Visual speech detection using OpenCV

[1]Muhammad Usman Ghani Khan, [2]Sajid Mahmood, [3]Mahmood Ahmed, [1]Yoshihiko Gotoh,
[1]The University of Sheffield, [2]Al-KICS UET Lahore , [3]University of Engineering & Technology Lahore
usmanghanikhan@gmail.com, sajid.mahmood@kics.edu.pk, Mahmood@uet.edu.pk, y.gotoh@dcs.shef.ac.uk

Abstract

Visual information from the human face; lip-movements and tongue provide us with lots of information about the spoken message and helps in understanding the verbal communication. The visual speech detection overcomes some of the persistent problems and inaccuracies encountered by users that creep in when there is background noise. In noisy environment we pay more attention to the lips which dramatically improves our understanding of what other people are saying. This research is focussed towards creation of a speech detector which works solely on video data. This work is part of speaker identification problem in videos. We propose speaker identification using visual clues only. Based on visual information, presence of speech can be extracted in video sequences.

Keywords: Speech detection, Classification, OpenCV, Video processing

1. Introduction

One of the most important means of communication is our voice, but sometimes when we are in a noisy environment or at some distance away from a person to which to communicate, then we pay more attention to the lips so as to understand what he is speaking. Moreover, many sounds that are confusing by ear are easily distinguishable by eye, such as N and M. In all such situations vision based speech detection can be a very useful mean for communication.

There have been significant efforts to improve the performance of the speech recognizers by using different techniques, one of which is to mimic human behavior by extracting the visual cues from video data (movements of lips and tongue, facial expressions etc.). Audio-video speech-reading systems use the information from the mouth area to enhance the understanding of speech. The visual channel is not dependent on the audio sounds, therefore is unaffected by presence of any background noise or cross-talk among the speakers. [1]

After about 30 years of research in the field of speech detectors and recognizers, we could identify applications ranging from speaker dependent, isolated word recognizers, to speaker independent, large vocabulary and continuous speech recognizers. [2]

Automatic speech recognition (ASR) systems have become quite widely used in HCI (Human-Computer Interaction), they are still significantly limited, depending on the environment in which they are being used. [3] However we can not predict their behaviour in noisy conditions; some of them will cope with a moderate level of noise, others will fail completely when they encounter minor acoustic interference. [4]

One possible approach is to introduce another data source which would complement the acoustic information and would be invariant to the most common sources of corruption - other people speaking, reverberation effects, transmission channel distortions created by the hardware capturing the audio signal. [5] One option is to use visual information since it has been proven that combining audio and visual data improve the results considerably. It is known that visual information from the face, tongue, and lip-movements of a talker provides information about the spoken message and enhances the intelligibility of speech. [2].

Automatic speech-reading systems use the information coming from the extraction of important features in the mouth area.

There are two main ways of extracting the features:

- Use the data of all the pixels from the mouth area, as the input for a recognition system (Hidden Markov Models, neural networks or different ML techniques). Next, the selected system learns the typical values of pixels for particular lip movements in order to distinguish between them. [7]
- Use image processing techniques in order to find certain feature points in the image (lip corners etc.). This approach can involve simple techniques such as: thresholding, integral projection, or sophisticated methods, for example: active contour models, active shape models or 3D lip models. [2] [4]

Our contribution is that we developed a full fledge working system using OpenCV framework which is an open source framework for vision based applications. Several methods and ideas were explored and implemented to produce a system for vision based speech detection. This work is suitable for scenarios where there is no speech or noisy environments.

2. Proposed Methodology

Figure 1 demonstrates steps for the detection of speech based on visual features only. For face detection haar-cascade classifier was applied. Once face is detected, head motion tracking is captured using mean shift and cam-shift algorithms. Lower face parts such as lips, tongue, chin and jaw were detected using geometry structure of the face. Idea is to capture the movement of these facial parts. Based on the movements of these lower parts different classification algorithms like SVM, MLP, random forest trees were applied which classified a given video as either speech video or speechless video.

