IJIRR

International Journal of Information Research and Review
Vol. 04, Issue, 08, pp.4459-4463, August, 2017

CASE REPORT

VISUAL OBJECT SPEECH RECOGNITIONFOR ISOLATED ARABIC WORDS

***[1]Nour Sami Ghadban, [2]Jafar Alkheir and [1]Mariam Saii**

[1]PhD student, Department of Computer and Control Engineering, Tishreen University, Syria
[2]Asocciate Professor - Department of Computer and Control Engineering, Tishreen University, Syria

ARTICLE INFO

Article History:

Received 14th May, 2017
Received in revised form
22nd June, 2017
Accepted 10th July, 2017
Published online 30th August, 2017

Keywords:

Visual object speech processing,
Visual features, DCT,
Mouth location/tracking.

ABSTRACT

Due to the increasing demands for the symbiosis between human and robots, humanoid robots (HRs) are expected to offer the perceptual capabilities as analogous as human being. One challenge of HRs is its capability of communication with people. Recently, the author presented a novel system for the visual speech recognition. The visual recognition problem is central to computer vision research. This article views first algorithm for Arabic visual object speech recognition. Visual features are width height, marker distances, dct of lips .The Idea is to find the green markers (objects) around the lips to form a masking polygon that will crop out irrelevant areas of image after converts this image to labeled image. and these objects will track them to extract features. Image filtering techniques introduces many artifacts (extra pixels that are not makers). Missing marker positions due to aliasing in video, we will estimate position of missing markers. Algorithm work for all test and training sets, different lighting, framing conditions

Copyright©2017, Nour Sami Ghadban. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.*

INTRODUCTION

Previous work & Research in the field of Arabicvisual speech recognition

The studies presented in the Arabic language are still new in this field. The studies differ according to the characteristics extracted from the oral image. The following is an explanation of the studies presented in the field of visual recognition in Arabic:

The research (Sadeddine, 2013), designed a system to identify Arabic audio tracks to extract visual features. Discrete cosine transform (DCT), a well-known compression tool, was used to reduce the size of the lip matrix, a conversion that converts the image from a spatial area to a Bandwidth, and DCTs are the input system. The study (Fatma Zohra Chelali, 2011) used 2D-DCT coefficients to extract lip image features. After identifying the oral area, the area of the mouth was changed to 120 * 120 and DCT was applied. The 100 most energy-conserved laboratories were used for classification. Depends on the geometry parameters by taking the horizontal and vertical distance of the lips. The search (Seddik, 2013), in this article follows the following steps to take the features of the lips: first the images were converted from RGB space to the YIQ space and then they truncated the Q vehicle and then applied an optimal arrangement based on the histogram of the resulting image to convert the Q vehicle to the binary system.

Corresponding author: Nour Sami Ghadban
Department of Computer and Control Engineering,Tishreen University, Syria

Morphological processes were used to extract characteristic features that describe the movement of the lips. The Harris Corner Detection Method was applied to extract the corners of the mouth and 4 points were identified to describe the movement of the mouth where the horizontal and vertical distance that marks the word is calculated. The study (Elham, 2014) extracted Discrete Cosine Transform (DCT) Coefficients from the oral region. The facial area and the oral region were detected using the Viola-Jones algorithm. The research (Fatma Zohra Chelali, 2012) manually selected a frame around the mouth images taken from a 120 x 160-word video, then the image is converted to gray, and the mean and standard deviation of the pixel values is calculated. The horizontal and vertical distance of the inner and outer lips is taken as characteristic features of the lips. The researcher (AlaaSagheer, 2015) designed a speech recognition system based on sound and oral images, the system detects the face and mouth in real time, the researcher used the Self Organizing Map (SOM) technique to extract features and k-Nearest Neighbor technique in addition to Hidden Markov Model to identify the Speech, the system was applied to a database of Arabic language collected by the author consisting of 36 words and 13 sentences. In order to detect the face and mouth area, the Viola-Jones algorithm was used to extract features. Self Organizing Map (SOM) technology, a technique known in neuronal networks based on competitive learning without supervision, was used. The research presents a new way to automate lip reading and extract its features from videos of isolated words spoken in Arabic. The area of the face and then

the mouth region were detected based on the Viola-Jones algorithm. There is also the possibility of manual identification of the frame around mouth images.

