Challenges in robust speaker - independent lip detection

Stefan Badura*1,2, Michal Mokryš*1,2, Anton Lieskovsky*1,2, Roman Badura*2

*1 Itall s.r.o., Univerzitina 8498/25, 010 08
*2 University of Zilina, FRI, Univerzitina 8215/1, 010 26

Zilina, Slovakia

Abstract—Lip detection is a crucial part in the process of automatic lip-reading system. This paper discusses simple methods for lip detection based on color, wavelet properties and template matching. In this paper we try to show on complexity of solved task. We use different databases of frontal faces with their different sizes and also in different, non-uniform lighting conditions. The results vary based on used method and also they distinguish based on used frontal view if faces for given inputs.

Keywords—lip detection, haar, HSV, RGB, YCrCb, PCA, LDA, template matching

1. INTRODUCTION

Automatic lip reading is interesting and very challenging task. Human can learn this process intuitively by watching lip shape and mouth movements. Although kin, lip surroundings hold also important information, but the most informative are the lip’s shape and the lip’s parameters. Before estimating these properties, an important part is reduction of whole processed image to a lip area - region of interest (ROI). The dimensionality reduction can be processed with proper lip detection and localization that is rather difficult task. The complexity is caused especially by lip shape variation for different speakers, lip color changes and there are also many changes in illumination, rotation, size, etc. This paper describes process for automatic lip detection algorithms which are important in automatic lip reading systems. The main goal of this paper is to propose an overview for variety of methods possibly used in this process. Different properties of input images are considered. For example color, haar, geometric properties or templates matching in different spaces are compared. Two databases in different frontal face scales are used for methods verification (see figure 1). We are not intending to provide the top lip finder for certain dataset, we are looking for methods which are useful for real conditions, and which are robust against, size, illumination and other unwanted changes in input images.

Proper lip localization is key part in our model for lip reading feature extraction. When no proper position of lips is known, lip reading system works with a lot of useless data. Lip detection is important process from two points of view:

- Region of interest initializing.
- It is a preprocessing stage – initialization for other processes – e.g. lips fitting or description.

Most methods for lip detection use color information. In [1] a rule-based technique can be found, which uses normalized RGB color space. They defined quadratic polynomial discriminant function on r-g plane, for detection of lips pixels. In [2] a chromatic curve was introduced, to define a border between lip and skin color, such approach should provide resistance against some lighting changes. Next, authors in [5] use a normalized RGB chrominance diagram for the same purpose.

The HSV color space for lip detection is used in [11]. In [12] H value of HSV is combined with A value of CielAB color space into chrominance detection model, bivariate probability model is estimated. Another different method for extraction of lips boundary is propped in [7], where Gaussian mixture model (GMM) of pixels for lip and non lip region is trained. The R/G color space is used in [6] where the second order statistics of adjacent pixels, and an unsupervised clustering technique based on the expectation maximization (EM) algorithm is employed. The R/G histograms are used in [8].

In [3] a boosting system was introduced using haar features. The haar features are used also in [13], where a set of weak
haar feature based detectors are combined into one adaboost system. In [4] the watershed segmentation for lip detection in large area without face localization is used. Most papers use color properties of skin and lips. In [9,10] a Spatial Fuzzy C-Means (SFCM) clustering algorithm is proposed, where also spatial interaction of neighboring pixels is considered.

In [7] PCA template matching for lip corner detection is used. Authors in [14] PCA for mapping pixels of lips and non-lips region for RGB color scheme is used.

There are many other approaches and their combination. In our experiments we discuss some of them. The main problem is their robustness in different situation and condition. Generally could be said, that most methods provide high positive recognition percentage for one dataset in specific condition, but when the condition rapidly change, a new model must be trained. In the following paragraph we discuss possibilities of simple, often used method for lip detection in different conditions.

Next section provides a method’s overview for lip detection. The 3th section presents experiments and compares algorithm’s efficiency for different conditions. The last section concludes this paper and proposes feature work.

II. LIP DETECTION METHODS

Effective and simple lip detection is the main motivation for proposed experiments. Straightforward approaches used for lip detection are based on color properties. The lip’s color and its well separation from face color distribution seem to be a good approach for lip extraction. When considering color properties, different color schemes can be used. In our experiments we use RGB, HSV and YCrCb. The simplest method is based on finding thresholds for each color component for all pixels. Looking for threshold experimentally is complicated, therefore normalized RGB or chrominance diagrams are computed and parameters for boundary between colors are estimated from them. Considering differences in color components for lips and skin lead to design of various models, for example one tested is a pseudo hue model. When considering not just pixel properties but also higher level organization, region based methods are proposed, which try then to estimate region color properties. As it is shown in [16], a lip template database is generated from a set of lips images (see figure 2). Then the whole image is looked up with floating window (with the size of template) and minimal distance from template to unknown sub-image is searched. It is basic template matching problem. Difference is just in database template representation.

