Proc. of the Intl. Conf. on Advances in Computing and Communication – ICACC 2013
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-981-07-6260-5

Password Authentication through lip reading using human machine interface

Vibhanshu Gupta
Electronics and Telecommunication Dept.
V.E.S.Institute of Technology
Mumbai,India

Sharmila Sengupta
Electronics and Telecommunication Dept.
V.E.S.Institute of Technology,
Mumbai,India

Abstract— **This paper is about a lip reading technique for speech recognition by using motion estimation analysis. It presents a user authentication system based on password lip reading. Motion estimation is done for lip movement image sequences representing speech. In this methodology, the motion estimation is computed without extracting the speaker's lip contours and location. This leads to obtaining robust visual features for lip movements representing utterances. The proposed methodology comprises of two phases, a training phase and a recognition phase. In both the phases an n x n video frame of the image sequence for an utterance (password) is divided into m x m blocks. This method calculates and fits eight curves for each frame. Each curve represents motion estimation of this frame in a specific direction. These eight curves are representing set of features of a specific frame and are extracted in an unsupervised manner. The feature set consists of the integral values of the motion estimation. The feature sets are used to characterize specific utterances with no additional acoustic feature set. A corpus of utterances and their motion estimation features are built in the training phase. The recognition phase is accomplished by extracting the feature set, from the new image sequence of lip movement of an utterance, and compares it to the corpus using the mean square error metric for recognition.**

Keywords*—component, formatting, style, styling, insert *(key words)

I. Introduction

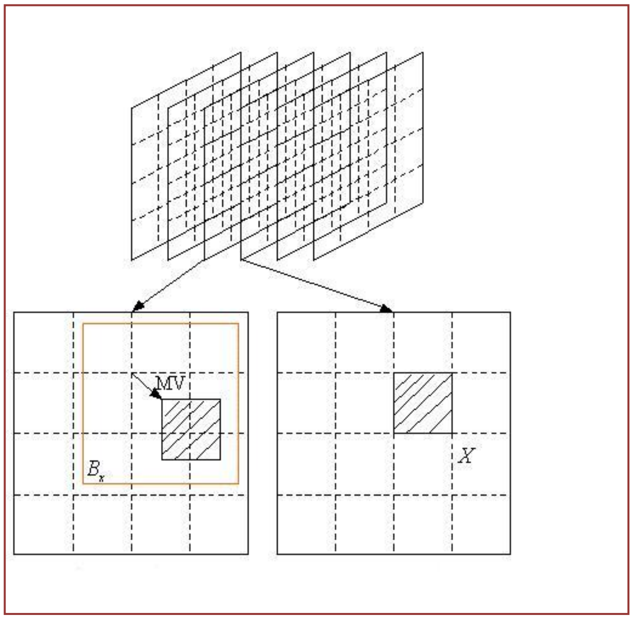
Automatic Speech Reading (ASR)[6] systems are playing successful roles in recognizing speech with high accuracy rates. Lip-reading [5] has become a hot topic for human computer interaction (HCI) and audio-visual speech reading (AVSR)[4]. Lip reading systems can be utilized in many applications such as hearing impaired aid and for noisy environment, where speech is highly unrecognizable and as password entry scheme. This paper concentrates on the visual only lip-reading system; feature extraction is a crucial part for a lip-reading system[1]. Various visual features have been proposed in general, feature extraction methods are pixel based where features are employed directly from the image or lip contour based, in which a detection model is used to extract the mouth area or some combinations of the two methods. The visual features can also be detected without extracting the lip locations and contours[2]. The proposed feature extraction methods utilize motion analysis of image sequences representing lip movement while uttering some speech.

II. Proposed Method:

The proposed scheme for implementation of project is the motion estimation analysis for robust speech recognition from lip reading. Visual features are extracted from the image sequences and are used for model training and recognition. Block based motion estimation techniques[3] are used to extract visual features blindly without any prior knowledge of lip location.