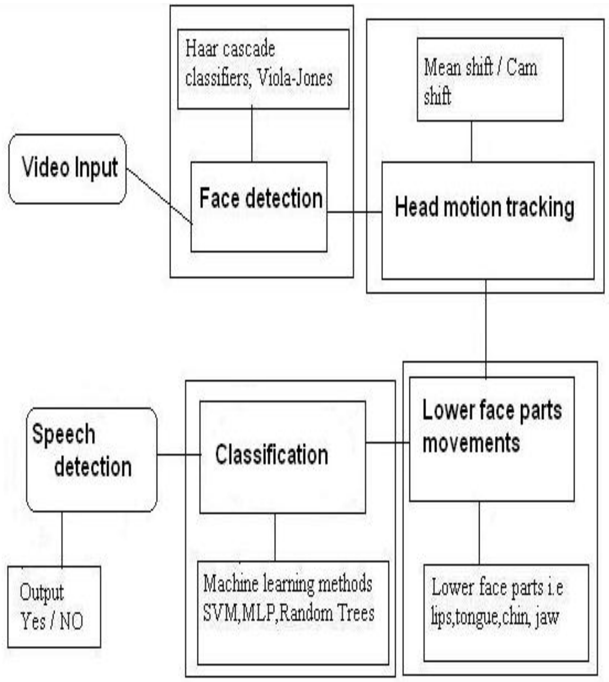

Haar cascade classifiers, Viola-Jones
Mean shift / Cam shift
Video Input
Face detection
Head motion tracking
Speech detection
Classification
Lower face parts movements
Output Yes / NO
Machine learning methods SVM,MLP,Random Trees
Lower face parts i.e lips,tongue,chin, jaw

Figure 1: Proposed methodology for detection of speech based on visual features

2.1 Face detection

We could describe the problem as follows: Find and identify one or more persons in the scene from given still or video images. Many algorithms have been proposed to deal with the problem, in some of them, both the training and the test set consist only of facial images. Just to mention a few: Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA) and Elastic Graphic Matching. The Hidden Markov Model (HMM) which has been used for speech, gestures and expressions recognition can also be applied to image-based face detection (precisely, the spatially embedded HMM is used). Video-based face recognition can also be performed with adaptive HMM. [7] [2]

One of the best performing face detection frameworks, which is operating extremely rapidly and achieves high detection rate, is the method proposed by Viola and Jones. [13]

The technique introduces new image representation called the “Integral image”, which allows for quick computation of features. The system can compute the integral image by using only a few operations per pixel. With the integral image Haar-like features can be computed at any scale in constant time. [5]

In OpenCV the actual Viola-Jones detector has been extended to use diagonal features [8]. It is often referred to as the Haar-classifier, because it uses Haar-like wavelets (consist of adding and subtracting rectangular image regions prior to thresholding the result).Our face detection is also based on the method of haar-classifiers. This face detection method was implemented and tested on several images resulting in satisfactory results.

2.2 Head Motion Tracking:

It is not always necessary to detect the position of a face in every frame. If a face does not appear in front of the camera, the chance that it will be there in the consecutive frame is relatively low. If, however the system finds a frontal face, taking in account the movement of the head, we can predict the location of the face. During the process of speech it is mostly the shape of the mouth and the lower part of the face that changes (as well as the orientation of the head, yet the upper part of the face stays stable). That assumption guarantees insignificant loss in performance and lots of computational savings. [10]

The mean-shift (kernel-base) algorithm [6] uses local extrema of the density distribution of a set of data. The method calculates a histogram for the tracked object, and then scans the back-projection of each frame to find the most possible location of the object. The algorithm ignores outliers in the data, which means, it ignores points that are far away from peaks. The algorithm can be approximated in a few steps:

1. Search window selection:
    - initial location
    - type (uniform, polynomial, exponential, or Gaussian)
    - shape (symmetric, skewed, rotated, rounded or rectangular)
    - size (possible cut off)
2. Find the centre of mass for the window (let’s call the point C).
3. Use the C point to centre the window.
4. Go back to step two as long as the window is moving (iterations usually restricted to some maximum number). [6] [9]

The Camshift (“continuously adaptive mean-shift”) tracker uses the same principles as the mean-shift; however the window size does not stay the same, as it adjusts itself. If your object is moving closer to the camera, the algorithm should adjust the size of the window, and track the object accurately. [6]

2.3 Lower-face movement analysis

Four points are usually used to perform tracking of the lips. One at mid-point of each lip, and one at each mouth corner. Even though finding and tracking the upper lip is reasonably straightforward, the lower lip causes a lot of problems. There have been several solutions proposed, to solve the problem (based on colour information etc.) however none of them deliver sufficient accuracy or robustness.