Objects in Image

Object detection is the identification of an object in an image or video. Computer Vision System Toolbox™ supports several approaches to object detection, including template matching, blob analysis, and the Viola-Jones algorithm. Template matching uses a small image, or template, to find matching regions in a larger image. Blob analysis uses segmentation and blob properties to identify objects of interest. The Viola-Jones algorithm uses Haar-like features and a cascade of classifiers to identify pretrained objects, including faces, noses, eyes, and other body parts. You can also train a custom classifier.

The functions Get information about the objects in an image:

Table 1. Functions Get information about the objects in an image

regionprops	Measure properties of image regions
bwarea	Area of objects in binary image
bwconncomp	Find connected components in binary image
bwconvhull	Generate convex hull image from binary image
bwdist	Distance transform of binary image
bwdistgeodesic	Geodesic distance transform of binary image
bweuler	Euler number of binary image
bwperim	Find perimeter of objects in binary image
bwselect	Select objects in binary image
graydist	Gray-weighted distance transform of grayscale image
imcontour	Create contour plot of image data
imhist	Histogram of image data
impixel	Pixel color values
improfile	Pixel-value cross-sections along line segments
corr2	2-D correlation coefficient
mean2	Average or mean of matrix elements
std2	Standard deviation of matrix elements
bwlabel	Label connected components in 2-D binary image
bwlabeln	Label connected components in binary image
labelmatrix	Create label matrix from bwconncomp structure

providing a "good" basis for isolating the important area of the mouth and understanding the spoken word:

- Determine the stages of mask implementation to identify the lips to isolate the mouth area.
- Extract features.

Part.1: lip mask

There are five steps to build a lip mask**:**

- Crop frame to lip region
- Find green markers
- Remove artefacts
- Reconstruct any missing markers

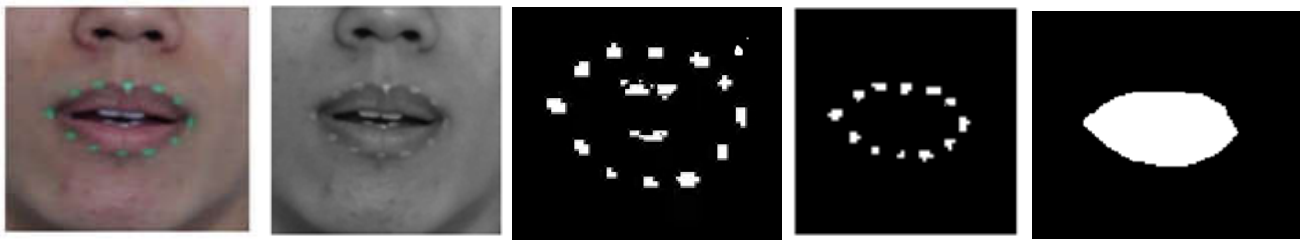

Figure 1. Steps to build a lip mask

In our search we convert image to Label connected components in 2-D binary image, find Area of objects in binary image, and measure properties of image regions.

MATERIALS AND METHODS

The new algorithm of motion tracking techniques used in Hollywood films drew the small green dots on the face,

- Working with markers as objects
- Construct mask polygon

Crop the frame: Need to crop frame to reduce amount of data to process. Imaging processing is slow. Also reduces artifacts in data (erroneous green pixels due to thresholding) that occurs outside of the mouth due to image filtering. Fixed bounding

box so that we can run the DCT on the frame, results would not be useful with varying size DCT.

So we can:

- Rough crop to find mouth markers.
- Find markers
- Find center point
- Make a fixed sized bounding box crop

We have two crops:

- **First frame:** A hard coded rectangle crop because we know where the mouth is in all our videos.
- **Subsequent frames:** a fixed sized bounding box is made at the center point to fit as tightly all the markers.

