GLCM Feature Extraction for Visual Speech Recognition

Ajit S. Ghodke[1], Yuting Zhang[1], Ritesh A. Magare[2]

[1]International Education College, Neusoft Institute Guangdong, Foshan, 528225, China

[2]Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Maharashtra State, 431004, India

Abstract: Visual data based speech recognition is a field with a great potential to help solve difficult issues like audio corruption by noise. This paper addresses an application of various GLCM features to extract region based features for motion estimation of images. The features like contrast, energy, homogeneity, mean, standard deviation, entropy, variance, smoothness and IDM have high discrimination accuracy and requires less computation time and hence efficiently used for pattern recognition like speech recognition application. In this paper we used the visual dataset which is prepared by Samsung galaxy nxt mobile phone.

Keywords: GLCM (Gray Level Co-Occurrence Matrix); IDM(Inverse Different Moment); Feature extraction; Visual speech recognition

1. Introduction

Speech is the way in human communicates or interacts with each other. Visual speech is described as an expression of a speaking human face that is appealing to the human eye. Identification of speech from visual knowledge is called as recognition of visual speech or lip reading. This is important to determine the person who speaks when acoustic knowledge is not available. This is also the security way to identify the speech of criminals on public places like railway stations, bus stops, and airports. There are three major components of a speech recognition system: extraction of information, probabilistic feature processing and classification. The general approach is to extract the main components of lip movement with regard to the properties of the lip form in order to establish a one-to - one correspondence between speech phonemes and lip shape visemes.

2. Literature Survey

In literature many authors have studied many features as shown in the following table1.

Table 1. Features and classification techniques used

Reference number	Database used	Features used	Classification techniques used
[1]	----	Visual features such as mouth open/closed, tongue visible/not-visible, teeth visible/notvisible, and several shape descriptors of the mouth and its motion are all rapidly computable in a manner quite insensitive to lighting Conditions.	Neural network
[2]	A USB based webcam was used for capturing the video of a person while speaking ten different words.	The lip information is extracted using lip geometric and lip appearance features.	The neural network is used for word identification
[3]	Own established independent database (English pronunciation of Numbers from zero to nine by three males and three females).	Visual Geometry Group (VGG) network to extract the lip image features.	The convolutional neural network (CNN) used to image feature extraction is combined with a Recurrent neural network (RNN) based on attention mechanism for automatic lip-reading recognition.
[4]	Lip-reading of Urdu speech words and phrases, we constructed a Video-speech corpus.	MFCC features	Video sequences using spatiotemporal convolution neural network ,Bi-gated recurrent neural network and Connectionist Temporal

			Classification Loss. audio that inputs the MFCC features to a layer of LSTM cells and output the sequence.
[5]	The English digits 0 to 9 as available in the CUAVE database.	Lip geometry feature extraction,Shape-based lip features obtained from a single video frame i.	HMM for classification
[6]	English digits from zero to nine when speaking	Lip contour and real-time recognize each English Digit when speaking.	-----
[7]	MIRACL-VC1	----	VGG and LSTM model
[8]	MIRACL-V1	---	We explore a CNN + LSTM Baseline model, a Deep Layered CNN + LSTM model, an ImageNet Pretrained VGG-16 Features + LSTM model, and a Fine-Tuned VGG-16 + LSTM model.
[9]	Audio-visual speech dataset comprising 300 Japanese words with six different speakers.	Convolutional neural network (CNN) as a visual feature extraction mechanism for VSR	A hidden Markov model in our proposed system recognizes multiple isolated words.
[10]	GRID corpus dataset in which the videos are recorded from 33 speakers.	The features recognized include the height and width of the lips, the outside and inside edges of the lips, and angles between specific lip points.	Convolutional Neural Networks.
[11]	The GRID audiovisual dataset	A Gaussian mixture model (GMM) baseline system is developed using standard image-based two-dimensional discrete cosine transform (2D-DCT) visual speech features	Convolutional neural networks

