International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:8, 2009

A new Adaptive Approach for Histogram based Mouth Segmentation

Axel Panning, Robert Niese, Ayoub Al-Hamadi and Bernd Michaelis

Abstract—The segmentation of mouth and lips is a fundamental problem in facial image analyisis. In this paper we propose a method for lip segmentation based on rg-color histogram. Statistical analysis shows, using the rg-color-space is optimal for this purpose of a pure color based segmentation. Initially a rough adaptive threshold selects a histogram region, that assures that all pixels in that region are skin pixels. Based on that pixels we build a gaussian model which represents the skin pixels distribution and is utilized to obtain a refined, optimal threshold. We are not incorporating shape or edge information. In experiments we show the performance of our lip pixel segmentation method compared to the ground truth of our dataset and a conventional watershed algorithm.

Keywords—Feature extraction, Segmentation, Image processing, Application

I. Inroduction

Lip segmentation is a basic method for various applications. It can be utilized for lip reading, supporting speech recognition or expression analysis (i.e. facial expression, estimation of emotional state, pain recognition). Each application has its own limitation concerning speed, accuracy and robustness. The requirements for facial expression recognition can be very different depending on application context.

Many algorithms are proposed to work on color images [2]–[5], [9], [10]. Initially a color transformation[1] is performed to exploit the different chromaticity of lips from skin. In common these transformations focus on red and green color information, concerning the RGB color space. The output of the color transformation is a single channel intensity map of the mouth region, where the mouth is highlighted. However, also some works have been published working on normal gray images from monochrome cameras [6], [8], [12]. Basically, the segmentation approaches can be classified into two groups. The first group is focusing on detection of lip edges in the mouth region of interest (ROI) [2], [4]. They apply active contour models to the color transformed ROI [4] or use deformable Templates [2]. Some works stabilize their Active Contours using support tracking points [1], [6]. The general assumption of edge based algorithm is, that the lips generate prominent edges at the skin-lip crossing. In monochrome images a simple shadow casting can already cause serious problems. Colored images and their mouth-highlighting transformed representation can work around this issue. But still there is no guarantee the edges of the lips create significant edges here. This might happen for many cases. Mainly for asian people this rule holds true. However, for european/caucasian this rule does not hold for all cases anymore, since the transition from skin to lip pixels does not form rough edges here for all subjects and conditions. Histogram based approaches do not suffer this drawback, since they are not dependent on rough edges. This approach is a consequent continuation of the initial color transformation. The ROI is binarized into lip and non-lip pixels, where non-lip pixels are mainly skin pixels. The crucial point in histogram based algorithms is the definition of the threshold. A very easy approach, mostly used for first rough mouth segmentation is a fixed threshold, found by statistical average of numerous samples [7]. A more adaptive approach sets up a watershed like rule, which defines 15 percent of the darkest pixels in their color transformed mouth ROI as lip pixels [9]. Other works assume a certain topology in the histogram. Following this idea they seek for a local minimum between a lip and a skin heap in the histogram and define the threshold here [5]. Hybrid approaches combine both, color and edge information. Commonly deformable templates or active contours with an energy minimization referring to edge information and average color intensity inside of the template (or contour) are used here [2], [3].

In this paper we presume that the detection of face and mouth ROI have been solved yet. Methods for face detection [15] and mouth ROI [14] are available. An algorithm is proposed, which will rely to color information only. First, we select an adaptive threshold, which uses a very simple approach but ensures that only skin pixels and no lip pixels are selected. Second, based on an adaptively generated skin model distribution another refined threhsold is obtained. We do not use edge information in this method. Incorporation of edge information can be very useful but can be also disturbing. Especially when the mouth appears in different modes (closed, opened, opened with teeth, etc.) the interpretation of edges leads to complicated model assumptions. Therefore the incorporation of edges needs some higher level assistance. In this paper we want to show, what is the maximum available result using color information only. A hybrid method can be based upon the results presented in this article.

The rest of the paper is organized as follows. In section II we give some insight to the behavior and statistics of different color spaces concerning mouth region. Section III describes our approach for an adaptive thresholding of histogram of the mouth region. The results of our methods and comparisons to conventional watershed algorithm, which also use color information only are presented in section IV. Section V gives a short summery and outlook.

