Comparative Analysis of Lip Features for Person Identification
Usman Saeed
COMSATS Institute of Information Technology
Lancaster Block, Defence Road,
Off Raiwind Road, Lahore, Pakistan.
+92 (42) 111-001-007
DrUsmanSaeed@ciitlahore.edu.pk

ABSTRACT
Humans unconsciously use visual cues from lip motion not only in aural conversations but also to communicate subtle emotions. In the past decades lip detection and segmentation has been extensively studied in various fields from speech recognition and human computer interaction to biometrics, but literature comparing visual lip features is almost non-existent. In this paper we present a comparative evaluation of different visual features from lips. Geometric and appearance based features were extracted and their relevance to identifying people was studied by feature selection methods.

Categories and Subject Descriptors

General Terms

Keywords
Biometrics, Feature Extraction and Selection.

1. INTRODUCTION
Audio/Video Speech Recognition (AVSR) and facial expression recognition have been the dominant fields of research in lip motion features. In AVSR, it has now been extensively proven that the recognition rate can be significantly improved by combining audio and visual features. These improvements have been mainly attributed to two sources, one is that visual features simply provide more data for recognition and the other is that they compensate when audio features are degraded by noise.

A relatively new addition to the use of lip features is in the field of person recognition [1], where the aim is to use only lip motion to recognize people based on physical and behavioral parameters. Studies have shown that given enough training data a behavioral pattern does emerge in lip motion that is text independent and can be used to recognize individuals. On the other hand it should not be forgotten that this field is still nascent with two main shortcomings. The first is that like in all biometric techniques based on behavioral parameters it requires larger amount of data per subject as compared to physical parameters, which unfortunately is not present in any publicly available databases today. The second is the lack of a study to compare various lip features for person recognition. Although some effort has been made in this regard but it has either been limited by the diversity of features extracted [6] or has sufficed by using a classifier as a feature selection method.

In this paper we aim to compare visual features from lip motion for their relevance to person recognition. For this propose we have extracted various geometric and appearance based lip features and compared them using three feature selection measures. The rest of the paper is divided as follows. In Section 2 we elaborate the previous work done, then we detail the proposed algorithms in section 3, after that we report and comment our results in section 4 and finally we conclude this paper with remarks and future works in section 5.

2. Previous Work
Literature relating to evaluation of lip features is almost non-existent, other than what has already been presented in the introduction. Therefore for the previous work we have decided to divide the problem into subsections and present the literature review accordingly.

2.1 Feature Extraction
Visual lip features can be broadly divided into two categories, static and dynamic. Static features provide a snapshot of the lip shape and appearance at a specific instant of time and characterize the physical aspect of the lip. The dynamic features represent the motion of the lip during speech or other labial activity and represent the behavioral aspect of lip.

2.1.1 Static
The static features can be further divided into appearance, shape and hybrid features.

2.1.1.1 Appearance
Appearance features are based on the Region of Interest (ROI) around the detected lip, which is normally a rectangle. Color or grey level pixel values could be used directly but the
dimensionality of the feature vector becomes prohibitively large for classification, thus a dimension transformation is highly useful. Examples of such transforms are PCA generating eigenlips[2], 2-D Fourier Transform, the Discrete Cosine Transform (DCT), the Discrete Wavelet Transform (DWT), Linear Discriminant Analysis (LDA), Fisher Linear Discriminant (FLD), and the Maximum Likelihood Linear Transform (MLLT)[3]. LDA and FLD take separation between the pattern classes into account thus provide most discriminant features.

2.1.1.2 Shape Based
Shape based features assume that most of the discriminatory information is contained in the lip shape, thus several studies have extracted compact representations of the lip shape. Although they are invariant to size and lighting but they are difficult to extract and computationally intensive. They are further divided as geometric and model based features.

Geometric: Geometric features, such as the height, width, perimeter of the mouth have been used for speaker recognition [4]. Distances or angles between key points located on the lip margins, mouth corners, or jaw have been used in [5].

Model Based: Model-based visual features are normally used when a parametric or statistical facial feature extraction algorithm is used. In model-based approaches, the model parameters are directly used as visual speech features such as Deformable templates [6] and Active Shape Models (ASM) [7].