We used some of the principles as described in [10], which can be summarised as follows. The position of the area is based on general head proportions and once the position is established, the created algorithm extracts the features from the lips area. The signal from a video file/camera is transferred into HSV colour space, which separates hue, saturation and value (Figure 2). The HSV colour space in comparison to the RGB is not intensity-sensitive. This makes the algorithm more robust to illumination changes, and makes it possible to analyse the changes in lighting. The saturation values are used; since using the hue value to separate the lips and the skin flesh from the oral cavity does not work very well.

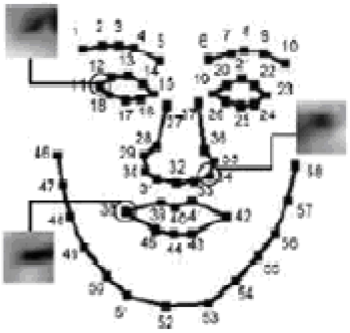


Figure 2: Topological Model of face [7]

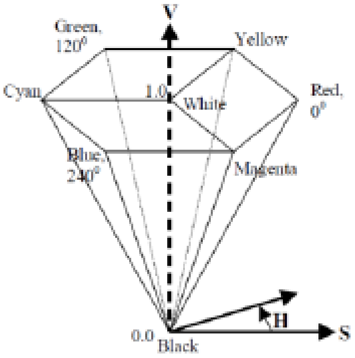


Figure 3: HSV colour space [8]

The oral cavity has the highest saturation (because it is the darkest), the values for the teeth are relatively low, and the skin values are positioned somewhere in between. Using that information, we can assume that when the mouth is open, the average luminance of the mouth area is lower than when the mouth is closed. Based on that assumption, speech detection can rely solely on the illumination change. Obviously it is easy to mislead the system, by simply moving one’s jaw or simulating speech activity. [10] [4]

2.4 Classification:

Once face detection and facial parts have been separated and we have information related to head motions and lower face parts, next step is to classify a given image in speech and speechless images. We tried several of the classification algorithms and results are presented in this section with brief review of every classifier.

2.4.1 Linear Discriminant Analysis [12]

LDA is a technique used in statistics and machine learning for classifying objects into one of two or more classes, based on a set of features. The method finds the linear combination of the features which best separates the given classes. In principle we assign an object to one of the predefined groups based on the observations on that object. This technique is sometimes applied to classification problem in speech recognition (which might be called a much more complex version of our problem), and therefore has been considered as one of the possible prediction algorithms from the start of the project.

The resulting combination of the features can be used as a linear classifier, or for dimensionality reduction. For the purpose of the system, LDA will be used as a classification method.

2.4.2 Support Vector Machine [12]

SVM are often used in many computer vision algorithms such as: *finding feature points, tracking and segmenting scenes.* The algorithms work by projecting the data into a higher-dimensional space (new dimensions are created by combining the features). It will then try to find the optimal linear separator between the classes. When finally the high-dimensional linear classifier is projected to the original space, it usually becomes quite nonlinear. This allows us to use linear classification techniques based on maximal between-class separation to produce nonlinear classifiers, and with enough dimensions we can almost always perfectly separate our data.

2.4.3 Random Trees [6]

OpenCV contains an implementation of Leo Breiman's theory of random forests. By collecting the class 'votes' at the leaves of each tree and selecting the class with maximum votes, the algorithm can learn more than one class at a time. This allow for using the technique to deal with visual speech detection.

Our work averages together many randomly arranged decision trees, and tries to overcome the fact that averaging together similar trees will not improve our results. Random trees can also be used for determining outliers.

2.4.4 Boosting [6]

Boosting is a machine learning meta-algorithm for performing supervised learning. The technique is based on the question: can a set of weak learners create a single strong learner? We define a weak learner as a classifier which only slightly correlates with the true classification.

OpenCV library provides four types of boosting (all are variations of the original algorithm).

