Color-based Lip Localization Method

Ahmad B. A. Hassanat, Sabah Jassim
Applied Computing
The University of Buckingham, UK
{ahmad.hassanat, sabah.jassim}@buckingham.ac.uk

ABSTRACT

This paper is concerned with lip localization for visual speech recognition (VSR) system. We shall present an efficient method for localization human’s lips/mouth in video images. This method is based on using the YCbCr approach to find at least any part of the lip as an initial step. Then we use all the available information about the segmented lip-pixels such as r, g, b, warped hue, etc. to segment the rest of the lip. The mean is calculated for each value, then for each pixel in ROI, Euclidian distance from the mean vector is calculated. Pixels with smaller distances are further clustered as lip pixels. Thus, the rest of the pixels in ROI will be clustered (to lip/non-lip pixel) depending on their distances from the mean vector of the initial segmented lip region.

The method is evaluated on a new-recorded database of 780,000 frames; the experiments show that the method localizes the lips efficiently, with high level of accuracy (91.15%) that outperforms existing lip detection approaches.

Keywords: lip detection, lip localization, object detection, Colour segmentation, Image segmentation

1. INTRODUCTION

Over the last few decades, the number of applications that are concerned with the automatic processing/analysis of human faces has grown remarkably. Many of these applications have a particular interest in the lips and mouth area. For such applications a robust and real-time lips detection/localization method is a major factor contributing to their reliability and success. Since lips are the most deformable part of the face, detecting them is a nontrivial problem, adding to the long list of factors that adversely affect the performance of image processing/analysis schemes, such as variations in lighting conditions, pose, head rotation, facial expressions and scaling.

The lips and mouth region are the visual parts of the human speech production system; these parts hold the most visual speech information, therefore, it is imperative for any VSR system to detect/localize such regions to capture the related visual information, i.e. we cannot read lips without seeing them first. Therefore, Lip localization is an essential process for any visual speech recognition (VSR) system. In this paper, we assume that all the face images are frontal and upright. All the faces have been detected to facilitate lip localization using our method, which is described in [1]. Hence, the face is more easily detected than the lips, and we know that lips are located in the lower centre part of the face.

The major aim of this paper is to investigate lip detection/localization problems and propose a robust method to address the needs of VSR systems. Lip detection methods that have been proposed in the literature so far have their own strengths and weaknesses. The most accurate methods often take more processing time, but some methods proposed for use in online applications trade accuracy for speed. The rest of this paper is devoted to reviewing existing trends in lip localization, and presenting our proposed lip localization method. Using experiments on an in-house video database, we shall demonstrate that our proposed method outperforms existing techniques.

2. EXISTING TRENDS IN LIP LOCALIZATION

Many techniques for lips detection/localization in digital images have been reported in the literature, and can be categorized into two main types of solutions:

1. **Model-based** lips detection methods. Such models include “Snakes”, Active Shape Models (ASM), Active Appearance Models (AAM), and deformable templates.

2. **Image-based** lips detection methods. These include the use of spatial information, pixel colour and intensity, lines, corners, edges, and motion.

We shall now review some of the most common lip detection schemes in these different categories.
2.1 Model-based lip detection methods

This approach depends on building lip model(s), with or without using training face images and subsequently using the defined model to search for the lips in any freshly input image. The best fit to the model, with respect to some prescribed criteria, is declared to be the location of the detected lips.

2.1.1 Snakes

Snakes or Active Contour Models were first introduced by Kass, Witkin, and Terzopoulos, [2]. Snakes have been proposed as a general method for shape detection. Their optimisation techniques are based on splines. The idea is to minimize the energies of the spline iteratively to fit local minima. Figure 1 illustrates the lip localization process using the snakes approach.

The snakes method has been applied efficiently for lips and mouth region detection by various researchers including [3], [4] and [5], who reported good lip localization representative results. However, a number of problems can be associated with the use of the snakes technique for lips and mouth region detection. e.g. they can fit to the wrong feature, such as the nose or the chin; especially if the initial position was far from the lip edges. Snakes like splines do not easily bend sharply so it is difficult for them to locate sharp curves like the corners of the mouth. They can be fooled easily by facial hair. Sometimes (depending on the detected object) the tuning of snake parameters is very difficult to achieve, and takes a long time (several seconds!) (see Figure 1).