A. *Motion Estimation Analysis:*

Motion estimation removes temporal redundancies among video frames and is a computation intensive operation in the video encoding process. Block based schemes assume that each block of the current frame is obtained from the translation of some corresponding region in a reference frame. Motion estimation tries to identify this best matching region in the reference frame for every block in the current frame.



X: Source block for block-matching

MV: Motion vector Bx : Search are associated with X

Fig. 1 block based motion estimation using MV field obtained with Blocks of 8 x 8 pixels

In fig. 1, the gray block on the right corresponds to the current block and the gray area on the left represents the best match found for the current block, in the reference frame. The displacement is called the motion vector. The search range specified by the baseline h.263 standard allows motion vectors to range between –15 pixels and 16 pixels in either dimension. The size of the search window is of size 32 x32 about the search centre. Block matching algorithm (BMA) for motion estimation has been widely adopted by the current video compression standards, such as h.261, h.263, mpeg-1, mpeg-2, mpeg-4 and h.264 [1] due to its effectiveness and simple implementation. The most straightforward BMA is the full search (FS), which exhaustively evaluates all the possible candidate blocks within the search window. However, this method is very computationally intensive, and can consume up to 80% of the computational power of the encoder. This limitation makes me the main bottleneck in real-time video coding applications including lip reading systems. Consequently, fast BMAs are used to decrease the computational cost with the expense of less accuracy in determining the correct motion vectors. Many fast BMAs were proposed, such as three-step search (TSS), four-step search (4SS), block-based diamond search (DS),Full Search (FS) algorithms etc.

B. *Full-Search Block-Matching Algorithm*

Full-search block matching algorithm (FS) finds the best match for a reference block in the current frame within a search area in the previous frame. The criterion for best match is the candidate block with the minimum amount of distortion when compared with the reference block. The measure used for calculating distortion is the sum of absolute differences (SAD) of intensity values between the two blocks. The SAD for the candidate block of size n x n at position (u,v) can be defined as:

$$\mathrm{SAD}\,(u, v) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \, u\,(\, i+u \,,\, j+v) \;-\; v\,(\, i, j\,) \, \right| \qquad (1)$$

where v (i,j) and u (i+u , j+v) are intensity values at position (i,j) of the reference block and (i+u, j+v) of the candidate block in search area s. The search area is formed by extending the reference block by a search range w on each side forming a search area of (2w+n)2 pixels. As a result, there are (2w +1) candidate blocks in both horizontal and vertical directions i.e. a total of (2w+1)2 candidate blocks have to be searched corresponding to each reference block. The distortion value is computed for each candidate block and the minimum value SAD min is found. The block matching process generates a motion vector (u,v) min and the corresponding distortion value SAD min.

III. The Proposed Technique for lip reading:

The proposed technique is composed of two phases, the first phase is the training phase which results in feature extraction from image sequences representing different utterance lip movement, the second phase is the recognition phase, where new utterance through lip movement is compared against the output of the training phase and recognized.

A. *Training Phase and Feature Extraction*

An image sequence is captured with the frame rate of 30 frames/sec and the resolution size of 360×240 .Block-based motion estimation technique is used, motion vectors representing motion of block is computed from a pair of consecutive images. Block matching algorithm can be full search or 3SS or FSS or DS. The previously mentioned algorithm will produce motion vectors with values from -3 to+3 in one of the eight geographical directions.This restriction in defining motion vectors is due to the fact that lip motion in utterance is very restricted at the rate of 15 frames per second, motion is very slow.