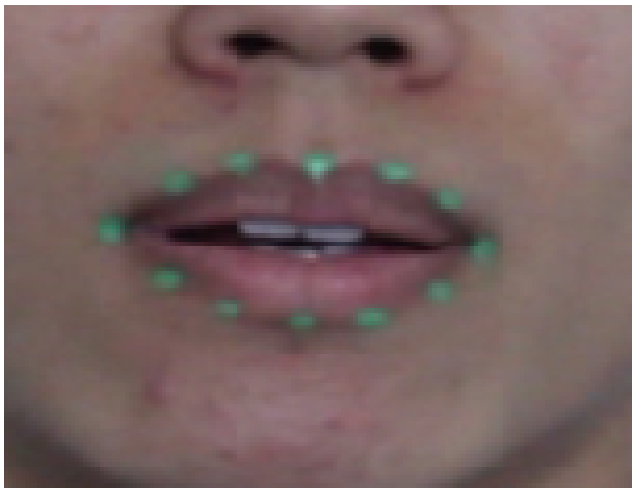

Figure 2. A hard coded rectangle crop

Extract green pixels

Two methods to filter out green pixels

1. Test pixel color ratio
   a. g > r && r > b
2. Green pixel intensity threshold
   a. Value = Diff green channel and grayscale
   b. Value > threshold = 1

Figure 3. Extract green pixels

Remove artefacts

Take logical intersection to remove as manyerroneous pixels as possible.

a. Remove inner mouth artefacts
   Teeth
   - Pass pixels r:g:b ratio near 1
   - Pass pixel intensity > average
     Dark areas inside mouth
   - Pass pixel intensity < average
   - Varying sized circle at center

Reconstruct any missing markers

Remove regions that stray too far from the ring of markers

- Sort regions by euclidean distance from center
- Get difference of adjacent values
- Region is considered stray when
  a. Is near the end of the list (far)
  b. difference > threshold

Need to have 12 full markers at all times

- If markers < 12, compare with previous frame to find which marker is missing.
- Use k-nearest neighbor search by matching their direction vectors
- Take old position and add weighted velocity of two closest neighbours
- More weight if neighbor is closer to missing marker
- More weight if marker at bottom of mouth
- Reestimate the center point at end

Working with markers as objects

Once we've got our 12 markers and they don't disappear we convert them into objects and obtain their properties, mainly the position of the centre of mass for each object.

Functions we use

- bwlabel() - inbuilt
- finds and labels white objects from the binary image
- regionprops() – inbuilt obtains properties of those objects
- in our case we use 'Centroid' - position of the centre of mass of each object
- vislabels() - by Steve Eddins*

Guide to sorting the markers in the clockwise order

- Obtain centre point by taking the mean value of all markers
- Get alpha angle value between centre point and each marker with regards to Y axis
- Sort the markers ascending based on that value
- Profit. Now you can connect the markers to make a mask

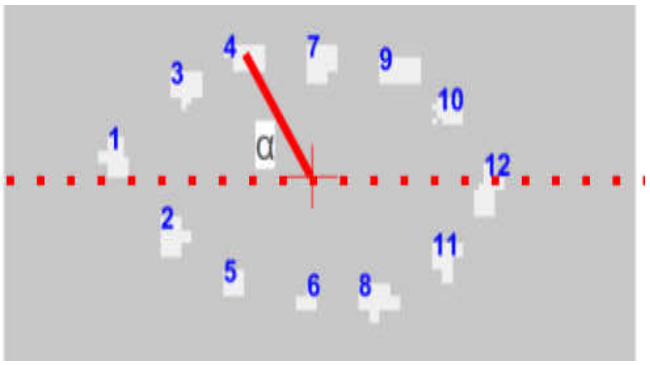

Figure 4. Alpha angle value between centre point and each marker with regards to Y axis

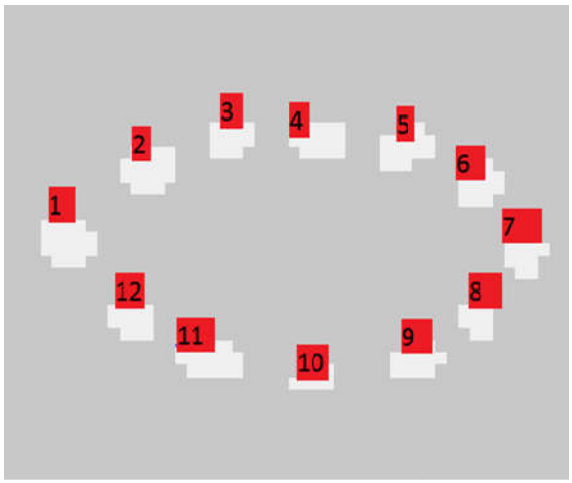

Figure 5. Sorting the markers in the clockwise order

Construct mask polygon

Once we got the markers ordered, we connect them and fill the inside creating a lip mask. When applied to the original frame we get a nice, isolated mouth area.