2.1.2 Active Shape Models

ASM can be used successfully for facial features detection, particularly the lips and mouth region. ASM was originally proposed by Cootes and Taylor in 1992, [6]. ASM are statistical models of the shapes of objects, which iteratively adjust to fit to the detected object in digital images. A shape is represented by a set of n labelled “landmarks”.

Using a training set of a marked object in images (in this study the object is the lips), ASMs build a statistical shape model of the global shape variation from the training set. This model is used to fit a model to new occurrences of the trained object. A principal component analysis (PCA) is used to construct such a model, [7]. A recent study [8], proposes an extended version of ASM for facial features detection; some of the extensions include fitting more landmarks, and the use of two dimensional landmarks templates instead of one dimension. Their representative results illustrate better performance by adding more landmarks.

Caplier, [9], proposes an algorithm, which can extract lip shape over an image sequence. The algorithm works under natural lighting conditions, and uses an ASM to describe the mouth. The mouth model is iteratively deformed under constraints in the light of image energies. Kalman filters were used on the independent mouth parameters, to give an initial shape close to the final one, which speeds up the convergence time. Another work [10] propose an enhanced version of ASM by employing colour information, which is used to localize the centres of the mouth and the eyes to assist the initialisation step. Their good representative results emphasized the power of colour as a rich source of information. Some other work that have been done on ASM includes [11], [12], and [13].

2.1.3 Active Appearance Models

Active Appearance Models (AAM) are the same as ASM, but instead of using the edge profile along a landmark, AAM are extended to include the grey scale information of the whole image along with the shape information, [14]. Matthews et al., [15], illustrated good lip detection results using AAM after 15 iterations. Unlike the Snakes method, AAM
converges using a small number of iterations because these models have some prior information about the shape, but they are not necessarily fast, because there are more parameters to be handled. Another more recent work involving lip detection using AAM is [16], which argue that the AAM approach was more consistent with the detection of the non-speech sections containing complex lip movements.

2.1.4 Deformable templates

As introduced in [17], a deformable template is a parameterised mathematical model used to track the movements of a given object. It is active in the sense that it can adapt itself to fit the given object. The deformable lip template adjusts its shape according to the value of a number of integrals along the relevant contours. In a manner similar to that of snakes, an energy function is defined to link edges, peaks, and valleys in the input image to the corresponding parameters in the template. The best fit of the elastic model is found by minimizing an energy function of the parameters, [18].

It is a useful shape model because of its flexibility, and its ability to both impose geometrical constraints on the shape and to integrate local image evidences. The work in [19] employs a deformable template for lip tracking, making use of several configurational and temporal penalty terms, which keep erroneous template deviations under control. Other works that have dealt with lip detection and/or tracking based on a deformable template include [20]. In this contribution, the mouth features are represented by the mouth corners and the lip contour parameters, which describe the opening of the mouth. This model facilitates the development of an algorithm to determine whether the mouth is open or closed.

There are several drawbacks when using deformable templates, such as: high computational time complexity, the unexpected shrinking of the templates and rotation of the template, [21].

2.2 Image-based lips localization methods

Since there is a difference between the colour of lips and the colour of the face region around the lips, detecting lips using colour information attracted researchers’ interest recently. Simplicity, not time consuming, and the use of fewer resources e.g. low memory; allowing many promising methods for lip detection using colour information to emerge. The most important information that researchers focus on include the red and the green colours in the RGB colour system, the hue of the HSV colour system, and the component of the red and blue in the YCbCr colour system. Some researchers used more information from the lip edges and lip motion. A well known mixed approach is called the “hybrid edge”, [5].

2.2.1 RGB approach

In RGB space, skin and lip pixels have different components. The red is dominant for both, the green is more than the blue in the skin colour, and skin appears more yellow than lips, [5]. Gomez et al., [22], propose a method for lip detection in which the image is transformed by a linear combination of red, green and blue components of the RGB colour space. Then they apply a high pass filter to highlight the details of the lip in the transformed image, after which both of the generated images are converted to obtain a binary image. The largest area in the binary image is recognized as the lip area.

The Zhang et al., [23] lip detection method is based on red colour exclusion and Fisher transform. Their approach works as follows: firstly, the face region is located using a skin-colour model, and motion correlation, then the lip region of interest (ROI) is defined as the area in the lowest part of the face. Secondly, the R-component in RGB colour space is excluded, then G-component and B-component are used as the Fisher transform vector to enhance the lip image. Finally, an adaptive threshold to separate the lip colour and the skin colour is set according to the normal distribution of the grey value histogram.

