Real Time Characterization of Lip-Movements Using Webcam

Suresh D S, Sandesh C

Director & Head of ECE Dept, CIT, Gubbi, Tumkur, Karnataka.

PG Student, ECE Dept, CIT, Gubbi, Tumkur, Karnataka.

Abstract: Communication plays an important role in the day-to-day activities of human beings. The main objective of this paper is to help the people who are unable to speak. Visual speech plays a major role in the lip-reading for listeners with deaf person. Here we are using local spatiotemporal descriptors in order to identify or recognize the words or phrases from the disable (dumb) people by their lip-movements. Local binary patterns extracted from the lip movements are used to recognize the isolated phrases. Experiments were made on ten speakers with 5 phrases and obtained accuracy of about 75% for speaker dependent and 65% for speaker independent. While being made comparison with other concepts like AV letter database, our method outperforms the other by accuracy of 65%. The advantages of our method are robustness and recognition is possible in real time.

Key words: LBP Top, Webcam, Visual speech, KNN classifier, Silent speech interface.

I. Introduction

In recent days, a silent speech interface has becoming an alternative in the field of speech processing. A voice based speech communication has suffered from many challenges which arises due to fact that speech should be clearly audible, it cannot be masked, includes lack of robustness privacy issues and exclusion of speech disabled person. So these challenges may be overcome by the use of silent speech interface. Silent speech interface is a device, where system enabling speech communication takes place, without the necessity of a voice signal or when audible acoustic signal is unavailable. In our approach, lip movements are captured by the use of a webcam, which is placed in front of the lips. Many research work focuses only on visual information to enhance speech recognition[1]. Audio information still plays a major role than the visual feature or information. But, in most cases it is very difficult to extract information. In our method we are concentrating on the lip movement representations for speech recognition solely with visual information.

Extraction of set of visual observation vectors is the key element on AVSR(audio-visual speech recognition) system. Geometric feature, combined feature and appearance features are mainly used for representing visual information. Geometric feature method represents the facial animation parameters such as lip movement, shape of jaw and width of the mouth. These methods require more accurate and reliable facial feature detection which are difficult in practice and impossible at low image resolution.

In this paper, we propose an approach for lip reading or lip movements, where human-computer interaction improves significantly and understanding in noisy environment also improves.

II. Local Spatiotemporal Descriptors for Visual Information

In this paper, we are concentrating on LBP-TOP in order to extract the features from the extracted video frames. The Local Binary Pattern (LBP) operator is a gray-scale which is not varied, always remains the same texture, simple statistic, which has shown the best performance in the classification of different kinds of texture. For an individual pixel in an image its respective binary will be generated and the thresholds will be compared with the its neighborhood value of center pixel as shown in Figure 1(a) [1].

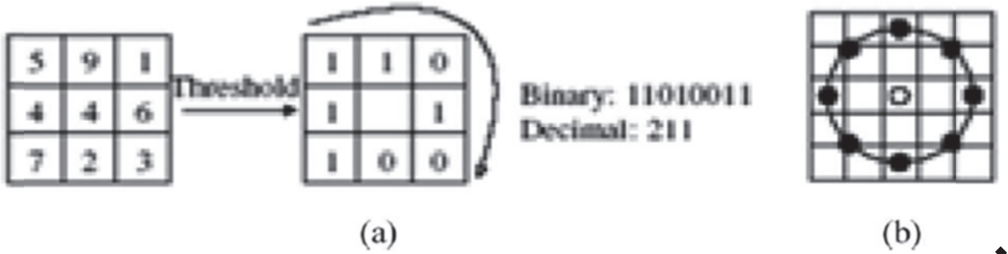

Figure. 1. (a) Basic LBP operator. (b) Circular (8,2) neighborhood.

$$LBP_{P,R} = \sum_{p=0}^{p=1} s(g_p - g_c)2^P,\ s'(x) = \begin{cases} 1, & x>0 \\ 0, & x<0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where gc denotes the grey value of center pixel{xc, yc} of the local neighborhood and gp is related to the grey values of P equally spaced pixels on a circle of radius R. A histogram is created in order to collect up the occurrences of various binary patterns. The definition of neighbors can be extended to circular neighborhoods with any number of pixels as shown in Figure 1(b). By this way, we can collect larger-scale texture primitives or micro-patterns, like spots, lines and corners.