[1]Some important transformation rules are described in Table I in section II

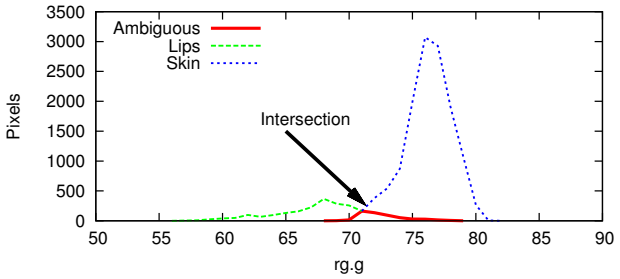


Fig. 1. The histogram of manual labeled skin and lip pixels for one sample image. Class C (teeth) is not displayed here. The intersection is the bulk of pixels which are ambiguous and can lead to classification errors.

II. Color Histogram Statistics

As mentioned in Section I color transformation to extract the mouth borders or at least to support such is commonly used. Our approach focuses on plain color based segmentation without using any edge or shape information. The mouth segmentation will be solved by plain histogram analysis. Basically, a mouth ROI can contain three classes of pixels: class A the lip-pixels; class B skin pixels and class C the teeth pixels. B and C can be joined to a single class: non-lip pixels. In our experiments we used 56 images under varying illumination conditions. The images are manually labeled with a ground truth for the classes A, B and C. The approximate sizes of the sample images are 160x80 pixels. With respect to different histogram structures for different mouth states the samples contain images from three different types of mouth states: (a) closed mouth, (b) opened mouth (also with teeth) and (c) mouth with pressed lips (see Fig. 2). However, to our knowledge we are unable to find any paper addressing the problem of teeth and their potential impact to the histogram of a mouth ROI.

A method exclusively based on histogram analysis needs a color transformation where the intersection $I = A \cap (B \cup C)$ of the lip and non-lip pixels is minimal (see Fig. 1). Section I referenced different approaches utilizing color for mouth highlighting. Some of the most common formulas are listed in Table I. Using the ground truth of our database, a comparative statistic can analyze their ability to separate lip from non-lip pixels, based on color information only. Motivated by prior experiments the green channel of the normalized rg[2] color space ($rg.g$) was also considered here. The rg color space is widely used for skin detection. It is not much different from $R/(R+G)$ but improves the results significantly. As results from this analysis the green channel from $rg.g$ was superior to all others (Table I). The worst results were achieved by the $YCbCr$ based color transformations. The best results were achieved by the normalized green channel from the $rg.g$ color space.

Intersection of skin and lip color in the mouth ROI with respect to different color transformations, with and without teeth appearance. The percentage is relative to the whole histogram and gives the false classified pixels (see Fig. 1) using optimal threshold by ground truth. $Luv.u$ is the u channel of the Luv space and $rg.g$ is the green channel of the rg space. The analysis also revealed that each of the both classes (skin and lip pixels) is following approximately the rules of a Gaussian distribution in the histogram. Under advantageous conditions lips and skin pixel form two well noticeable bell curves in the histogram with a noticeable local minimum in between (Fig. 2a). This can motivate approaches like [5], searching for this minor local minimum. However, these optimal cases cannot be assumed in general. The general structure of the histogram can vary in different scenarios (Fig. 2). More complex situations can create numerous minor local minima instead of only one major minimum. In other cases the smaller bell curve related to the lip pixels can be directly attached to the larger bell which represents the skin pixels without producing any local minimum (Fig. 2b). This multiple behavior can be observed independently from the chosen color transformation. However, surprisingly the appearance of the teeth had just a minor impact to the separability (see Table I) using a histogram threshold. Actually the teeth pixels occupy histogram slots in the upper intensity range of skin pixels after $rg.g$ color transformation. In some few cases they produce another small bell curve (attached or even separated) in the right of the skin pixel curve.