2.2.2 HSV approach

The RGB colour system is most commonly used in practice, while the HSV is closer to how humans describe and interpret colour. Hue (H) represents the dominant colour as perceived by someone, apparently, hue is a key feature for detecting lips using the HSV colour system, and since the hue value of the lip pixels is smaller than that of the face pixels. Coianiz et al., [24], use the filtering hue method to detect the lips in coloured images. Researchers normally detect the face then use it to detect the lips, while Jun and Hua, [25], proposed a reverse technique, by which they detect the face using the lips. Their method determines skin area based on a threshold value in an HSV colour model, and then it searches the rough lip area based on the skin chrominance adaptively. After that it employs an R/G filter to find the lip, depending on the geometrical relation between face and lip; the face is finally located.
2.2.3 Hybrid edge

The “hybrid edge” is the combination of the colour and luminance information. It was first produced by in [5]. This work uses a pseudo hue definition, which demonstrates the difference between the lips and the skin colour. The use of the pseudo hue with gradient vector of the intensity of the upper, middle and lower sections of the lip, allows a reliable estimation of several key points positions. Then 5 cubic polynomial curves are fitted using these points to represent the outer lips boundary. A more recent study [26] uses the hybrid edge as a part of a quasi-automatic system to extract and analyse lip-motion features for speaker identification.

2.2.4 YCbCr approach

Nasiri et al., [27], propose a novel lip detection method using Particle Swarm Optimisation to obtain an optimised map based on the fact that lip region has high Cr and low Cb values. After mapping the image to the YCbCr colour space, two sets for lip and face pixels are created to be clustered by a clustering algorithm. Jaroslav, [28], exploits the fact that lips are more red than other parts of the face and uses a specific equation to maximize the Cr and the Cb components and the output image is thresholded to locate the mouth area.

3 THE PROPOSED LIPS DETECTION APPROACH

Among the model-based lip-detection methods, the active shape models are the most common and best performing technique for lip detection. However, it is affected by factors such as facial hair and does not meet some of the functional requirements (e.g. in terms of speed). “The implementation of Active Shape Model ASM was always difficult to run in Real-Time”, [29]. The same thing applies to the AAM approach. In fact, AAM is slower than ASM.

In grey level images and under diffuse lighting conditions, the external border of the lip is not sharp enough (see Figure 2a), and this makes the use of techniques based on the information provided by these images alone ineffective, [24]. Moreover, the ASM and the AAM algorithms are sensitive to the initialisation process. When initialisation is far from the target object, they can converge to local minima (see Figure 2a). Other problems, like the appearance of the moustache, beard, and unexpected accessories like mouth rings, contribute to the problems and challenges that model-based lip detection methods undergo. Figure 2b illustrates these problems for the ASM scheme.

Since the visual speech recognition (VSR) problem needs several pre-processing steps, e.g. face and lips detection, it is vital for the VSR system to have faster solutions for these steps in order that the final solution can work in real time.

In order to overcome the above-mentioned difficulties, we shall incorporate colour information into our proposed scheme, in spite of being vulnerable to variations in light conditions. A lip localization method using colour information is proposed in this study, called “the nearest colour” method. This method is based on using the YCbCr approach to find at least any part of the lip, then using the pixel values of this area to train the system, thus, the rest of the pixels will be clustered depending on the training data. The ROI is searched to find the nearest pixels to the initially detected pixels to emphasize and expand the area to include the whole detected lip area.

* “MSAM” is a state-of-art ASM free library, developed by Milborrow and Nicolls, [8], and can be downloaded from the following link: http://www.milbo.users.sonic.net/stasm/download.html
The question is, what method to use for the initial detection? For the RGB, luminance and chromatic information are mixed, and the RGB values of both the lip and the face are overlapped, i.e. it is difficult to extract the lip from the face using the RGB values of their pixels, see Figure 3.

![Figure 3](image3.png)

**Figure 3.** (left) RGB histogram for lip pixels, (middle) RGB histogram for face pixels, (right) intensity histogram for both lips and face pixels.

For the Hue in the HSV, even after the warping process, the lip hue values are very close to the face hue values, see Figure 4.