2.1.1.3 Hybrid
Shape and appearance based features have been combined into a single model for improving performance. PCA appearance features are combined with ASMs to create AAMs [8]. In [9], the lip feature vector is formed by concatenating the Karhunen Loeve Transform (KLT) coefficients of the inner-outer lip contour points with the texture information.

2.1.2 Dynamic
Dynamic features that capture the behavioral rather than the physical aspect of the lip have been used for some time in the speech recognition community and now have recently been added to the person recognition domain. Optical flow is the most common and easy to extract visual feature. In [10] dense optical flow is calculated, then these dense velocity vectors are quantized by allowing only 3 directions \((0^\circ, -45^\circ, 45^\circ)\), and only 20 values resulting in a feature vector of 40 parameters. These quantization values were obtained by fuzzy c-means clustering. Another method of extracting dynamic features is to first extract static features, then to derive dynamic features from these static features by taking derivative over a window. In [11], the radial magnitudes are measured from points around the circumference of the lip, stepping pixel by pixel to the mid point of the principal diagonal. The final lip signature is then derived by taking the DCT of the radial magnitudes. Dynamic features measuring the rate of change are then extracted by first and second order differentials, over a sliding window.

2.2 Feature Selection
The objective of feature selection is to choose a subset of features from a much larger set of features, given an evaluation criterion to compare the performance of the new subset. Traditionally several factors have been the motivation behind feature selection, first and foremost has been to reduce the number of features to save time and computational complexity for the classification process, secondly to improve classification results by removing irrelevant parameters, and lastly to prevent over-fitting due to parameters which are mostly noise. Feature selection algorithms can be divided into the following categories.

2.2.1 Filter Schemes
These methods act as pre-filters to the classification task and involve using certain generic characteristic of the data such as correlation, information gain or entropy [12], to select and rank subsets of features. Mostly this group of techniques is used with large number of features as they have better computational complexity. They also tend to provide better generalization as compared to other techniques at comparable accuracy.

2.2.2 Wrapper Schemes
These techniques are feature selection techniques [13] that are specifically based on a classifier and use the classifier's ability to evaluate the relevance of the feature subset. This leads to feature subsets that are highly accurate for a certain classifier at the cost of generalization. These are also computationally expensive so may be inappropriate for large variable sets.

3. Proposed Method
The proposed system can be divided into three main parts which are lip detection, lip feature extraction and feature selection. In the feature extraction phase both geometric and appearance based features are extracted. In the feature selection we have employed three techniques to rank relevant features, the details are as follows.

3.1 Lip Contour Detection
In this section we present a lip detection method to extract the outer lip contour. The algorithm combines edge based and segmentation based algorithms using OR fusion. Given a database image containing a human face the first step is to select the mouth Region of Interest (ROI) using the tracking points provided with the database. Next two detection methods are applied to the mouth ROI independently, one based on edge detection and the other on segmentation. Finally the results from the two methods are fused to obtain the final outer lip contour.

3.1.1 Edge Based Detection
The first algorithm is based on a well accepted edge detection method, it consists of two steps, the first one is a lip enhancing color transform and the second one is edge detection based on active contours (cf. Figure 1). Several color transforms have already been proposed for either enhancing the lip region independently or with respect to the skin. Here, after evaluating several transforms we have selected the color transform proposed by [14]. It is based on the principle that blue component has reduced role in lip / skin color discrimination. It has been defined as:

\[
I = \frac{2G - R - 0.5B}{4}
\]

Where R,G,B are the Red, Green and Blue components of the mouth ROI. The next step is the extraction of the outer lip contour, for this we have used active contours [15]. Active contours are an edge detection method based on the minimization of an energy associated to the contour. This energy is the sum of
internal and external energies; the aim of the internal energy is to maintain the shape as regular and smooth as possible. The most straightforward approach grants high energy to elongated contours (elastic force) and to high curvature contours (rigid force). The external energy models the edge of the object and is supposed to be minimal when the active contours (snake) are at the object boundary. The simplest approach consists of using regularized gradient as the external energy. In our study the contour was initialized as an oval half the size of the ROI with node separation of four pixels.