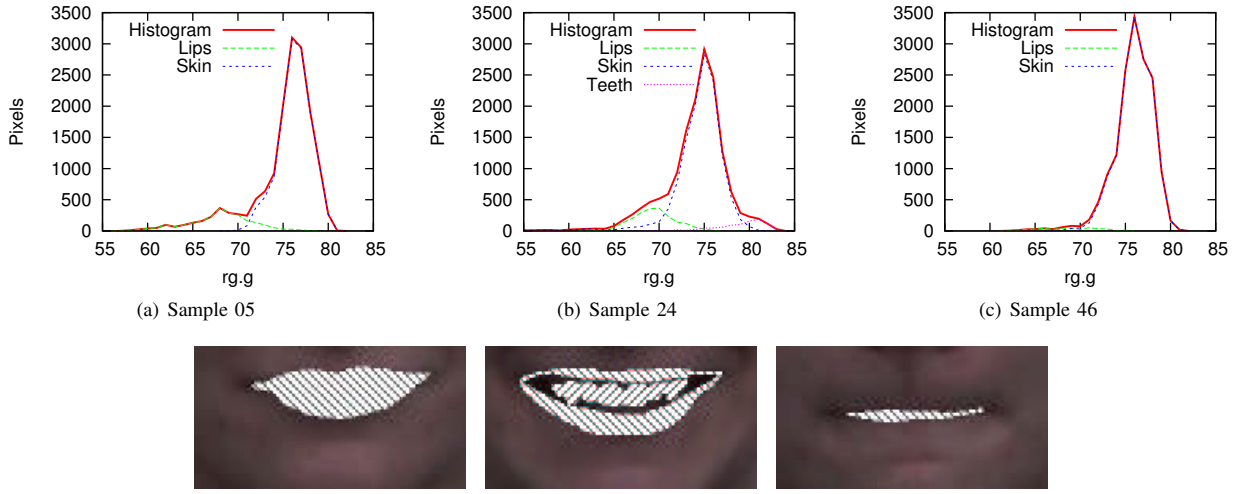
TABLE I
Aptitude of Color Models

Used in	Transf.	With teeth	No Teeth
[2]	$Luv.u$	4.61 %	3.16%
[11]	G/B	5.97 %	2.59%
[3]	G/R	2.30 %	1.45%
[4]	Cr	11.75 %	10.97%
[4]	Cr/Cb	13.08 %	11.16%
[9]	$R/(R+G)$	2.36 %	1.48%
not found	$rg.g$	**0.09%**	**0.38%**

III. Automatic Threshold Detection

Section II affiliates the $rg.g$ color space is optimal for lip segmentation from skin background. In this section we will present a simple but very efficient method to find an optimal threshold automatically. The problem is to identify the optimal threshold from the real histogram. Both pixel classes are Gaussian distributed after transformation to $rg.g$. One crucial constraint of a histogram based approach for mouth segmentation is to avoid wrong positive lip pixels (skin pixels classified as lip). They produce a kind of flow out which causes more damage to the segmentation than lip pixels which are classified as skin pixels. The reason can be found in the general shape structure of the histogram. Wrong positive lip pixels are basically caused by a too high allocated threshold, with respect to our $rg.g$ transformation (other transformations like the R/(R+G) the skin pixels have lower intensity than the lip pixels). With increasing intensity of $rg.g$ also the probability of adding a high amount of pixels to the lip pixel class in one single step is increasing (see Fig. 3). Therefore the algorithm prefers a too low threshold rather, instead of too high, because absolute number of wrong pixels will be smaller in that case. In contribution to that, the focus lies on the approximation of

[2] rg is the normalized RGB Space with $r = R/(R+G+B)$ and $g = G/(R+G+B)$



(a) Sample 05 (b) Sample 24 (c) Sample 46

Fig. 2. (Lower image row): Ground truth for our database. (a) normal mouth state, (b) open mouth with appearing teeth, (c) pressed lips with almost none lip pixels left. (Upper row): the corresponding histograms after transformation to $rg.g$ color space. The real histogram is the fat black line. The dotted lines represent skin/theeth (no-lip) and lips. These information are only gathered by ground truth here and are a-priori unknown in application case.

the skin pixels distribution to estimate the lower limit of the skin pixel intensities (see Fig. 3).