![Figure 4](image4.png)

**Figure 4.** Warped Hue histogram for both lips and face pixels.

In YCbCr, chromatic and luminance components can be separated, and lip pixels have high Cr and low Cb, [27]. To foster this property the following equation is used:

\[
Lip - Map = Cr^2 \left( Cr^2 - \eta \frac{Cr}{Cb} \right)^2
\]

\[
\eta = 0.95 \frac{\sum Cr^2}{\sum \frac{Cr}{Cb}}
\]

where \(Cr^2\) and \(Cr/Cb\) all are normalized to the range \([0,1]\).

The “Lip-Map” formula (Equation 1) is designed to increase the values of the pixels where Cr is high and Cb is low (the case of the lips pixels); now the lip-map values of the ROI can be sent to the clustering algorithm, the higher values are lip-pixels, and the lower ones are non-lip pixels.

![Figure 5](image5.png)

**Figure 5.** YCbCr chromatic components histogram. (left) lip pixels (right) face pixels.

The Cr component for the lips is slightly higher than that for the face, and the difference between Cr and Cb for the lip is higher than that for the face. The previous Lip-Map equation was designed to highlight these properties by increasing the output value where Cr is high and Cb is low; \(Cr^2\) gives emphasis to pixels with higher Cr, and \(Cr/Cb\) completes the idea that lip regions have high Cr and low Cb values, [27].

Although this formula does not work properly (it cannot detect the whole lip region) for various kinds of images, and it is not flexible under various lighting conditions, [27], (according to the experiments of this study) it is still better to be used
as a starter for the proposed method rather than the other approaches. It can highlight at least (in the worst case) a small part of the lip (without being tricked by other details like the facial hair), which is enough for the initialization step of the proposed method.

For the reasons mentioned, the Lip-Map is used as an initial step for the proposed method, which works as follows:

1. Apply the Lip-Map formula on the ROI pixels.
2. Initially, cluster (K-mean) the output to lip or non-lip pixels (see Figure 6b).
3. For all the lip pixels calculate the mean of the r, g, b, warped hue, Cb, Cr, the X co-ordinate and the Y co-ordinate.
4. For each pixel in ROI, calculate the Euclidian distance from the mean vector in step 3.
5. Sort the distances array, assuming that at least the highest half of the distances array belong to non-lip pixels.
6. For all the pixels of the highest distances, calculate the mean of the r, g, b, warped hue, Cb, Cr, the X co-coordinate and the Y co-ordinate. Now we have 2 feature vectors, one representing the lip and the other representing the non-lip pixels.
7. For each pixel in ROI, calculate its distance to both feature vectors.
8. If the pixel is nearest to the lip feature vector then classify it as lip pixel, else classify it as face pixel.
9. Repeat 8 and 9 until no more pixels are clustered in ROI (see Figure 6c).
10. Create a binary image 1s representing the lip region, and apply the opening binary filter (erosion followed by dilation) to remove the noise (see Figure 7).
11. Consider the largest white area as the lips region.
12. In some cases where there are 2 large white neighbouring areas, consider both together as the lips region.

Figure 6. The different stages of the “nearest colour”, a) face detection followed by ROI defining, b) initial clustering using the Lip-Map, c) binary image of ROI resulting from the nearest colour algorithm, d) final lip detecting.

Figure 7. Applying the binary opening filter on the segmented ROI in Figure 6c.

After executing the previous algorithm, the mouth box can be easily determined using the output binary image, see Figure 6d. The colour information calculated in step 3 is used to separate the lip pixels from the face pixels, depending on their colours. While the x and y coordinates are used to discard (reduce the effect of) the pixels that have the same colour as the lips, but far away from the segmented area, such as some of the nose and chin pixels.
4. EXPERIMENTAL RESULTS AND EVALUATION

The proposed method and some of the mentioned approaches were tested to detect lips in a new-recorded video database (called the in-house database). This database was not only designed for this paper purposes, but also for a wider investigation, carried out at the University of Buckingham on VSR [30]. The results of this evaluation will enlighten us to choose the best approach for lip detection to be used within the investigated VSR system.

4.1 In-house database

This database consists of 26 participants of different races and nationalities (Africans, Europeans, Asians and Middle-Eastern volunteers), who were students and staff members from the Applied Computing Department in the University of Buckingham.