![Figure 1: a) Mouth ROI, b) Color Transform, c) Edge Detection.](image)

### 3.1.2 Segmentation Based Detection

In contrast to the edge based approach the second approach is segmentation based on a color transform in the YIQ domain. As in the first approach we experimented with several color transform presented in the literature to find the one that is most appropriate for lip segmentation. [16] have presented that skin/lip discrimination can be achieved successfully in the YIQ domain, which firstly de-couples the luminance and chrominance information. They have also suggested that the I channel is most discriminant for skin detection and the Q channel for lip enhancement. Thus we transformed the mouth ROI form RGB to YIQ color space and retained the Q channel for further processing (cf. Figure 2).

In classical active contours the external energy is modeled as an edge detector using the gradient of the image, to stop the evolution of the curve on the boundary of the desired object while maintaining smoothness in the curve. This is a major limitation of the active contours as they can only detect objects with reasonably defined edges. Thus for the second method we selected a technique called “active contours without edges” [17], which models the intensities in different region of the image and uses it as the stopping term in active contours. More precisely this model [17], is based on Mumford–Shah functional and level sets. In the level set formulation, the problem becomes a mean-curvature flow evolving the active contour, which will stop on the desired boundary. However, the stopping term does not depend on the gradient of the image, as in the classical active contour models, but is instead based on Mumford–Shah functional for segmentation.

![Figure 2: a) Mouth ROI, b) Color Transform, c) Region Detection](image)

### 3.1.3 Error Detection and Fusion

Lip detection being an intricate problem is prone to errors, especially the lower lip as reported by [18]. We faced two types of errors and propose appropriate error detection and correction techniques. The first type of error, which was commonly observed, was caused when the lip was missed altogether and some other feature was selected. This error can easily be detected by applying feature value and locality constraints such as the lip cannot be connected to the ROI’s boundary and cannot have an area value less than one-third of the average area value in the entire video sequence. If this error was observed, the detection results were discarded.

The second type occurs when the lip is not detected in its entirety, e.g. missing the lower lip, such errors are difficult to detect thus we proposed to use fusion as a corrective measure, under the assumption that both the detection techniques will not fail simultaneously.

The detection results from the above described methods were then fused using OR logical operators. The outer lip contours are used to create binary masks which describe the interior and the exterior of the outer lip contour. Table 1 presents the commonly observed errors and the effect of OR fusion on the results.

<table>
<thead>
<tr>
<th>Error Type 1</th>
<th>Error Type 2</th>
<th>No Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Segmentation Based" /></td>
<td><img src="image" alt="Edge Based" /></td>
<td><img src="image" alt="OR Fusion" /></td>
</tr>
</tbody>
</table>

### 3.1.4 Experiments and Results

In this section we elaborate the experimental setup and discuss the results obtained. Tests were carried out on Valid Database [19] which consists of five recording sessions of 106 subjects using the third utterance. One image was extracted from each of the five videos to create a database of 530 facial images. The reason for selecting one image per video was that the database did not contain any ground truth for lip detection, so ground truth had to be created manually, which is a time consuming task. The images contained both illumination and shape variation; illumination from the fact that they were extracted from all five videos, and shape as they were extracted from random frames of speaker videos.

To evaluate the lip detection algorithm we used the following two measures proposed by [20], the first measure determines the percentage of overlap (OL) between the segmented lip region A and the ground truth A_G. It is defined as
\[
OL = \frac{2(A \cap A_o)}{A + A_o} \times 100
\]

Using this measure, total agreement will have an overlap of 100%. The second measure is the segmentation error (SE) defined as

\[
SE = \frac{OLE + ILE}{2 \times TL} \times 100
\]

OLE (outer lip error) is the number of non-lip pixels being classified as lip pixels and ILE (inner lip error) is the number of lip-pixels classified as non-lip ones. TL denotes the number of lip-pixels in the ground truth. Total agreement will have an SE of 0%.

Initially we calculated the overlap and segmentation errors for edge and segmentation based methods individually, and it was visually observed that edges based method was more accurate but not robust and on several occasions missed almost half of the lip. On the other hand segmentation based method was less accurate but was quite robust and always succeeded in detecting the lip.

### Table 2: Lip detection Results

<table>
<thead>
<tr>
<th>Lip Detection Method</th>
<th>Mean Segmentation Error (SE) %</th>
<th>Mean Overlap (OL) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation Based</td>
<td>17.8225</td>
<td>83.6419</td>
</tr>
<tr>
<td>Edge Based</td>
<td>22.3665</td>
<td>65.6430</td>
</tr>
<tr>
<td>OR Fusion</td>
<td>15.6524</td>
<td>83.9321</td>
</tr>
</tbody>
</table>

OR fusion was then applied and the best results were observed for OR fusion (cf. Table 2), around 15%. An aspect of the experiment that must be kept in mind is the ground truth. Although every effort was made to establish an ideal ground truth but due to limited time and resources some compromises had to be made.