Let $h(i)$ be the value of the the i'th bin of the histogram and h_{max} the value of $h(i)$ with maximum peak in the histogram, x the intensity of a pixel in the color space transformed ROI. In order to get the Gaussian distribution G_{skin} of the skin pixels we calculate the $mean_H$ and the standard deviation σ_H from all x ROI satisfying the following two conditions:

$$h(x) > th_{oc} \tag{1}$$

$$x > th_{med} \tag{2}$$

with

$$th_{oc} = h_{max}/4 \tag{3}$$

$$th_{med} = (h_{low} + h_{high})/2 \tag{4}$$

see Fig. 3a for explanation. The normal probability density function is defined as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\Pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5}$$

Condition (1) is introduced with respect to the size ratio of lip pixels to skin pixels. Condition (2) is a median for the covered histogram bins with h_{low} being the first histogram bin $h(i)$ having more than 1 pixel (h_{high} is the equivalent from the back). These conditions operate as an a-priori hard coded threshold related to (1) the size ratio of mouth to skin and (2) the knowledge that lip pixels that are darker than skin pixels in $rg.g$. This can be considered as rough pre segmentation. For example, the median in condition (2) itself is a suitable adaptive threshold to divide lip from skin pixels. The two conditions were designed to avoid wrong positive lip pixels by any means, adjusted using ground truth of our base data. So ambiguities are excluded here by statistical examinations. Therefore the G_{skin} is based on reliable skin pixel information. In the next processing step this first rough result is refined using G_{skin}. This is the main contribution of our work enhancing it to be superior to other methods, which only use a-priori estimation on mouth size (e.g. [Kim06]). This Gaussian model allows the algorithm to select a foot point at the bell curve of the skin pixels with the following condition:

$$t = \underset{i}{argmin}\,(f(i) > \epsilon) \tag{6}$$

where $f(i)$ is the density funciton to G_{skin}. With respect to outliers and practical reasons we defined the threshold $\epsilon = 0.01$. In words the approach is to avoid wrong positive lip pixels. So the algorithm seeks the left footpoint of the gaussian distribution (which represents the skin pixels), where $(f(i)$ becomes insignificantly small.

IV. Experiments and Results

TABLE II
Variance of the Results

Resultbase	TP-μ	TP-σ^2	FP-μ	FP-σ^2
Ground truth	81.75%	3.88%	2.262%	0.002%
Proposed	80.24%	2.55%	2.200%	0.031%
Watershed 13%	79.16%	1.98%	3.917%	0.223%
Watershed 15%	87.74%	0.93%	5.933%	0.339%

In our experiments the algorithm was tested on the same 56 images which already were base of the statistical pre-analysis. The images have a size of approximately 160x80 pixels. The mouth sizes vary between in width between 120 and 150 pixels. The mouth height varies between 20 and 70 pixels. The higher variance is due to the opening of the mouth as greater impact to the height than the e.g. smiling has impact to the width. The images are partially from the Faces Database from CIT [13] and partially from own recored data with various illumination conditions (see Fig. 6). Our method was compared with the ground truth and a conventional watershed algorithm which are discussed in following:

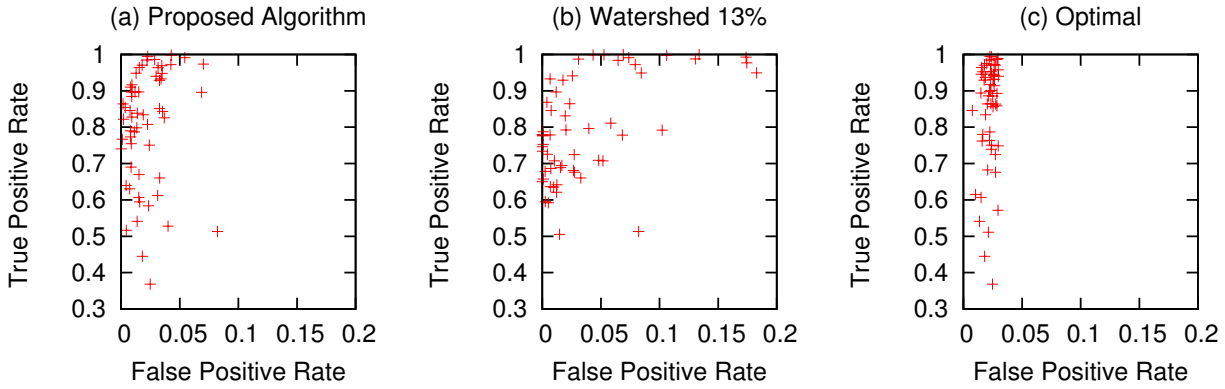


Fig. 5. The distribution of all single results in TPR/FPR space. Our proposed algorithm (a) has significantly less FPR than the watershed (b). The optimal thresholds were optimized according FPR ($< 5\%$) by the manual labeled ground truth.