Each participant recorded 2 videos (sessions 1 and 2) at different times (about a 2 month period in time between the two recordings). The participants were 10 females and 16 males, distributed over different ethnic groups: 5 Africans, 3 Asians, 8 Europeans, and 10 Middle Eastern participants. Six of the males having both beard and moustache, and 3 males having moustache only. Video frame rate is 30 frames per second, and the average time for each video is about 8 minutes and 33 seconds. Each video contains about 15000 frames, so the total number of frames in the database ≈ 15,000 frame * 2 Sessions * 26 persons = 780,000 frames.

The videos were recorded inside a normal room, which was lit by a 500-watt light source, using Sony HDR-SR10E high definition (HD) 40GB Hard Disc Drive Handy-cam Digital Camcorder – 4 Mega Pixels. The videos were de-interlaced then compressed using Intel IYUV codec, converted to AVI format, and resized to (320 x 240) pixels, because it is easier to deal with AVI format, and it is faster for training and analyzing the videos with smaller frame sizes.

4.2 Experiments and results

No specific researcher image-based method was evaluated; instead, the whole approach, which is used by several researchers, has been evaluated by this study. For the image-based approaches, the evaluation of each was built on clustering the ROI pixels into lips or non-lips pixels; using the information obtained from that approach, the clustering method was the k-mean clustering algorithm.

For the model-based approach, ASM was evaluated on behalf of this approach. This time the whole method was applied as it is, using the MASM software, designed, and developed by Milborrow and Nicolls, [8]. The method was the same as the one introduced by Cootes et al., [6], but with adding more landmarks. Results are improved by fitting more landmarks.

The proposed method was tested to evaluate lip detection for about 780,000 video frames from the in-house database. The evaluation was done by human subjectively, by observing the detected lips allowing about 3 pixels plus/minus in all directions. If the mouth box was drawn to include the real mouth within this margin it is considered as true detection, otherwise, it is false detection.

It is a very difficult task to evaluate all the video frames manually. Instead, 10 random samples from each video was evaluated, each random sample of size 100 random video frames. The result of the lip detection method is observed and approximated by human intelligence. The observer has to decide whether the lips are detected or not in each frame. Then the observer averages the results for each video, then for each person.

The in-house database, which was created to investigate the lip-reading problem, consists of several ethnic groups. Most of the investigated methods depend on the use of colour information as a key factor. Hence, the lip detection results should be affected by the variation in the skin and lip colour due to ethnicity. The results in Table 2 depict the performance of the different lips detection schemes (including ours) for the different ethnic groups.

Both tables show the superiority of the proposed “nearest colour” algorithm, with accuracy of about 91% and a significant difference (13%) from its next rival algorithm. This method is qualified more than the other ones to detect the lips in 91% confidence for the final step (the lip reading). Figure 8 illustrates some correctly detected lips of some subjects in the in-house database, and Figure 9 illustrates the false detection. The nearest colour method makes use of the YCbCr (just for the initial step) to get some of the real lips pixels. The colours of these pixels may differ from person to person, which is not important for this method, because it compares these pixels with the same image pixels, trying to find other similar pixels to declare them as lip pixels. The difference between the lips and the skin colour does not
matter, because each pixel is classified depending on its distance to either the lips averages, or to the skin averages, regardless of how close these averages are.