- CvBoost :: DISCRETE (discrete AdaBoost)
- CvBoost :: REAL (real AdaBoost)
- CvBoost :: LOGIT (LogitBoost)
- CvBoost :: GENTLE (gentle AdaBoost)

Experiments reveal that the real and gentle algorithms work best in most situations. Real form of AdaBoost utilizes confidence-rated prediction, and yields high efficiency with categorical data. Gentle algorithm puts less weight on outliers and because of that gives good results with regression data.

2.4.5 Multilayer Perceptron [11]

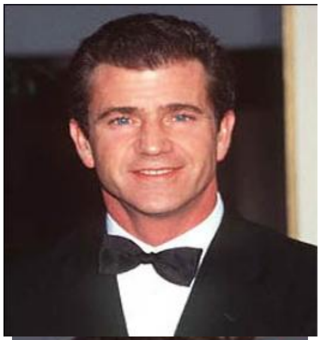
By using a network of artificial nodes organized in layers, MLP simulates the brain activity.
Backward propagation is the most often used training algorithm. The technique tries to minimize the prediction error by looping through a learning cycle (recalculating neuron weights). After each iteration, we adjust the weights to better match the desired output and assign error to all neurons.

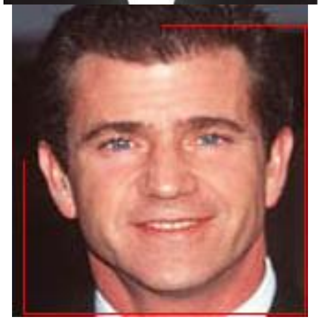
3. Results and discussion

The test set is composed of 5 completely unrelated video files.

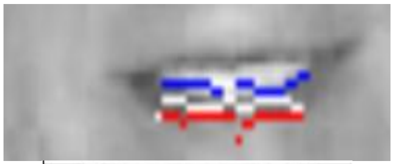
- Small part of a news program- few faces appearing simultaneously.
- Long recording from an online video corpus – only one face appearing, long periods of silence, rapid head movement
- Short clip from a movie trailer- high resolution video showing different faces, lots of action/movement
- Part of a television interview- faces of the interviewees
- Short TV report – only one face, small and obstructed face

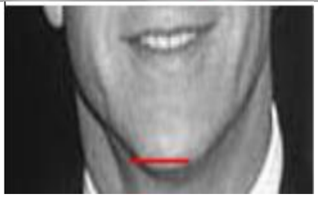
Following frames show the detection of speech in a video taken from a video. First video is converted to a frame. From frame human face is detected. Then facial parts like chin, upper and lower lips and complete jaw positions are identified. Threshold is applied for finding the lip area and mouth openness calculation. These steps are repeated for multiple frames. Based on these calculations feature vectors are calculated. These feature vectors are input to classification module where several classification algorithms are implemented and produce the final output which is presence or absence of speech in the given video sequence.



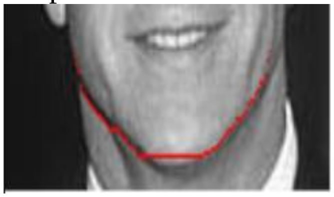


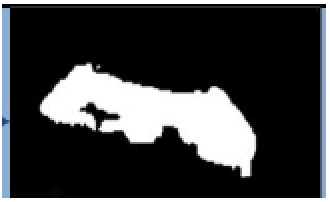
Frame showing person from video Result of face detection