### 3.2 Feature Extraction

In the feature extraction phase both geometric and appearance based features are extracted.

#### 3.2.1 Geometric Features

The geometric features that we have extracted for this study include area, length of major and minor axis, eccentricity, orientation and length of the perimeter of the outer lip contour. The algorithm takes as input the outer lip contour detected in the previous module, next the shape of the lip is characterized by several features which include the area contained inside the outer lip contour, the length of the major and minor axis and the perimeter of the lip contour. Eccentricity and orientation were also calculated from the major and minor axis to be used as a person independent feature. The output of the module for each frame of the video is organized in a feature vector for the feature selection module.

#### 3.2.2 Appearance Based Features

We have employed several appearance based techniques to extract visual feature vectors. As these techniques result in a wide range of different type and size of features, so to provide a fairer comparison we have limited the size of all feature vectors to a standard of 300 features per frame.

### Pixel Intensity Profiles

Based on the lip contour seven vertical and one horizontal scan were defined (cf. Figure 3). Pixel intensity levels from these scan lines were then concatenated to form a feature vector of size 300.

![Figure 3: Pixel intensity profiles.](image)

**Mean Removed PCA (MR-PCA):** as proposed by [21] consists of calculating the average mouth image from a video sequence and then subtracting it from each frame. This enables us to remove unwanted variation that is static and subject dependent. First a mean mouth image was calculated as

\[
\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t
\]

Then new mean removed frames were calculated as

\[
y^* = y_t - \bar{y}
\]

These mean removed mouth frames are then used to create a PCA subspace where 300 of the maximum modes of variations are preserved.

**Optical Flow:** We have calculated the dense optical flow with Lucas-Kanade method and then to reduce the feature vector to the standard size of 300 we apply an averaging filter.

**Spatio-Temporal Templates (STT):** and similar techniques [22] are generated by accumulating frame by frame image difference in a video sequence. STT captures both the location and time of motion occurrence and removes static objects. First intensity values were subtracted in consecutive frames. Then these difference images were binarized. Finally time information was added implicitly by multiplying frames with a time factor, which was a linear ramp function in our case.

**Discrete Fourier Transform (DFT):** Two dimensional fast Fourier transform was calculated for each mouth frame and 300 of the highest frequencies were preserved as the feature vector.

**Independent Component Analysis (ICA):** was used to create a subspace preserving 300 maximum variation and then feature vectors were projected into this space.

**Discrete Wavelet Transform (DWT):** A second order 2-D DWT was applied to each mouth image using Haar wavelets converting low-low coefficients successively.

**Discrete Cosine Transform (DCT):** was applied to each mouth frame and then 300 highest coefficients were selected by a zigzag scan.

### 3.3 Feature Selection

The next module evaluates feature extracted from the previous module for relevance to person identification. We have selected one advance technique and two basic ones which are described below.

#### 3.3.1 mRMR Feature Selection

Most of the filter based techniques are based on the concept of simple ranking, in which features are first ranked and the top most
ranking features are selected. Whereas it is quite possible that the
selected features are highly correlated and thus redundant. Minimal-Redundancy-Maximum-Relevance (mRMR) proposed
by [23] aims to solve this problem by first selecting features set S
that has maximum relevance between feature $x_i$ and target class $c$
by means of a similarity measure such as mutual information.

$$\max D(S, c) = \frac{1}{|S|} \sum_{i,j \in S} I(x_i; c)$$

Then reducing redundancy by selecting features that are
maximally dissimilar to each other as

$$\min R(S) = \frac{1}{|S|} \sum_{i \neq j \in S} I(x_i; x_j)$$

Finally the above two criteria are combined and optimized as

$$\max \Phi(D, R) = D - R$$

### 3.3.2 Bhattacharyya Distance (BD)

Bhattacharyya distance measures the similarity between two
discrete probability distributions. For discrete probability
distributions $p$ and $q$ over the domain $X$, it is defined as:

$$D_B(p, q) = -\ln(BC(p, q))$$

Where BC is the Bhattacharyya coefficient and is defined as

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

### 3.3.3 Mutual Information (MI)

Mutual information of two random variables $X$ and $Y$ is the
quantity that measures the mutual dependence of the two variables
and for discrete variables can be defined as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

where $p(x, y)$ is the joint probability distribution function of $X$ and
$Y$, and $p(x)$ and $p(y)$ are the marginal probability distribution
functions of $X$ and $Y$ respectively.