The simplest approach is using lip images (as shown in figure 2) as templates, but it does not lead to efficient lip detection and it is time consuming. For dimensionality reduction varieties of data projection exist. Histograms generated from sub-images are one of them, or general tasks of linear regression of 1D or 2D data. 2D projections consider also pixels correlation, therefore they can be suitable. Well known methods for our purposes are principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Coefficients of projected lips images into subspace generated with mentioned projections are not dimensional and they are informative. These coefficients serve as templates and they can be searched in unknown image.

Often used approaches for lip detection are based on haar properties. These features, rather than using the intensity or color values of a pixel, use the change in contrast values between adjacent rectangular groups of pixels. An example of haar like features can be seen in figure 3.

The face has fixed structure (eyes, nose and mouth position), this fact can be efficiently utilized. Information about eye symmetry, nose position can lean to lip detection model design. Vertical distance between eyes and horizontal distance between mouth and eyes are in proportion, that can be computed.

Next class of methods combines more simple algorithms into one robust system. Different types of properties could be mixed into one single adaboosting model.

The affectivity of lip detection system can be enhanced using algorithms for face detection or other techniques for ROI reduction.

In next section we propose experiments with described methods. The goal is to examine its possibilities for lip detection.

III. EXPERIMENTS

In this section some experiments for the lip detection task are presented and discussed. Important parameters which are interesting from our point of view are:

- Robustness against lightening, size changes and model’s parameters stability – model generalization.
- Computational time.
- Final percentage of positive lip detection in unknown image.
The first examined parameter is the most important, because it is not problem to design method with high final percentage for uniform data in general, but the problem is noise in the data. Variation in illumination and different scales of inputs are not required but in real situation they often occur. In our experiments two datasets were tested. Some examples are shown in the figure 1. Images in the first database contain information just about mouth region. Images in the second dataset depict whole person’s face. We are interested in if given tested algorithm for one situation can be easily used in different conditions, just with its little modalities (a little modality could be e.g. database change). Unwanted behavior is when all parameters for given lip detector have to be searched always, when a change occurs.

All experiments were executed on personal computers with configurations: i3, 1.33 GHz, 4GB RAM; i7, 2GHz, 8GB RAM. The first database for our experiments was collected in our office using web cam (will be referenced as db1). All video were recorded in different times, with different lightening conditions for three speakers together 63 images. In those inputs a noise can be seen, where images are not uniformly illuminated. The second database is taken from the University of Aberdeen, where 116 unique people are shown in 690 images (referenced as db2) mostly frontal view.

A. Finding for lip’s color properties

Methods for finding lips based on their pixel color characteristics are often first choice, when lip detection system is designed. In our experiments we tested RGB, HSV, YCrCb color spaces. The RGB color space is not suitable because according [13] it is susceptible to light and brightness. From experiments it was clear that skin and lips have different distribution of pixels in RGB space. Green is stronger in skin then in lips. The difference between red and green is greater for lips than for skin. In [15] a pseudo hue model that describes these variations was proposed:

\[ h(x, y) = \frac{R(x, y)}{G(x, y) + R(x, y)} \]  

Where \( h \) is new hue value for all pixels of R,G components of RGB color space. Pseudo hue gives satisfied results when images are in constant illumination, otherwise the percentage remains around 50%.

As it was already mentioned, skin color varies widely, but the difference in chrominance is much smaller than in brightness. Therefore RGB chrominance diagrams are designed. The final percentage for chrominance lip detection algorithm moved around 50% of positive recognition for both databases. The example of not well separated pixels based on chrominance can be seen in the figure 4. The chrominance curve has sensitive to illumination changes.

When the chrominance diagram is used, a problem with luminance still remains; therefore the HSV color scheme is used. The HSV should be resistant against lighting changes, but with experiments we did not obtain better results than by RGB pseudo hue model or chrominance.

Another tested color space is YCrCb, where difference between red and blue components is better considered. The result remains unchanged.

B. Template matching – projection

All previously described methods are based on pixels processing. If the size of searched region is known then region based methods could be designed.

Searching for lip region can be considered as template matching. The difference is in template representation. First attempts are with templates as row images. Template is represented by mouth corner (left or right). These corners (templates) are sequentially searched in unknown image and the candidate with minimal distance is considered as corner (found pattern, see figure 5).

From our experiments the final percentage was not high, while the computational time was high and the resistance against illumination, size and rotation changes was not reliable. Another attempt lied on using histograms for each color component in HUE color space. Similarly to RGB also between H and S dependence for lips and non lips regions can be seen (see figure 6). The H value has peak near 0 for lips and S get close to the value 1. Histograms of given regions are

Figure 4 Pixel detected as lips for chrominance diagram algorithm.

