arabic language. This dwindling, also, may be caused by one or all of the following factors:

1. The estimated covariance was too small.
2. The number of learning data was not enough.
3. The optimum number of HMM state is not known for each Arabic phoneme.

On the other hand, FDCT based method is not shift invariant recognition methods [3], [12]. So a precise tracking for mouth area (or ROI) should do before starting experiment. In contrast, the capability of HCM to work through shift, invariant and rotate objects is pointed out in a previous report [8]. Figure 4 shows snapshots for some images in our database where the users shift their faces right or left up or down without any restrictions.

Fig. 4. Snapshots for shift invariant (not centered) objects

9. CONCLUSION

In this paper, we proposed a novel visual speech features representation system. The proposed system is a combination of HCM, in 2 dimensions, and HMM. The system performance is examined using multiple sentences of Arabic language. In parallel, we implemented fast DCT based method for same data base. Using this database, FDCT showed its higher accuracy using only 9 coefficients. Fair comparison using same database turned out that our system accuracy results are higher than SOM, 3 dimensions, and FDCT, 9 dimensions. This means that HCM able to extract the visual feature better than both FDCT and SOM. Additionally, HCM can still better in case of shift invariant object recognition whereas FDCT or SOM drops drastically in this situation. Finally, we showed, experimentally, the non-generalization of FDCT.

REFERENCE

[1] S. Takahashi, T. Morimoto, S. Maeda, and N. Tsuruta: "Dialogue Experiment for Elderly People in Home Health Care System," Lecture Notes in Computer Science, Springer (TSD 2003 Proc), 2807,pp.418—423.

[2] L.R. Rabiner, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc of the IEEE, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

[3] X. Zhang, R. Mersereau, M. Clements and C. Broun: "Visual Speech Feature Extraction for Improved Speech Recognition," proceeding of the IEEE international Conf. on Acoustic, Speech and Signal Processing ICASSP02, 2002.

[4] G. Potamianos, H. P. Graf and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lip Reading", in the Proc. of the Int. Conf. on Image Processing, Vol. 3, pp. 173-177, 1998.

[5] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," IEEE Transaction on Multimedia, Vol. 2, No.3, 2000.

[6] M. T. Chan, "HMM-Based Audio-Visual Speech Recognition Integrating Geometric- and Appearance-Based Visual Features," Proc. of IEEE Workshop on Multimedia Signal Processing, pp. 9-14, 2001.

[7] I. Matthews, T. Cootes, A. Bangham, S. Cox and R. Harvey, "Extraction of Visual Features for Lip-Reading," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, 2002.

[8] N. Tsuruta, R. Taniguchi and M. Amamiya," Hyper-column Model: A Combination Model Hierarchical Self- Organizing Maps and Neocognitron for Image Recognition," System and Computer in Japan, Vol. 31, No. 2, 2000.

[9] T. kohonen, "Self Organizing Maps," 3rd edition, Springer series in information sciences, 2001.

[10] T. Chen," Audiovisual Speech Processing, Lip Reading and Lip Synchronization," IEEE Signal Processing Magazine, pp. 9-21, Jan. 2001.

[11] K. Rao and P. Yip,"Discrete cosine transform : algorithms, advantages, applications," Academic Press, 1990.

[12] M. Heckmann, K. Kroschel, C. Savariaux and F. Berthommier, "DCT-Based Video Features For Audio-Visual Speech Recognition," Proc. Of Inter. Conf. on Spoken Language Processing, ICSLP02, pp. 1925-1928, 2002.

[13] E. D. Petajan, "Automatic lip-reading to enhance speech recognition," IEEE Global Telecommunications Conf: 265-272, 1984.

[14] B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," IEEE Communication Mag., Vol. 27, pp. 65-71, 1989.

[15] D. G. Stork, G. Wolff and E. Levine," Neural networks lip-reading system for improved speech recognition," in Proc. of Int. Joint Conference on Neural Networks, pp. 285-295, 1992.

[16] A.J. Goldschen, O.N. Garcia and E. Petajan," Continuous optical automatic speech recognition by lip-reading," In the 28th Annual Asilomar conference on Signal Systems, and Computer, Vol. 1, pp. 572-577, 1994.

[17] R. Gonzalez and R. Woods, "Digital image processing," 2nd ed. Prentice Hall international, 2002.

[18] G. Potamianos, A. Verma, C. Neti, G. Iyengar and S. Basu, "Cascade Image Transform for Speaker Independent Automatic Speechreading," Proceedings of the IEEE International Conference on Multimedia and Expo, 2000.

[19] A. Sagheer, N. Tsuruta, R. Taniguchi, S. Maeda, "Visual Speech Features Representation for Automatic Lip-Reading," Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP05, Vol 2, pp: 781 – 784, 2005.

[20] K. Fukushima, " Neocognitron," Biol Cyberetics, Vol. 36, No. 4, pp. 193-202, 1980.

[21] J. Lampinen and E. Oja, "Clustering Properties of Hierarchical Self Organizing Maps," J. of Math. Imaging and Vision, Vol. 2, 1992.

[22] HMM Tool Kit, [Online] : http://htk.eng.cam.ac.uk