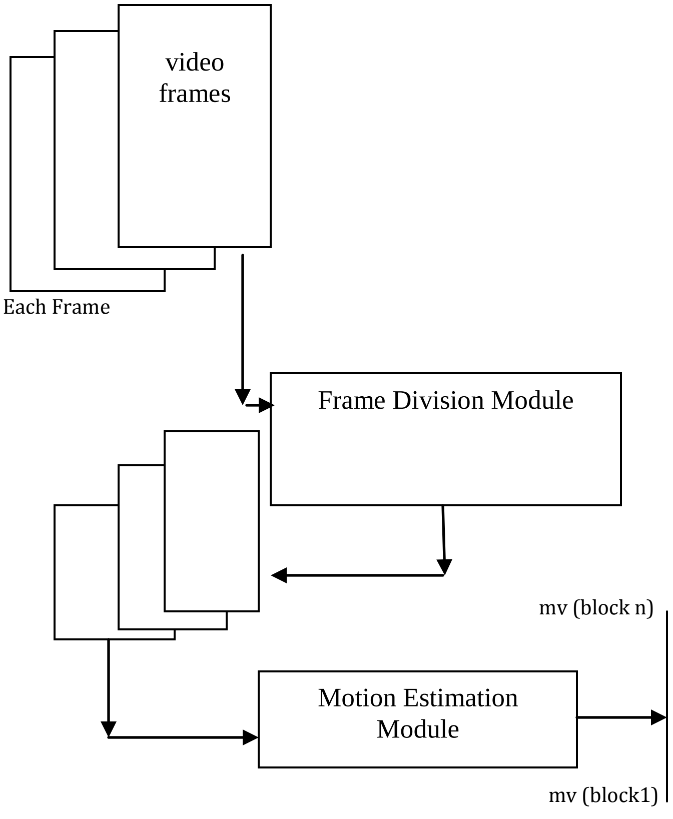


Fig. 2 Block1 (Motion estimation and motion vector extraction for one video sequence)

B. *Algorithm –Training Phase:*

The diagram in fig. 2 illustrates block 1 in the training phase algorithm. Each video frame, of the utterance lip movement sequence, is fed to the frame division module into 8x8 blocks, each of these blocks (block1 to block n) are fed to the motion estimation module to for motion analysis and production of the motion vector of such block. A set of n motion vectors {mv (block1), …. mv (block n)} are produced.

Many videos for the same utterance are fed iteratively to block 1 to calculate average motion vectors for each block in a frame of these videos. The set of the average motion vectors are fed to block 2. where eight curves are built from the average motion vectors for each frame. Each curve is the value of the average motion vector of a block versus its particular location in a video frame. Each curve represents the average motion of blocks in the video frame in a certain direction (we restricted the directions to the eight geographical directions) and this curve is fitted to be a continuous curve. Such curve represents a motion feature of the utterance video. The motion feature will be represented by the area under this curve by taking the integration value of this curve, which is done by the area calculation module, the area is calculated as in equation 1. Number of motion features for an utterance video is equal to 8f, where f equals to the number of frames in the utterance video. Each utterance and its motion features are stored in the utterance database.