Local texture descriptors have obtained tremendous attention in the analysis of facial image because of their robustness to challenge such as pose and illumination changes. In our approach a temporal texture recognition using local binary patterns were extracted from the three orthogonal planes(LBP-TOP). LBP-TOP method is more efficient than the ordinary LBP. In ordinary LBP we are extracting information or features in two dimension, where as in LBP-TOP we are extracting information in three dimension i.e, X, Y and T. For LBP-TOP, the radii in spatial and temporal axes X, Y, and T, and the number of neighboring points in the XY, XT, and YT planes can also be different and can be marked as RX, RY and RT, PXY, PXT, PYT. The LBP-TOP feature is then denoted as LBP-TOP PXY, PXT, PYT, RX, RY, RT. If the coordinates of the center pixel gtc,e are (xc, yc, tc), than the coordinates of local neighborhood in XY plane gXY,p are given by (xc-RXsin(2πp/PXY),yc+RYcos(2πp/PXY), tc), the coordinates of local neighborhood in XT plane gXT,p are given by (xc-RXsin(2πp/PXT),yc, tc –RT cos(2πp/PXT)) and the coordinates of local neighborhood in YT plane gYT,p (xc, yc-RYcos(2πp/PYT),, tc –RT sin(2πp/PYT)). Sometimes, the radii in three axes are the same and so do the number of neighboring points in XY, XT, and YT planes. In that case , we use LBP-TOPP,R for abbreviation where P=PXY=PXT=PYT and R=RX=RY=RT[1].

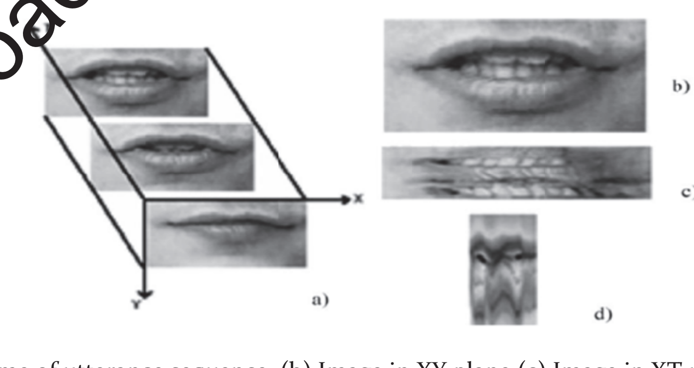

Figure. 2. (a) Volume of utterance sequence. (b) Image in XY plane (c) Image in XT plane(d) Image in TY plane.

Figure 2(a) demonstrates the volume of utterance sequence. Figure 2(b) shows image in the XY plane. Figure 2(c) is an image in the XT plane providing a visual impression of one row changing in time, while Figure 2(d) describes the motion of one column in temporal space[1]. An LBP description when computed over the whole utterance sequence it encodes only the occurrences of the micro-patterns without any indication about their locations. In order to overcome this effect, a representation which consists of dividing the mouth image into several overlapping blocks is introduced. Figure. 3 also gives some examples of the LBP images. The second, third, and fourth rows show the LBP images which are drawn using LBP code of every pixel from XY (second row), XT (third row), and YT (fourth row) planes, respectively, corresponding to mouth images in the first row.