### 4. Experiments and Results

In this section we elaborate the experimental setup and discuss the
results obtained from the feature selection module. Tests were
conducted on a Valid Database [19] which consists of five
recording sessions of 106 subjects using the third utterance.
Videos were first converted to sequence of images at a frame rate
of 25fps and then feature extraction techniques (ICA, etc.) were
applied to individual frames to extract feature vectors. Then
feature vectors were concatenated and analyzed using feature
selection methods.

Table 3 provides the ranking results according to the feature
selection techniques described before. Just looking at the first four
columns of Table 3 at first glance the results may appear to be
random and uncorrelated but on a deeper analysis some
conclusions can be deduced such as the Discrete Fourier Transform has average performance across all feature selection techniques.

<table>
<thead>
<tr>
<th>Rank</th>
<th>mRMR</th>
<th>BT</th>
<th>MI</th>
<th>Rank Fused</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ICA</td>
<td>PCA</td>
<td>DWT</td>
<td>ICA</td>
</tr>
<tr>
<td>2</td>
<td>Intensity Profiles</td>
<td>DCT</td>
<td>DFT</td>
<td>DFT</td>
</tr>
<tr>
<td>3</td>
<td>STT</td>
<td>ICA</td>
<td>Geometric</td>
<td>STT</td>
</tr>
<tr>
<td>4</td>
<td>DFT</td>
<td>DFT</td>
<td>PCA</td>
<td>DCT</td>
</tr>
<tr>
<td>5</td>
<td>DCT</td>
<td>Intensity Profiles</td>
<td>STT Intensity Profiles</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>DWT</td>
<td>STT</td>
<td>DCT</td>
<td>PCA</td>
</tr>
<tr>
<td>7</td>
<td>Geometric</td>
<td>Optical Flow</td>
<td>ICA</td>
<td>DWT</td>
</tr>
<tr>
<td>8</td>
<td>Optical Flow</td>
<td>Geometric</td>
<td>Optical Flow</td>
<td>Geometric</td>
</tr>
<tr>
<td>9</td>
<td>PCA</td>
<td>DWT</td>
<td>Intensity Profiles</td>
<td>Optical Flow</td>
</tr>
</tbody>
</table>

ICA always performs better than intensity profile method and that
optical flow has the one of the lowest performance for all the three
methods. Therefore to present a much clearer picture of the results
we decided to fuse the ranking results form the three techniques. The
fusion is a weighted sum of the ranks from the three
techniques and is given as

$$\text{Rank Fused} = (2 \times \text{mRMR}) + \text{BD} + \text{MI}$$

The mRMR technique is given twice the weight as compared to the
other techniques, the reason being that we believe the mRMR
is the most advanced technique and is better able to take the
mutual information and redundancy into account.

In the last column of Table 3 we can clearly see the fused ranking;
here we would like to comment about the best and the worst
performing technique. As expected the best performing technique
is ICA which is a supervised technique and takes the class
information into account. The two worst performing techniques
are the geometric features and optical flow. The reason behind
the poor performance of geometric features is they contain much less
information as compared to other features such as DFT i.e. size of
a geometric feature is 6 while that of DFT is 300, thus it is not a
fair comparison from the beginning. Whereas the poor performance of the optical flow features can be attributed to the
fact that they are purely behavioral feature and lack the physical
aspect of lip shape and appearance.

### 5. Conclusions

In this paper we have presented a comparison between various
feature extraction techniques and their relevance to person
recognition. Geometric and appearance based features were
extracted and three feature selection techniques were used to
come up with them. We observed weak correlation between the results
form the three methods and thus decided to fuse the ranking
results. A supervised technique ICA, gave the best performance,
where as geometric and optical flow features performed poorly.
Thus the instinctive notion is reinforced that appearance contains more information for person recognition than shape and behavior.

6. References