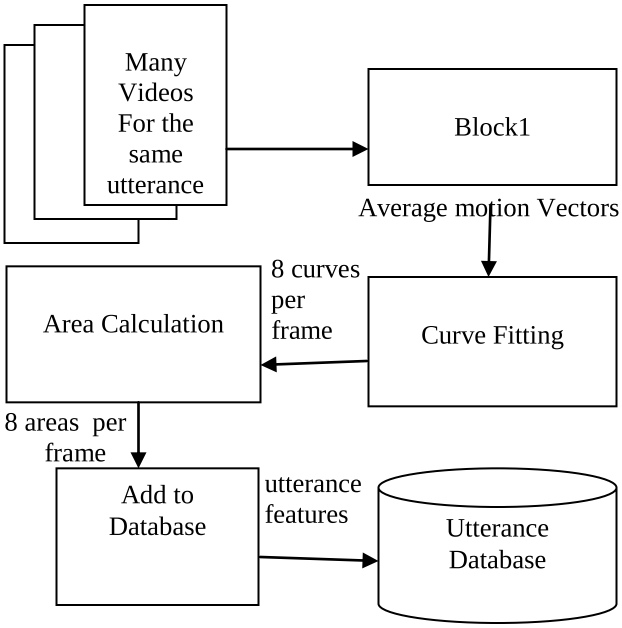


Fig .3 populating the utterance database

C. *Algorithm –Training (word: w)*

```
 Begin
For word w repeating the following steps j times by different speakers
{
1. Recording a video of lip reading the word w by a speaker Sj;
2. Dividing the video into n frames;
3. Do for J(k= 0, k=k+2, k< n)
{
a. frame k is divided into m blocks each of size 8x8 pixels;
b. motion vectors, M, of all the m blocks are calculated between frame k and frame k+1;
M = set of motion vectors = {mvi, i=1 to 8}, i is one of the eight principle geographical directions;
c. for i= 1 to 8 do
{ 1. draw and fit a discreet graph:
DG(k) between mvi and the location of the block, location of the block is numbered in a spiral fashion starting from the center of each 64x64 block;
2.Area (k) = ∫ DG (k)      --------------   (2)
}
}
4. feature set j = {area (k), i =1 to 8 ,¥ k}
}
Training-feature-set (W) =
{Average (area (k)), i= 1 to 8 ,¥ k}
End
```

D. *Explanation of the training algorithm*

1. The video sequence of the lip read word has n frames; each frame is divided into blocks of 8x8 pixels for the motion vectors calculations, as we assume that an 8x8 block moves translations motion as a one unit.
2. To draw the graphs DG (k), a frame is divided into blocks each of 64x64 pixels, or 8x8 blocks of size 8x8 pixels. Location of the block is numbered in a spiral fashion starting from the center of each 64x64 block.
3. There are 8 curves representing the motion vectors of the video sequence.
4. To calculate the integration of each curve, area under the curve is calculated by an approximate method.

IV. Recognition Phase*:*

Lip reading recognition phase is similar to the training phase, it starts with the unknown utterance lip movement video, these utterances should be recognized. This video is fed to block 1, where motion vectors for this video are calculated and fitted into 8f curves. Integral values of these curves are the areas under these curves and are fed to the comparison module. The comparison module compares the motion features of the input video against the utterance stored in the database by using the mean square error function. MSE is calculated between the input utterance features and every utterance features stored in the database. MSE is calculated as shown in equation 2. MMSE is computed to choose the candidate utterance that is most similar to the input utterance. MMSE is calculated as in equation 3.