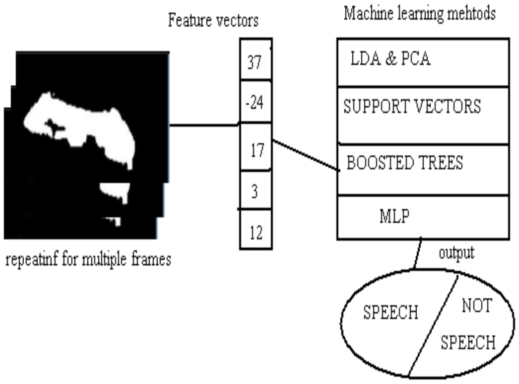


Upper-lower lips detection Chin detection

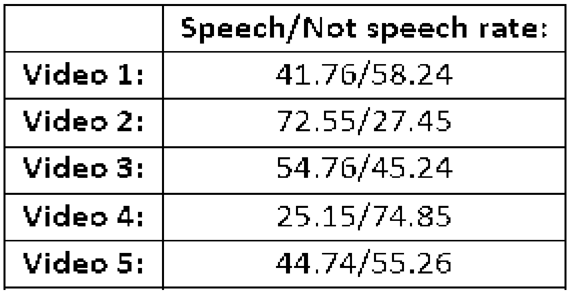




Complete jaw Thresholding lips area and Mouth openness calculation



Here we give the properties of the video files used for the experiments. For every video duration of speech and non speech is provided in seconds. There is not any specific ratio of speech to non speech length. These videos were chosen randomly to find the evaluate the proposed methodology.



	Speech/Not speech rate:
Video 1:	41.76/58.24
Video 2:	72.55/27.45
Video 3:	54.76/45.24
Video 4:	25.15/74.85
Video 5:	44.74/55.26

Table 1: Input videos properties

On the above mentioned videos, several classification algorithms were applied. Details are provided in table 2. SVM proved to be the best classification algorithm as presented by the table.

	SVM	LDA	RANDOM TREES	BOOSTED TREES	MLP	AVERAGE
Video 1	82.8	87.2	71.1	71.4	67	75.9
Video 2	78.4	39.2	79.6	77.1	67	56.26
Video 3	45.2	56.3	53.2	48.4	32.5	47.12
Video 4	74.9	78.1	63.8	62.9	57.2	67.38
Video 5	97.4	68.4	23.7	47.4	84.2	64.22

Table 2: Performance of classifiers when applied on the given video sets

4. Conclusion:

We used face detection, face parts detection and analysis and colour thresholding to achieve the ultimate goal. In the final system, faces are being found by using the Haar classifier, then the mouth position is calculated, and the mouth area is analyzed. Information coming from that region (the level of mouth openness) is passed to several machine learning algorithms which make decisions based on the previously processed data. The system could be easily extended for experimenting with other visual tasks like lip reading, human tracking and video surveillance. The software detects speech with 60-75% accuracy (depending on the specifications of the video material).

References

[1] R. Kaucic, A. Blake. Accurate, Real-time, Unadorned Lip Tracking. Oxford : Dept. of Engineering Science, University of Oxford, 1998.

[2] Tian Gan, Wolfgang Menzel, Shiqiang Yang. An Audio-Visual Speech Recognition Framework Based on Articulatory Features. 2007.

[3] J Bruce Millar et al. Stereo Vision Lip-Tracking For Audio-Video Speech Processing. Canberra : Computer Sciences Laboratory and Robotic Systems Laboratory, 2001.

[4] Roland Gocke, et al. Automatic Extraction of Lip Feature Points. Canberra : Computer Sciences Laboratory and Robotic Systems Laboratory,Australian National University, 2000.

[5] Rowan Seymour, Darryl Stewart and Ji Ming. Comparison of Image Transform-Based Features for Visual Speech Recognition in Clean and Corrupted Videos. Belfast : School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, 2008.

[6] Gary Bradski, Adrian Kaehler. Learning OpenCV Computer Vision with the OpenCV Library. Sebastopol : O'Reilly Media, Inc, 2008. Used with permission of O'Reilly Media, Inc. All rights reserved..

[7] Xiaoming Liu, Tsuhan Chen. Video-Based Face Recognition Using Adaptive Hidden Markov Models. Pittsburgh : Electrical and Computer Engineering, Carnegie Mellon University, 2003.

[8] Rainer Stiefelhagen, Uwe Meier, Jie Yang. Real-time lip-tracking for lip-reading. Karlsruhe : Interactive System Laboratories, 1997.

[9] Statistical Pattern Recognition Toolbox. Examples. [Online] http://cmp.felk.cvut.cz/cmp/software/stprtool/examples.html.

[10] Giridharan Iyengar, Chalapathy Net. A Vision-based Microphone Switch for Speech Intent Detection. Yorktown Heights : Human Language Technologies IBM T. J. Watson Research Center, 2001.

[11] http://en.wikipedia.org/wiki/Multilayer_perceptron

[12] Machine learning. Wikipedia. http://en.wikipedia.org/wiki/Machine_learning.