Figure 6. Lipmask

Figure 7. Lip mask extraction in motion

Part.2 Extract features

Feature 1 - Width and Height of Lips

How to obtain the feature:

- Find the extremas of all the markers centre of mass positions (both in X and Y)
- Subtract the smallest values from the largest values
- X value subtraction is width Y value subtraction is height Final feature vector: [width, height]

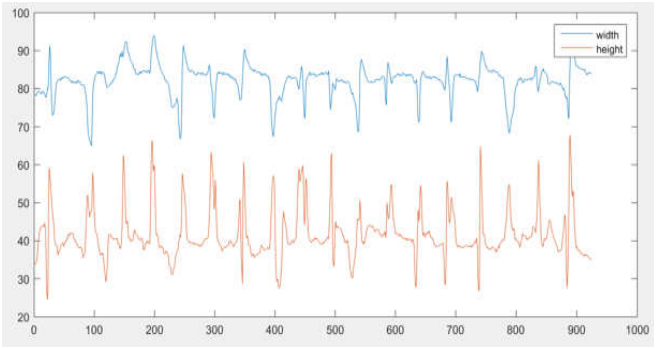

Figure 8. Width and heightfeature vector for one word

Feature 2. Relative Distance of Markers to Centre Point

How to obtain the feature:

- Use the matrix of sorted marker objects to calculate the Euclidean distance between each marker and the centre point

Final feature vector: 12 values (each marker tracked independently)

Feature 3 - Inner Mouth Pixel Count

The third feature vector is the inner mouth pixel count. Original algorithm picks up nostrils and specular reflection on the nose. In order to get rid of that we take a binary intersection between the original teeth/dark area mask and the newly acquired lip mask. What we get in the end is the isolated mask which accurately shows the openness of the mouth.

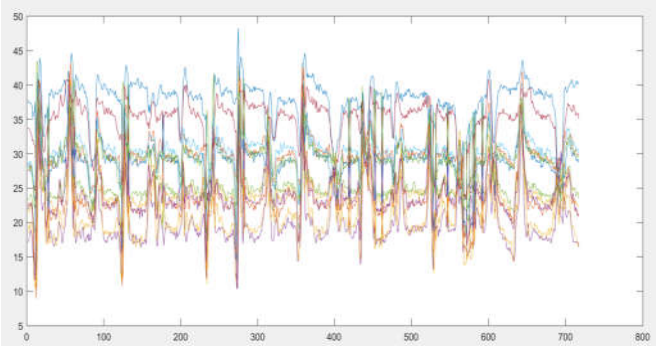

Figure 9. Inner Mouth Pixel Count feature vector for one word

Feature 4 - DCT representation

How to obtain the feature:

- Apply the lip mask to the original frame
- Convert to greyscale
- Apply dct()
- Extract low-frequency information from the result

Final feature vector: 25 values (5x5 matrix) representing the low frequency mouth information.

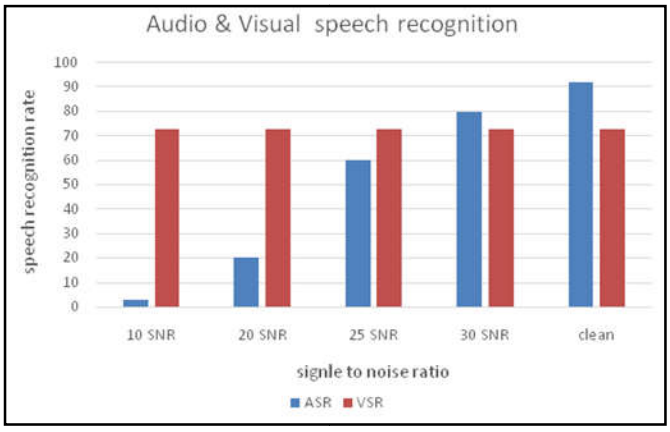

Figure 10. Comparison of speech recognition results using audio features (ASR)to visual features (VSR)in different signal to noiseratios