Table 1. Results of lip detection using different approaches.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Hue</th>
<th>rgb</th>
<th>Hybrid Edge</th>
<th>YCbCr</th>
<th>ASM</th>
<th>Nearest Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female_01</td>
<td>67.74%</td>
<td>45.83%</td>
<td>97.56%</td>
<td>84.09%</td>
<td>22.73%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Female_02</td>
<td>64.37%</td>
<td>39.02%</td>
<td>71.76%</td>
<td>100.00%</td>
<td>95.83%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Female_03</td>
<td>89.52%</td>
<td>97.01%</td>
<td>98.90%</td>
<td>95.56%</td>
<td>94.44%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Female_04</td>
<td>92.11%</td>
<td>88.46%</td>
<td>85.29%</td>
<td>100.00%</td>
<td>73.33%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Female_05</td>
<td>69.92%</td>
<td>48.89%</td>
<td>86.21%</td>
<td>95.35%</td>
<td>90.48%</td>
<td>89.86%</td>
</tr>
<tr>
<td>Female_06</td>
<td>61.29%</td>
<td>51.48%</td>
<td>88.37%</td>
<td>87.35%</td>
<td>100.00%</td>
<td>87.80%</td>
</tr>
<tr>
<td>Female_07</td>
<td>52.68%</td>
<td>32.14%</td>
<td>82.09%</td>
<td>43.90%</td>
<td>54.17%</td>
<td>61.90%</td>
</tr>
<tr>
<td>Female_08</td>
<td>66.92%</td>
<td>100.00%</td>
<td>80.00%</td>
<td>88.10%</td>
<td>14.29%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Female_09</td>
<td>64.92%</td>
<td>97.01%</td>
<td>98.90%</td>
<td>95.56%</td>
<td>50.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Female_10</td>
<td>72.66%</td>
<td>47.73%</td>
<td>54.88%</td>
<td>85.71%</td>
<td>4.35%</td>
<td>92.86%</td>
</tr>
<tr>
<td>Male_01</td>
<td>62.50%</td>
<td>86.89%</td>
<td>24.49%</td>
<td>71.43%</td>
<td>52.94%</td>
<td>68.18%</td>
</tr>
<tr>
<td>Male_02</td>
<td>100.00%</td>
<td>11.29%</td>
<td>42.86%</td>
<td>12.50%</td>
<td>64.71%</td>
<td>80.95%</td>
</tr>
<tr>
<td>Male_03</td>
<td>65.55%</td>
<td>4.88%</td>
<td>21.62%</td>
<td>19.51%</td>
<td>71.43%</td>
<td>68.18%</td>
</tr>
<tr>
<td>Male_04</td>
<td>100.00%</td>
<td>98.48%</td>
<td>92.86%</td>
<td>100.00%</td>
<td>50.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Male_05</td>
<td>72.66%</td>
<td>4.88%</td>
<td>21.62%</td>
<td>19.51%</td>
<td>71.43%</td>
<td>68.18%</td>
</tr>
<tr>
<td>Male_06</td>
<td>100.00%</td>
<td>98.48%</td>
<td>92.86%</td>
<td>100.00%</td>
<td>50.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Male_07</td>
<td>72.66%</td>
<td>4.88%</td>
<td>21.62%</td>
<td>19.51%</td>
<td>71.43%</td>
<td>68.18%</td>
</tr>
<tr>
<td>Male_08</td>
<td>85.71%</td>
<td>97.01%</td>
<td>98.90%</td>
<td>95.56%</td>
<td>90.48%</td>
<td>89.86%</td>
</tr>
<tr>
<td>Male_09</td>
<td>91.20%</td>
<td>25.64%</td>
<td>40.00%</td>
<td>52.50%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Male_10</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_11</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_12</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_13</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_14</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_15</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_16</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>Male_17</td>
<td>84.92%</td>
<td>52.50%</td>
<td>68.18%</td>
<td>82.50%</td>
<td>73.68%</td>
<td>92.68%</td>
</tr>
<tr>
<td>average</td>
<td>77.62%</td>
<td>50.87%</td>
<td>68.65%</td>
<td>64.58%</td>
<td>64.22%</td>
<td>91.15%</td>
</tr>
</tbody>
</table>

Table 2. Categorical results of lip detection using different approaches

<table>
<thead>
<tr>
<th>Category / Method</th>
<th>Hue</th>
<th>RGB</th>
<th>Hybrid Edge</th>
<th>YCbCr</th>
<th>ASM</th>
<th>Nearest Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td>74.45%</td>
<td>53.90%</td>
<td>87.92%</td>
<td>72.44%</td>
<td>42.30%</td>
<td>87.60%</td>
</tr>
<tr>
<td>Asian</td>
<td>79.64%</td>
<td>29.76%</td>
<td>76.56%</td>
<td>67.07%</td>
<td>97.92%</td>
<td>98.08%</td>
</tr>
<tr>
<td>European</td>
<td>73.95%</td>
<td>86.44%</td>
<td>78.17%</td>
<td>83.01%</td>
<td>62.67%</td>
<td>96.34%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>80.56%</td>
<td>67.89%</td>
<td>92.47%</td>
<td>82.22%</td>
<td>79.90%</td>
<td>97.10%</td>
</tr>
<tr>
<td>Beard And Moustache</td>
<td>74.96%</td>
<td>32.22%</td>
<td>40.29%</td>
<td>47.95%</td>
<td>48.90%</td>
<td>86.38%</td>
</tr>
<tr>
<td>Moustache</td>
<td>78.93%</td>
<td>16.39%</td>
<td>27.43%</td>
<td>20.66%</td>
<td>59.32%</td>
<td>86.43%</td>
</tr>
<tr>
<td>Female</td>
<td>76.54%</td>
<td>66.28%</td>
<td>87.21%</td>
<td>86.10%</td>
<td>67.86%</td>
<td>93.96%</td>
</tr>
<tr>
<td>Male</td>
<td>78.30%</td>
<td>41.25%</td>
<td>57.04%</td>
<td>51.13%</td>
<td>61.94%</td>
<td>89.39%</td>
</tr>
<tr>
<td>Average</td>
<td>77.62%</td>
<td>50.87%</td>
<td>68.65%</td>
<td>64.58%</td>
<td>64.22%</td>
<td>91.15%</td>
</tr>
</tbody>
</table>