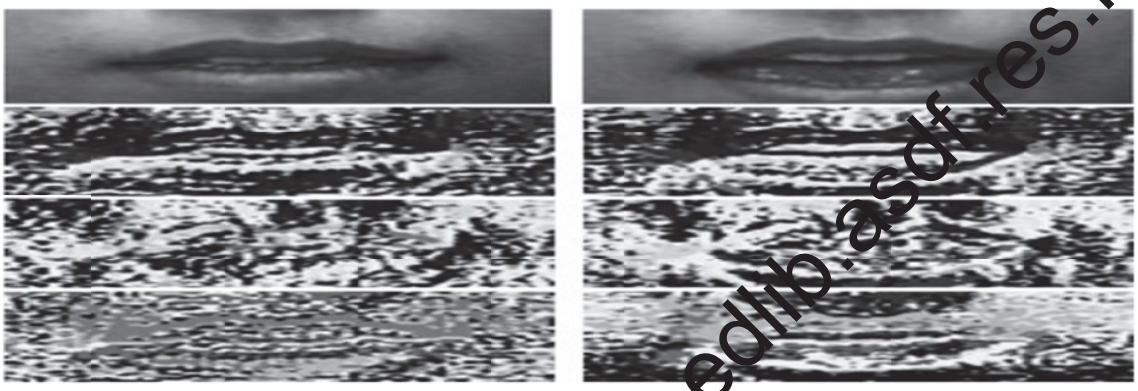

Figure 3. Mouth region images (first row), LBP-XY images (second row), LBP-XT images (third row), and LBP-YT images (last row) from one utterance[1].

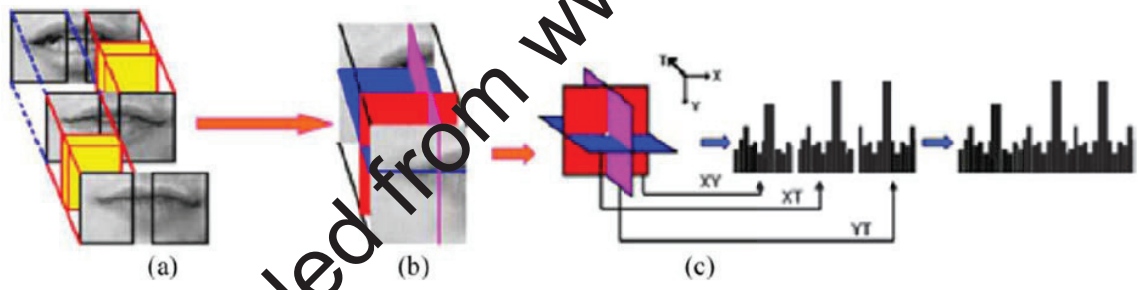

Figure. 4. Features in each block volume. (a) Block volumes. (b) LBP features from three orthogonal planes. (c) Concatenated features for one block volume with the appearance and motion[1].

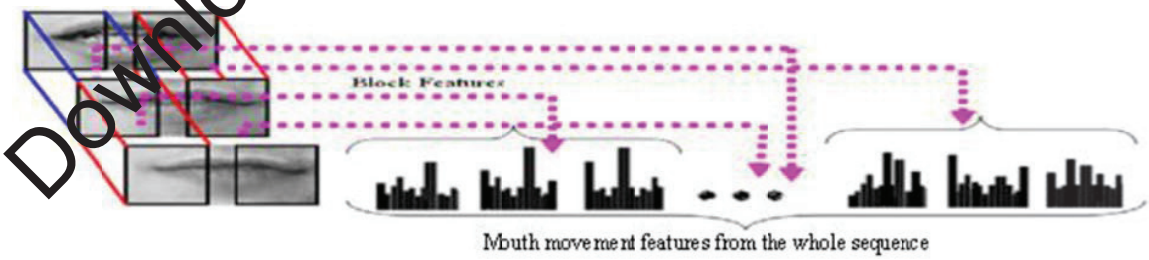

Figure. 5. Mouth movement representation[1].

When a person utters a command phrase, the words are pronounced in order, for instance “you-see” or “see-you”. If we do not consider the time order, these two phrases would generate almost the same features.

To overcome this effect, the whole sequence is not only divided into block volumes according to spatial regions but also in time order, as shown in Figure. 4(a) shows. The LBP-TOP histograms in each block volume are computed and concatenated into a single histogram, as shown in Figure. 4. All features extracted from each block volume are connected to represent the appearance and motion of the mouth region sequence, as shown in Figure. 5.