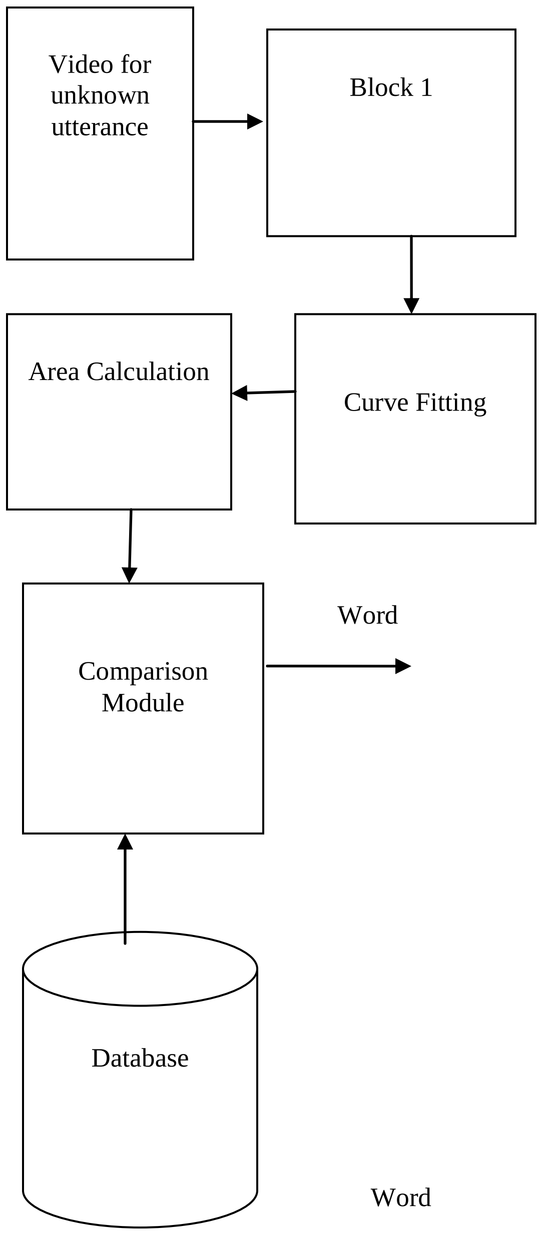


fig. 4 lip reading recognition phase

A. *ALGORITHM RECOGNIZE (WORD: VIDEO)*

```
Begin
Divide the word video into n frames;
1. Do for (k= 0, k=k+2, k< n)
{
a. frame k is divided into m blocks each of size 8x8 pixels;
b. motion vectors, m, of all the m blocks are calculated between frame k and frame k+1;
M = set of motion vectors = {mvi, i=1 to 8}, i is one of the eight principle geographical directions;
c. for i= 1 to 8 do
{
1. Draw and fit a discreet graph:
DG(k) between mvi and the location of the block, location of the block is numbered in a spiral
fashion starting from the center of each 64x64 block;
2. Area i (k) = DG (k) i ;
}
}
2. Feature (word) = {f area (k) ,i =1 to 8 , ¥ k}
3. Calculate mean square error MSEj for the input utterance and utterance j as follows:
```

$$MSEj = \text{---} \sum_{x}^{1} \sum_{i=1}^{2} {}_{k=1} (f\ Area\ i\ (k) - Area\ i\ (k)\} \quad \text{------------(eq.3)}$$

```
Calculate minimum mean square error MMSE as follows:
MMSE= min (MSEj )      j = 1 to x        ------------(eq.4)
4. if MMSE > threshold then the word is
Unrecognizable otherwise the WORD = Wd (Wd is a word in the corpus corresponds to the minimum MSE)
END
```

V. Experimental Results:

The images contain face and the mouth area of the speaker and are digitized at 30 frames/sec, 720 x 480 pixels and 8 bits/pixel. The first utterance of each word by a speaker can be used as the training set for the motion vector feature set extraction in training phase. The second instances can be used as the test set for testing the recognition phase. In the end, a total of 8 features/block, with a total of 8 curves/utterance were fed to the recognizer in the form of areas.

G is the number of blocks in each video and is calculated equation 5 as follows.

f is the number of frames per second, t is the duration time of the video of the utterance,

b is the size of the blocks is the size of the frame,

Where, $G = ((s / b) \times t \times f)$ ------------------------- (Eqn.5)

The expected experimental results must show that the proposed method achieves training phase and recognition phase using full search block matching algorithm and a lip reading password recognition system is designed as user authentication to Human Computer Interface.If the password threshold is less than MMSE then only it is authorized otherwise authentication will be unsuccessful.Some results are given in the table.

Password	Threshold	MMSE	Authentication Successful	Authentication failure
barge	**0.8**	**0.7**	**yes**	**no**
beg	**0.8**	**0.9**	**No**	**yes**
book	**0.8**	**0.6**	**yes**	**no**
count	**0.8**	**0.5**	**yes**	**no**
cup	**0.8**	**1.0**	**No**	**yes**

Table -1 (Results)

REFERENCES:

[1] S.Furui, “speech recognition technology in the ubiquitous/ wearable computing environment,” in proc. icassp2000, vol. 6,2000, pp. 3735–3738.

[2] Hanan Mahmoud “Reducing Shoulder-surfing by using silent speech password entry” technical report, KSU, center of excellence in information assurance, November 2008.

[3] C. Miyajima, K. Tokuda, and T.Kitamura, “audio-visual speech recognition uses MCE-based HMMs and model-dependent stream weights,” in proc. icslp2000, vol. 2, 2000, pp. 1023–1026.

[4] K. Iwano S. Furui T. Yoshinaga, S. Tamura,“Audio-visual speech recognition using lip movement extracted from side-face images,”proc. Auditory Visual Speech Processing(AVSP), pp. 117–120, 2003.

[5] G. Potamianos p. Lucey, “Lipreading using profile versus frontal views,” IEEE multimedia signal processing workshop, pp. 24–28, October 2006.

[6] T. Chen, “audiovisual speech processing. Lip reading and lip synchronization,” ieee signal processing mag., vol. 18, pp. 9–21, January 2001.