For the other approaches, the larger (apparent) the difference between the lip colour and the skin colour, the more accurate the clustering method will be. An interesting result is the relatively good results of the Hue approach (77.62%).
which is better than both the YCbCr, and the Hybrid edge approaches results. The Hue approach performance is better because it is not much affected by the appearance of the facial hair (74.96% and 78.93%), while it is less than 50% for both of the YCbCr and the Hybrid edge approaches. Because the hue of the facial hair is nearest to the skin colour rather than to the lip colour, this is what makes the Hue approach the only approach which performs better with males (78.30%) rather than females (76.54%).

Nevertheless, the proposed “nearest colour” method is affected less than the Hue approach when it comes to the facial hair (about 86%), because the facial hair colour values (including their hues) are closer to the skin colour. Therefore, they are considered as face pixels.

Another interesting result is the (87.92%) accuracy rate of the Africans by the “Hybrid Edge”, which is slightly higher than the “nearest colour” (87.60%). This is the worst result in terms of ethnic groups, compared to some other ethnic groups (more than 96%). This is because the colour of the lips and the skin for the Africans is almost similar, and very overlapped. This is not the case when it comes to the other ethnic groups. The “Hybrid Edge” is backed up with the edges information as well as the hue information. So, using sources of information other than the overlapped colour, such as the edges, makes the results better for the Africans (87.92%).

In general, the female results are better for most of the methods, although most of the female participants were not wearing full makeup, but still the females’ lips are more highlighted than males in general, and in the evaluation database females’ lips are even highlighted by makeup or by nature. Another more important reason is the lack of facial hair in the case of females, which greatly affects the detection method.

The RGB approach gives the worst results, because luminance and colour information are mixed together in the 3 channels, the red, green and the blue. This makes such an approach greatly affected by light conditions, shadows, wrinkles and facial hair.

![Figure 8. “Nearest Colour” algorithm accurate lips detection samples.](image)

Surprisingly, the performance of the ASM method (64%) was much worse than the proposed method (-27%), and worse than most of the image-based methods. Figure 10 illustrates accurate lip detection, and Figure 11 illustrates inaccurate lip detection for subjects from the in-house database. The low performance for ASM is due to the problems associated with the use of model-based approaches in general, and the use of ASM in particular, which were mentioned earlier. This proves and justifies the use of an image-based approach instead, for the new proposed methods in this study.

The convergence to local minima is one of the most common problems of the ASM. This problem can be noted in Figure 11 where a local minimum happened as a result for The appearance of the facial hair, which makes some contours look like the lips contour, the inner contour of the mouth, especially when the mouth is widely opened or other facial features like the nose or the chin contour.
5. CONCLUSION AND FUTURE WORK

In this paper we reviewed some of the most well known state-of-the-art lips detection methods, divided into two main categories: the model-based methods, and the image-based methods. Some of these approaches were implemented and evaluated using a special database, which is recorded for the lip-reading problem investigation. We also proposed a new method for lip detection, the “nearest colour” method. The results demonstrated that the nearest colour algorithm outperforms the state-of-the-art lips detection methods, and it can be applied in real time for online applications, such as the VSR.
A 91.15% lip localization accuracy is not good enough for a robust lip-reading system. Therefore, we are intending to work further on this area, trying to increase the accuracy by incorporating other type of information such as the edges and the motion, and decrease the processing time of the proposed method.

ACKNOWLEDGMENTS

The first author would like to thank the Jordan Army and Mu'tah University for sponsoring his PhD programme of study. He would also like to thank Dr. Harin Sellahewa for his valuable comments during the preparation of this paper.

REFERENCES