A histogram of the mouth movements can be defined as follows

$$H_{b_{f,d,j,i}} = I\{fj(x,y,t)=i\}, \; i=0,\ldots,\; nj-1;\; j=0,1,2 \qquad \ldots(2)$$

III. Multiresolution Features and Feature Selection

Multi resolution features will provide more accurate information and also the analysis of dynamic event will be increased. By using these multi-resolution features, it helps in increasing the number of features greatly. When the features from the different resolution were made to be connected in series directly, the feature vector would be very long and as a result the computational complexity will be too high. It is obvious that all the multi resolution features will not contribute equally, so it is necessary to find out features like which location, with what resolution and more importantly the types such as appearance, horizontal motion or vertical motion that are very important. We need feature selection for this purpose. In changing the parameters, three different types of spatiotemporal resolution are presented: 1) Use of a different number of neighboring points when computing the features in XY (appearance), XT (horizontal motion), and YT (vertical motion) slices; 2) Use of different radii that can capture the occurrences in different space and time scales; 3) Use of blocks of different sizes to create global and local statistical features[4][5].

IV. Our System

Our system consists of three stages, as shown in Figure 6. The first stage is a detection lip movements. The second stage extracts the visual features from the mouth movement sequence. The role of the final stage is to recognize the input utterance using an KNN classifier.

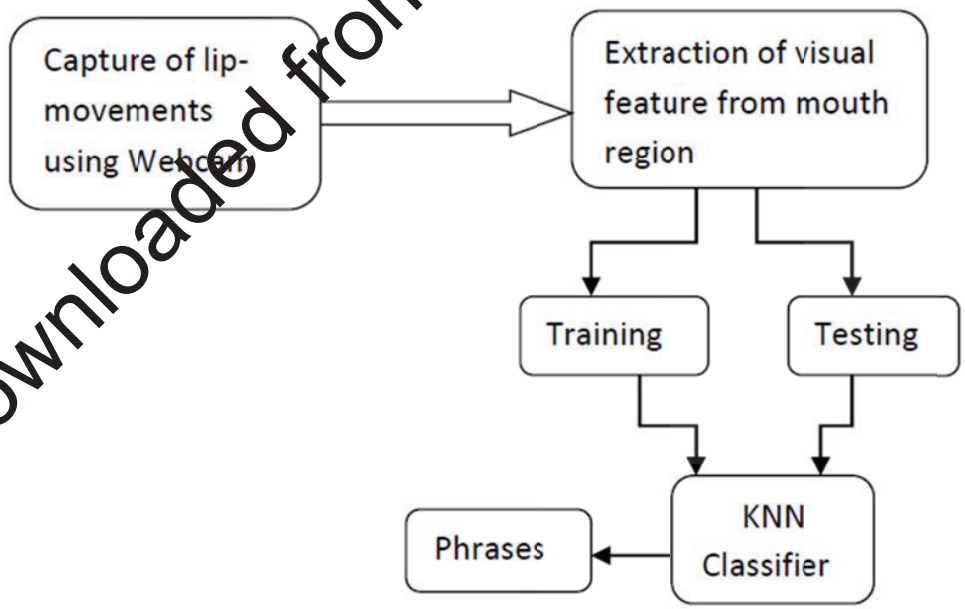

Figure 6. System diagram

In our approach webcam is used to capture the lip-movements. Further we will extract the visual feature from the mouth region. These features are then fed to the classifier. For speech recognition, a KNN classifier is selected since it is well founded in statistical learning theory and has been successfully applied to various object detection tasks in computer vision. Since the KNN is only used for separating two sets of points, the -phrase classification problem is decomposed into two-class problems, then a voting scheme is used to accomplish recognition. Sometimes more than one class gets the highest number of votes; in this case, 1-NN template matching is applied to these classes to reach the final result. This means that in training, the spatiotemporal LBP histograms of utterance sequences belonging to a given class are averaged to generate a histogram template for that class[1].

V. Experiments Protocol and Results

For more accurate evaluation of our proposed method, we made a design with different experiments, including speaker-independent, speaker-dependent, multi resolution.

1. Speaker-Independent Experiments: For the speaker-independent experiments, leave-one-speaker-out is utilized. We made a training of particular phrase using one speaker and we made to say the same phrase with 10 different speakers and when testing was done we were successful in getting the same phrase from 6 speakers. The overall results were obtained using M/N (M is the total number of correctly recognized sequences and N is the total number of testing sequences). When we are extracting the local patterns, we take into account not only locations of micro-patterns but also the time order of lip-movements, so the whole sequence is divided into block volumes according to not only spatial regions but also time order. Figure. 7 demonstrate the performance for every speaker. The results from the second speaker are the worst, mainly because the big moustache of that speaker really influences the appearance and motion in the mouth region.