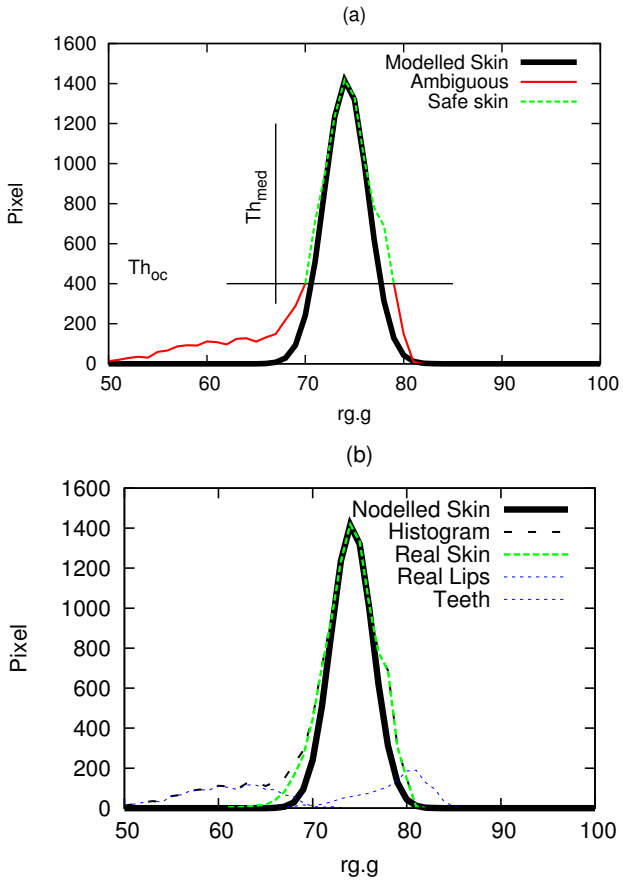


Fig. 3. Based on our rough assumptions, we can define a class of reliable skin pixels. From that one we can approximate a modeled skin pixel distribution G_{skin} (dotted line). G_{skin} is stretched here by a factor to visualize it's similarity to the genuine skin pixel distribution taken from our ground truth.

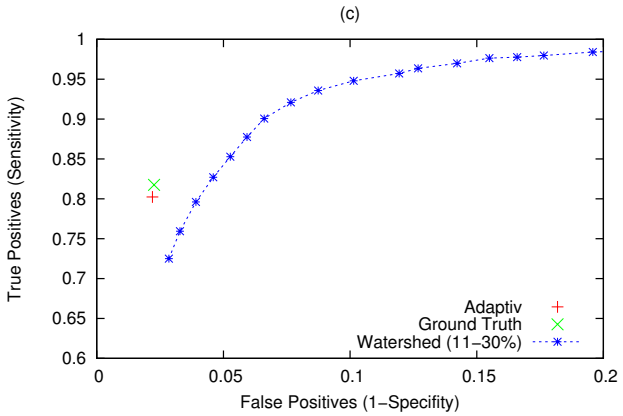


Fig. 4. The cross on the left marks the quality of our proposed algorithm. The curves represent the ROC plots for different watershed percentages ranging from 10 to 30% in different color spaces. It's obvious here, that our adaptive threhsold is superior to threhsolds selected by a general watershed percentage. This is independent from the choosen color space or the choosen percentages for the watershed.

1) The assumed **optimal** is achieved by the **ground truth**. For each image of our samples the optimal threshold was selected consulting the ground truth, such that threshold t is obtained by following condition:
$t = \underset{t}{argmin}\left(FP(t)^2 + (1 - TP(t))^2\right)$ or in words: choose that t, which generates FP- and FP-rate plotting closest to top left corner in the ROC curve. The meaning of $(FP(t))$ is here the false positive rate with respect to tested threshold t (TP accordingly).
2) A pure color based **watershed** algorithm, following [5], was implemented. In common a watershed is parametrized by the percentages for the target class, according to the whole histogram. So the histogram is added slot by slot starting from the lowest value. Once a percental threshold (compared to all pixels in the whole histogram) is reached the algorithm stops. This is like an application of a-priori knowledge about ratio of the mouth size to the size of the region of interest, where the histogram has been taken from. Our experiments included variations of the watershed with percentages from 10% to 30%. Also different promising color transformations were applied, to exclude the chance of inappropriate color model for watershed.