Figure 5 Template matching. In the first row the original image is shown. In second row from the right: searched template, found sub-image, MSE of searching process, where dark areas presents high match.
generated and next projected into subspace using different basis (coefficients of harmonic base, polynomial, etc.).

From experiments we obtained, that histograms are not so stable. For different skin color peaks of histograms move, but H and S are in small correlation. Next experiment was designed to overlap mentioned histograms movements. A database from projected coefficients of histograms for different templates was created [16]. This approach provides friendly computational time with some resistance against illumination changes and also satisfied generalization for different data set. Problem occurs when new speakers were added into system, the final percentage rapidly decreased (see figure 7).

Using histograms has negative property: pixels space distribution and correlation is lost, that loses information. Then using histograms transmit the error into projection when the information is lost, and therefore other techniques were tested, those eliminate these disadvantages and remain still computationally satisfactory.

C. Well known projections

The most common used projections which were tested in our experiments are PCA and LDA.

PCA is trained on lips templates (see figure 2). In the figure 8 a projection of lips and non-lips images into 3 subspace is depicted (69% of information is preserved). The final percentage is low for both databases (less than 20%). When we try to preserve at least 90% of information (10 coefficients of subspace are used) the final percentage grows up to 85% for db1. For db2 the final results moved just around 40%. Higher score does not increase even when the ROI is reduced just to face region. For db2 we tried to execute other experiments where other parts of face are searched (nose, mouth corners), but the results were not satisfactory.

Other tested projection is LDA. This method tries to find projection that maximize between class scatter while minimizing within class scatter (see figure 9). In our experiments two classes were trained, where the first one was represented by lips templates and the second one by non-lips sub-images. Then we measured the distance between lips and non lips candidates (representatives) for unknown sub-image. In the figure 9 is shown plot, where lips and non-lips are clustered with LDA. The final percentage moved around 77% for db1 and around 67% for db2.

D. Haar features

Another approach often used in many works is based on extraction of so called haar features. This kind of methods provide some valuable results, but they are very sensitive to
noise, model parameters and sudden contrast changes. In our experiments the result for positive lip detection moved around 50% for db1 and less 30% for db2 with simple haar detector. If haar features were combined with effective blob selection and corner detection the final percentage moved up to 70%. In the figure 10 some results for this method are shown.

![Found lips - positives](image)

![Found lips - negative](image)

Figure 10 Lip detection using haar features. The first and the second rows show positive lip detection and the row negative detection with a lot of noise.

E. Boosting models

In some cases it is better to combine some simpler approaches to obtain better result. In our experiments we combined 4 simple approaches described in previous text into one adaboosting model (lip detection based on rgb chrominance diagram, haar features, lip corner template matching and the last one is face characteristic distribution). The Figure 11 shows an example of finding face characteristic, where eyes position is estimated.

![Lip detection –face characteristics detection](image)

Figure 11 Lip detection –face characteristics detection. The first and the last images show false positive eye detection.

The RGB chrominance diagram and haar feature were used for localization of lip candidates. Next corner template matching was applied for proper lip detection over all candidates. The face characteristic as eyes was used at the end as verification procedure. The final result moved around 90% for db2. Problem occurred when it should be generalized for the db1. Parameters had to be set experimentally and for example the last step cannot be executed because eyes are not visible.

Despite the disadvantages, we see possibility of using boosted methods for efficient robust lip detection system.

IV. CONCLUSION

The most difficulties in the process of lip detection were caused non uniform illumination of input images

From our experiments a summarization can be made as follows:

- Models using pixel color properties do not lead to robust and effective lip detection model.
- Although some experiments provide high positive results rate for one dataset, but if method is used in different situation, the whole approach is not well generalized and the results decreases rapidly.
- Algorithms are sensitive to illumination changes, illumination noise and their parameter initialization.
- Lip detection results decrease if there are people with different face color, when men wear beard or women lip stick, etc.
- All mentioned disadvantages lead to high parameter sensitivity for those methods.
- A template matching is computationally demanding, it could be executed in real time, when some projections are used. When we try to design satisfactory model a large database must be created and then the whole process becomes even more complicated.
- Well known projection provide more satisfactory results, with stable behavior, problem is with real time execution, especially when more coefficients are used.
- Some tested methods are useful just as preprocessing step (pseudo HUE) for lip localization, ROI reduction

In this paper we proposed an overview of well known and often used methods in the process of lip extraction. Our goal was to experimentally verify the usability of different algorithms in different situation. Experiments should provide us a basis for next processing in the task of automatic lip/reading.

ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract No. VMSP-II-09.

REFERENCES