Summary

The system was tested on 1512 sample video for isolated words spoken in Arabic. The results showed that the new algorithm has a known rate of up to 73% and proves that it is complementary to the audio signal and will provide clear results in the development of audiovisual systems. Visual speech recognition systems are not dependent on their sound environment. They are not affected by audio noise, which reduces the recognition rate in audio systems. In comparison with a system that relies on extracting audio features only and for the same database and subjected to noise at different rates, 21% when the signal to noise ratio was 10 dB and the increase was 70% when the signal to noise ratio was 0 dB.

REFERENCES

AlaaSagheer, Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications, *Appl. Math. Inf. Sci.* 9, No. 6, 2885-2897 (2015).

Cox, S. I. Matthews, and J. A. Bangham, \Combining noise compensationwith visual Information in speech recognition," in Auditory-Visual Speech Processing, (Rhodes), 1997.

Dodd, B. and R. Campbell, eds., Hearing by Eye: The Psychology ofLipreading. London, England: Lawrence Erlbaum Associates Ltd., 1987.

Dodd, B. and R. Campbell, eds., Hearing by Eye: The Psychology ofLipreading. London, England: Lawrence Erlbaum Associates Ltd., 1987.

Dodd, B. and R. Campbell, eds., Hearing by Eye: The Psychology ofLipreading. London, England: Lawrence Erlbaum Associates Ltd., 1987.

Elham, S. SalamaReda A. El-Khoribi Mahmoud E. Shoman, Audio-Visual Speech Recognition for People with Speech Disorders,*International Journal of Computer Applications* (0975 – 8887) Volume 96– No.2, June 2014

Fatma Zohra Chelali, KhadidjaSadeddine Amar Djeradi, Visual Speech Analysis, Application to Arabic Phonemes Special Issue of *International Journal of Computer Applications* (0975 – 8887) on Software Engineering, Databases and Expert Systems – SEDEXS, September 2012.

Fatma Zohra Chelali,Audiovisual speech/speaker recognition,application to Arabic languageMultimedia Computing and Systems (ICMCS), 2011

Frowein, H. W., G. F. Smoorenburg, L. Pyters, and D. Schinkel, \Improved speech recognition through videotelephony: Experiments with the hard of hearing," *IEEE Journal of Selected Areas in Communications,* vol. 9,pp. 611{616, May 1991.

Hennecke, M. 1996. eds.), pp. 103{114, Springer-Verlag, 1996.

http://uk.mathworks.com/matlabcentral/fileexchange/19665-visualize-output-of-bwlabel/content/vislabels.m

Lavagetto, F. Converting speech into lip movements: A multimedia tele- phone for hard hearing people," *IEEE Transactions on Rehabilitation Engineering,* vol. 3, pp. 90{102, March 1995.

Lippmann, R. P. 1997. “Speech recognition by machines and humans,” Speech Commun, vol. 22, pp. 1-15.

Luettin, J., G. Potamianos, and C. Neti, \Asynchronous stream modelingfor large vocabulary audio-visual speech recognition," in International Con-ference on Acoustics, Speech and Signal Processing, vol. 1, (Salt Lake City),pp. 169 {172, May 2001.

McGurk, H. and J. MacDonald, \Hearing lips and seeing voices," Nature,pp. 746{748, December 1976.

McGurk, H. and J. MacDonald, \Hearing lips and seeing voices," Nature,pp. 746{748, December 1976.

Sadeddine, K FZ Chelali, R Djeradi, A Djeradi,Visual Speaker Verification System Depending on Arabic Syllables,Speech communication and signal processing laboratory HouariBoumedienne University of sciences and Technologies, USTHB 2013

Seddik, A.F. M El Adawy A Computer-Aided Speech Disorders Correction System for Arabic Language , - ieeexplore.ieee.org ,2013

Tren, D. M. W.Lewis, “Lip Region Detection,” 2001.

Yao-Jiunn Chen, Y.-C. L. 2007. “Simple Face-detection Algorithm Based on Minimum Facial Features,” in IEEE Industrial Electronics Society (IECON), Taipei, Taiwan.

****