The results, compared to our proposed algorithm, are shown

in a ROC plot in Fig. 4. In a ROC plot the optimum results is the most top left. The one cross in the left marks the results of our algorithm. The average hit rate of 80% appears moderate, but is the results of some outliers (see Fig. 5). Compared to the watershed based algorithm the proposed method can eliminate the completely disturbed outliers with FP rates greater than 10% almost completely. As outlined in section II the primary aim is prevention from false positive lip pixels. Even it has in average the half FP rate, but maintaining the average quality of 80% hitrate (Table II). Our results are very close to the results achieved by choosing an optimal threhsold consulting the groundtruth of our dataset. This can be seen as proof, that our algorithm gets most out of the color information and working with a plain color-/histogram-based algorithm. Better results would need to add more features/knowledge to the system. The best results with watershed algorithms were achieved with percentages rates between 12% and 15% (so 13% and 15% have been choosen exemplary for table plots). However, this is no big surprise, since the average ratio of mouthpixels inside the images over all samples was 17.2% and an average optimal hitrate of 80%. To achieve the same true positive hit-rate as out adaptive the watershed algorithm produces almost twice false positives(see Table II).

V. Summary and Conclusion

We presented a new adaptive pure color based method for lip pixel segmentation. It performs better than conventional watershed based methods since it can basically adapt better to the various ratios of mouth to skin in the mouth ROI. To summarize we can say, that drawback of the watershed is it's strict dependence from the mouth size vs ROI ratio. Our method also utilizes this, but the refinement step can compensate errata here. The different illumination samples did not include completely disturbed scenarios like cross fading with full loss of color information. So this algorithm is however still based on "'good-natured"' illumination but can handle here different levels of illumination adaptively. We did explicitly not compete with hybrid systems, which might have better results and focused to color based methods only. Also this work should be considered as first step toward a hybrid mouth segmentation method including additional information than only color. In future work we will develop a method using color *and* edge information.

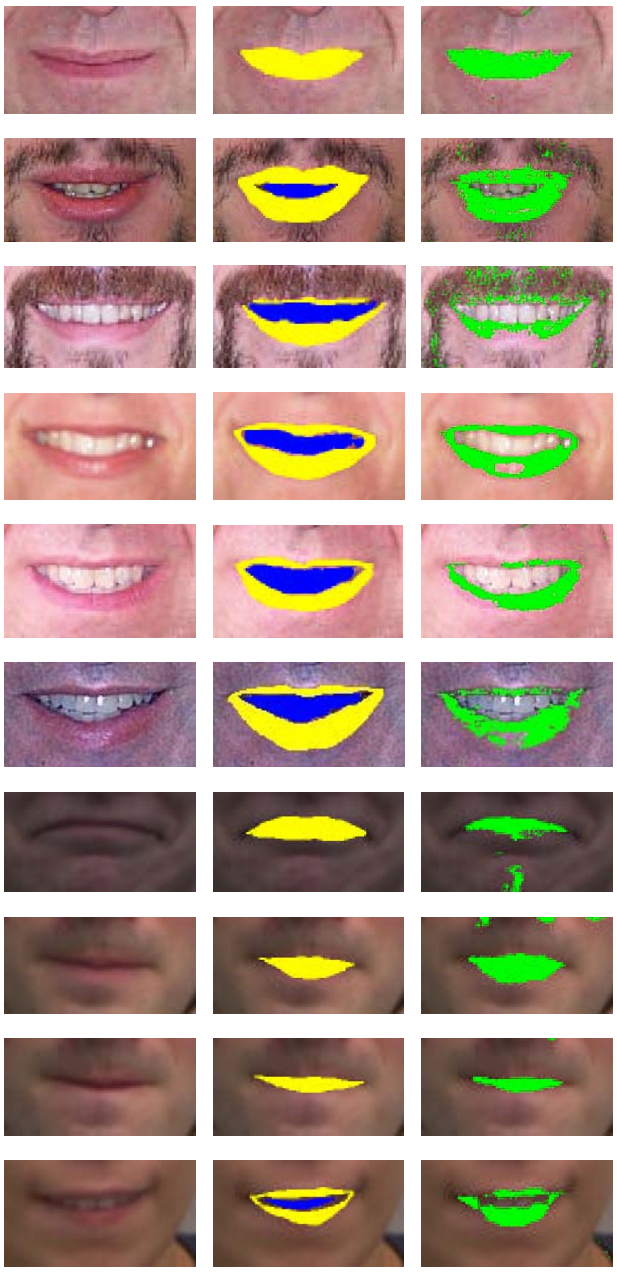


Fig. 6. Results of different illumination and mouth poses. Left: original, Center: groundtruth, Right: adaptive threhsold