3. Methodology

Following diagram a shows the proposed system

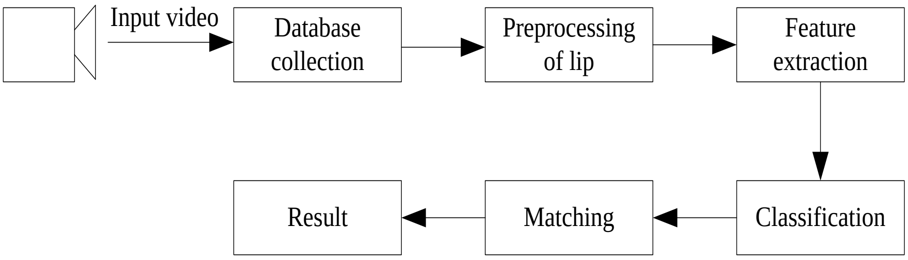
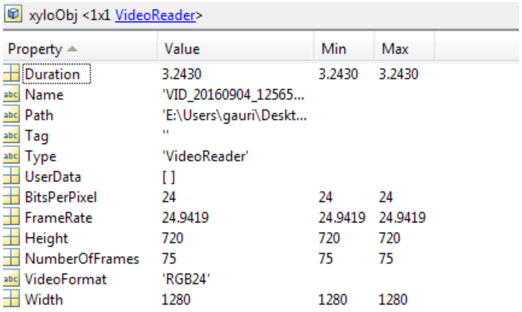


Figure 1. System architecture for visual speech recognition

3.1. Input video

To read a recorded video file following matlab command is used

XyloObj = Video Reader ('E: \Users \gauri \Desktop \RM \imp123 \video \VID_20160904_125655.mp4');



xyloObj <1x1 VideoReader>

Property	Value	Min	Max
Duration	3.2430	3.2430	3.2430
Name	'VID_20160904_12565...		
Path	'E:\Users\gauri\Deskt...		
Tag	''		
Type	'VideoReader'		
UserData	[ ]		
BitsPerPixel	24	24	24
FrameRate	24.9419	24.9419	24.9419
Height	720	720	720
NumberOfFrames	75	75	75
VideoFormat	'RGB24'		
Width	1280	1280	1280

Figure 2. Input video information

'Video Reader' command in MATLAB creates reader object for the video file. This object contains the metadata or the parameters related to the video such as Frame rate, Height and Width of the frames, duration of the video etc.

To read all the frames in the video, we can use two methods. The first method is to find the number of frames in the video and read it. The second method is to read the frames until no more video frames are available. in this, the frame rate is assumed to be constant throughout the duration of the video. At constant frame rate, the number of frames in the video is obtained by direct multiplication of frame rate and video duration. In our example, the video is 3.2430 seconds long and the frame rate is 24.0. Multiplying 3.2430 and 24 gives 75 that is 75 frames available in the video.

3.2. Database collection

In this we have recorded videos of 10 subjects which include male and female with different ages. Each subject has spoken 5 sentences from 5 different areas such as College, Government Office, Hospital, House and Restaurant and each sentence is repeated 5 times. We have selected 5 frequently used sentences from each area as shown in table 1. All the sentences are spoken in English language. The total sentences recorded in a database are 1250.

The format of the videos recorded is mp4 and all these videos are recorded by Samsung galaxy nxt mobile phone 13 megapixel camera by hand. The distance of the video recording is 4 feet.

3.3. Preprocessing

After database collection preprocessing has been done.

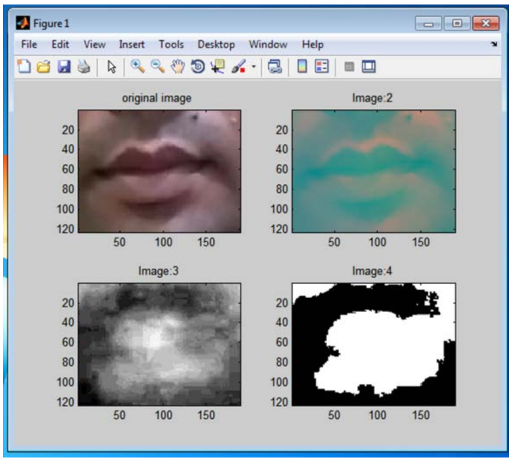


Figure 3. Preprocessing

3.3.1. Input image

Imread ( ) reads a grayscale or color image from the file specified by the string filename. If the file is not in the current folder, or in a folder on the MATLAB path, specify the full pathname.