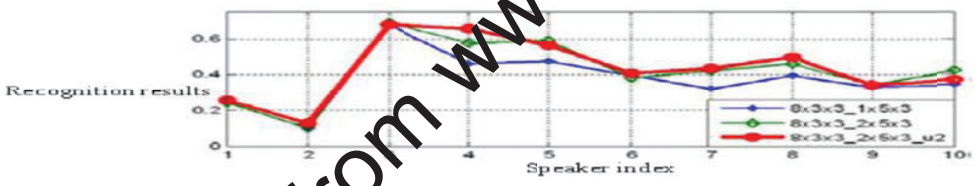

Figure. 7. Recognition performance for every speaker

2) Speaker-Dependent Experiments: For speaker-dependent experiments, the leave-one utterance-out is utilized for cross validation in our approach when a particular speaker is trained with some list of phrases and after testing is performed with the same speaker 70% of same phrases were identified correctly.

3) One-One versus One-Rest Recognition: In the previous experiments on our own dataset, the ten-phrase classification problem is decomposed into 45 two class problems ("Hello"- "See you", "I am sorry"- "Thank you", "You are welcome"- "Have a good time", etc.). But using this multiple two-class strategy, the number of classifiers grows quadratically with the number of classes to be recognized like in AVLetters database. When the class number is N, the number of the KNN classifiers would be N (N-1)/2.

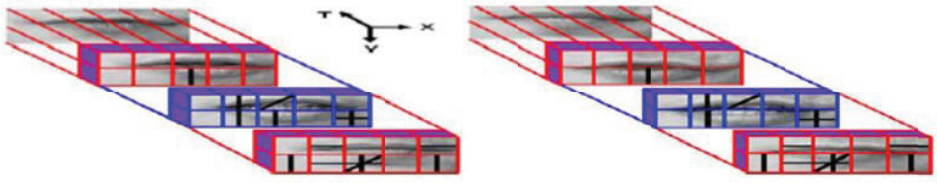

Figure. 8. Selected 15 slices for phrases "See you" and "Thank you". "|" in the blocks means the YT slice (vertical motion) is selected, and "_" the XT slice (horizontal motion), "/" means the appearance XY slice[1].

Figure-8 shows the selected slices for similar phrases “see you” and “thank you”. These phrases were the most difficult to recognize because they are quite similar in the latter part containing the same word “you”. The selected slices are mainly in the first and second part of the phrase; just one vertical slice is from the last part. The selected features are consistent with the human intuition.

Table I

Results from One-to-One and One-to-Rest Classifiers on Semi-Speaker-Dependent Experiments (Results in The Parentheses are From One-to-Rest Strategy)

Features	Blocks	Third test(O-R)	Three-fold(O-R)
$LBP\text{-}TOP_{8,3}$	1x5x3	58.92	63.45
$LBP\text{-}TOP^{u2}_{8,3}$	1x5x3	62.01	67.23

Conclusion

A real-time capture of lip-movements using webcam for recognition of speech was proposed, in order to help the disable person. Our approach uses local spatiotemporal descriptors for the recognition of input utterance. LBP-TOP is used to extract the features from the captured images. Experiments were made on ten speakers and the lip movements were converted into speech very efficiently. Compared to other approach, our method outperforms the other by accuracy of 70%.

References

1. Lipreading With Local Spatiotemporal Descriptors IEEE Transactions On Multimedia, Vol. 11, No. 7, November 2009.
2. T. Ahonen, A. Hadid, and M. Pietikäinen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, 2006.
3. G. Zhao and M. Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, 2007.
4. G. Zhao, M. Pietikäinen, and A. Hadid, “Local spatiotemporal descriptors for visual recognition of spoken phrases,” in Proc. 2nd Int. Workshop Human-Centered Multimedia (HCM2007), 2007, pp. 57–65.
5. G. Zhao and M. Pietikäinen, “Boosted multi-resolution spatiotemporal descriptors for facial expression recognition,” Pattern Recognit. Lett., Special Issue on Image/Video-Based Pattern Analysis and HCI, 2009, accepted for publication.