Acknowledgment

This work was supported by DFG-Schmerzerkennung (FKZ: BR3705/1-1), DFG-Transregional Collaborative Research Centre SFB/TRR 62, and BMBF Bernstein-Group (FKZ: 01GQ0702), NIMITEK-C4 (LSA: XN3621E/1005M)

References

[1] A. Al-Hamadi, A. Panning, R. Niese, and B. Michaelis. A model-based image analysis method for extraction and tracking of facial features in video sequence. In *The 4th International Multi-conference on Computer Science and Information Technology CSIT 2006, Spo. by IEEE, Amman,Vol.3*, pages 499–509, 2006.

[2] S. Arca, P. Campadelli, and R. Lanzarotti. A face recognition system based on local feature analysis. In *Audio- and Video-Based Biometric Person Authentication*, pages 182–189, 2003.

[3] C. Bouvier, P.Y. Coulon, and X. Maldague. Unsupervised lips segmentation based on roi optimisation and parametric model. In *IEEE International Conference on Image Processing*, pages IV: 301–304, 2007.

[4] Jingying Chen, Bernard Tiddeman, and Gang Zhao. *Advances in Visual Computing*, volume 5359/2008 of *Lecture Notes in Computer Science*, chapter Real-Time Lip Contour Extraction and Tracking Using an Improved Active Contour Model, pages 236–245. Springer Berlin / Heidelberg, 2008.

[5] P. Cisar and Zelezny M. Using of lip-reading for speech recognition in noisy environments. In *Speech Processing*, pages 137–142, Prague, 2004. Academy of Sciences of the Czech Republic.

[6] N. Eveno, A. Caplier, and P.Y. Coulon. Accurate and quasi-automatic lip tracking. *Circuits and Systems for Video Technology*, 14(5):706–715, May 2004.

[7] Erhan AliRiza Ince and Syed Amjad Ali. An adept segmentation algorithm and its application to the extraction of local regions containing fiducial points. In *ISCIS*, pages 553–562, 2006.
[8] K.S. Jang, S. Han, I. Lee, and Y.W. Woo. Lip localization based on active shape model and gaussian mixture model. In *Pacific-Rim Symposium on Image and Video Technology*, pages 1049–1058, Hsinchu , TAIWAN, 2006.
[9] J.Y. Kim, S.Y. Na, and R. Cole. Lip detection using confidence-based adaptive thresholding. In *International Symposium on Visual Computing*, pages I: 731–740, 2006.
[10] S.H. Leung, S.L. Wang, and W.H. Lau. Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Transaction on Image Processing*, 13(1):51–62, January 2004.
[11] Trent W. Lewis and David M.W. Powers. Lip feature extraction using red exclusion. In Peter Eades and Jesse Jin, editors, *Selected papers from Pan-Sydney Area Workshop on Visual Information Processing (VIP2000)*, volume 2 of *CRPIT*, pages 61–67, Sydney, Australia, 2001. ACS.
[12] D. Nguyen, D. Halupka, P. Aarabi, and A. Sheikholeslami. Real-time face detection and lip feature extraction using field-programmable gate arrays. *IEEE Trans. Systems, Man and Cybernetics, SMC-B*, 36(4):902–912, August 2006.
[13] California Institute of Technology. Faces 1999 (front). http://www.vision.caltech.edu/archive.html, 1999.
[14] A. Panning, A. Al-Hamadi, R. Niese, and B. Michaelis. Facial expression recognition based on haar-like feature detection. *Pattern Recognition and Image Analysis*, 18(3):447–452, 2008.
[15] Paul Viola and Michael Jones. Robust real-time object detection. *Second international workshop on statistical and computational theories of vision modeling, learning, computing, and sampling*, 2001.