3.3.2. Displaying the image

Imshow (I, [low high]) displays the grayscale image I, specifying the display range for I in [low high]. The value low (and any value less than low) displays as black; the value high (and any value greater than high) displays as white. Values in between are displayed as intermediate shades of gray, using the default number of gray levels.

3.3.3. Changing the color space

Makec form ('srgb2lab') creates the color transformation structure C that defines the color space conversion specified by type. To perform the transformation, pass the color transformation structure as an argument to the applyc form function. In this case the RGB color space is converted to LAB color space.

3.3.4. Finding gray thresh value

Graythresh (J(:,:,2)) computes a global threshold (level) that can be used to convert an intensity image to a binary image with im2bw. Level is a normalized intensity value that lies in the range [0, 1]. The gray thresh function uses Otsu's method, which chooses the threshold to minimize the in traclass variance of the black and white pixels. Multidimensional arrays are converted automatically to 2-D arrays using reshape. The gray thresh function ignores any nonzero imaginary part of I.

3.3.5. Converting to binary image

BW1=im2bw (J(:,:,2),L) converts the grayscale image I to a binary image. The output image BW replaces all pixels in the input image with luminance greater than level with the value 1 (white) and replaces all other pixels with the value 0 (black). Specify level in the range [0,1]. This range is relative to the signal levels possible for the image's class. Therefore, a level value of 0.5 is midway between black and white, regardless of class.

3.4. Feature extraction

In this processing the extracted feature of the lip pattern identifies whether lip open, lip closed, lip rounded or labial dental. Processing of the inner contour of the lip identifies whether lips, teethes are visible or not what is the height and width of the lip inner and outer contour.

Gray Level Co-Occurrence Matrix (GLCM) has proved to be a popular statistical method of extracting region feature from images. Following two tables demonstrate different GLCM features for the example sentences of selected 10 frames.

Table 2. Sentence: Thank you

Name of feature	**Lip_Image _a1**	**a2**	**a3**	**a4**	**a5**	**a6**	**a7**	**a8**	**a9**	**a10**
Contrast	0.0287	0.0175	0.0232	0.0173	0.0194	0.0144	0.0213	0.0231	0.0181	0.0142
Energy	0.4868	0.4831	0.4849	0.4853	0.5164	0.4948	0.4805	0.4839	0.4834	0.4884
Homogeneity	0.9857	0.9913	0.9884	0.9913	0.9903	0.9928	0.9894	0.9884	0.9909	0.9929
Mean	0.4121	0.5103	0.4360	0.4635	0.3649	0.4322	0.4720	0.4403	0.4735	0.4628
Standard_deviation	0.4922	0.4999	0.4959	0.4987	0.4814	0.4954	0.4992	0.4964	0.4993	0.4986
Entropy	0.9776	0.9997	0.9881	0.9962	0.9466	0.9867	0.9977	0.9897	0.9980	0.9960
Variance	0.1759	0.2139	0.1847	0.1686	0.1442	0.1604	0.1731	0.1442	0.1516	0.1206
Smoothness	0.9998	0.9999	0.9999	0.9999	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999
IDM	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000

Table 3. Sentence: May I help you

Name of feature	**Lip_Image _m1**	**m2**	**m3**	**m4**	**m5**	**m6**	**m7**	**m8**	**m9**	**m10**
Contrast	0.0196	0.0151	0.0145	0.0167	0.0234	0.0148	0.0163	0.0163	0.0174	0.0177
Energy	0.4838	0.5174	0.5307	0.4959	0.4892	0.4901	0.4929	0.4917	0.5041	0.4905
Homogeneity	0.9902	0.9925	0.9928	0.9917	0.9883	0.9926	0.9918	0.9919	0.9913	0.9911
Mean	0.4588	0.3709	0.3480	0.4191	0.4200	0.4505	0.4313	0.4361	0.3948	0.4353
Standard_Deviation	0.4983	0.4831	0.4764	0.4934	0.4936	0.4976	0.4953	0.4959	0.4888	0.4958
Entropy	0.9951	0.9514	0.9323	0.9810	0.9815	0.9929	0.9863	0.9882	0.9678	0.9879
0.1847	0.1790	0.1499	0.1499	0.1957		0.1830	0.1491	0.1648	0.1524	0.1779
Smoothness	0.9999	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
IDM	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000	255.0000

3.4.1. Contrast

Contrast returns a measure of the intensity contrast between a pixel and its neighbor over the whole image.

$$Contrast = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2 \tag{1}$$

3.4.2. Energy

Energy returns the sum of squared elements in the GLCM.

3.4.3. Homogeneity

Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM.

3.4.4. Mean

It is determined by adding all the data points in a population and then dividing the total by the number of points.

$$Mean = \frac{\sum fx}{\sum f} \tag{2}$$

3.4.5. Standard deviation

Standard deviation is a measure of variance within a data set.

$$SD = \sqrt{\frac{\sum \left| x - \overline{x} \right|^2}{n}} \tag{3}$$

3.4.6. Entropy

Entropy shows the amount of information of the image that is needed for the image compression. Entropy measures the loss of information or message in a transmitted signal and also measures the image information.

$$ENTROPY - \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} -P_{ij} * \log P_{ij} \tag{4}$$

3.4.7. Variance

Variance is a measure of the dispersion of the values around the mean. It is similar to entropy.

3.4.8. Smoothness

Measures the smoothness (homogeneity) of the. gray level distribution of the image.

3.4.9. IDM

Inverse Different Moment (IDM) is the local homogeneity. It is high when local gray level is uniform and inverse GLCM is high.

$$IDM - \frac{\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} P_{ij}}{1+(i-j)^2} \tag{5}$$

4. Conclusion

The Gray Level Co-occurrence Matrix (GLCM) method is used for extracting Statistical region Parameters i.e., contrast, energy, homogeneity, mean, standard deviation, entropy, variance, smoothness and IDM. By extracting the features of an image by GLCM approach, the time of image compression can be substantially decreased in the process of transforming RGB to Gray level image when compared to other DWT method, but however DWT is flexible method of compressing video as a whole. Such capabilities are useful in motion prediction of videos and in real time pattern recognition applications.

References

[1] Wolff G.J., Prasad K.V., Stork D.G., Hennecke M.. Lipreading by neural networks: Visual preprocessing, learning, and sensory integration. In Advances in Neural Information Processing Systems. 1994, 1027-1034.

[2] Rathee N. A novel approach for lip reading based on neural network. In 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). 2016, 421-426.

[3] Lu Y., Li H. Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Applied Sciences. 2019, 9(8), 1599.

[4] Faisal M., Manzoor S. Deep learning for lip reading using audio-visual information for urdu language. Arxiv Preprint Arxiv. 2018, 1802.05521.

[5] Ibrahim M.Z., Mulvaney D.J. Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping. Journal of Visual Communication and Image Representation. 2015, 30, 219-233.

[6] Shivani P., Smita J. Lip reading of digits using artificial intelligence. International Journal of Science Technology Management and Research. 2018, 3,(4), 2456-0006.

[7] Garg A., Noyola J., Bagadia S. Lip reading using CNN and LSTM. Technical report. Stanford University. 2016.

[8] Gutierrez A., Robert Z. Lip reading word classification. 2017.

[9] Noda K., Yamaguchi Y., Nakadai K., Okuno H.G., Ogata T. Lipreading using convolutional neural network. In Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[10] Khetarpal P., Moradian R., Sadar S., Doultani S., Pathan S. Lipvision: a deep learning approach. International Journal of Computer Applications. 2017, 975, 8887.

[11] Le Cornu T., Milner B. Voicing classification of visual speech using convolutional neural networks. In FAAVSP-The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing. 2015.

[12] Mohanaiah P., Sathyanarayana P., GuruKumar L. Image texture feature extraction using GLCM approach. International Journal of Scientific and Research Publications. 2013, 3(5), 2250-3153.