Automatic Segmentation of Lip Images Based on Markov Random Field

Meng Li
Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

Abstract

This paper addresses the problem of lip segmentation in color space that is a crucial issue to a successful lip-reading system. We present a new segmentation approach to lip contour extraction by taking account of the maximum a posterior - Markov random field (MAP-MRF) framework. We first examine various color models and select a simple color transform derived from LUX and 1976 CIELAB color space as an effective descriptor to characterize the lip region by its discriminative properties. Thus, the initial label set with respect to lip and skin region is available. Based upon the identified lip area, we further refine the lip region using both color and label information, as those two are combined within a Markov random field (MRF) framework. Finally, we extract the lip contour via convex hull algorithm with the prior knowledge of the mouth shape. Experiments show the efficacy of the proposed approach in comparison with the existing lip segmentation methods.

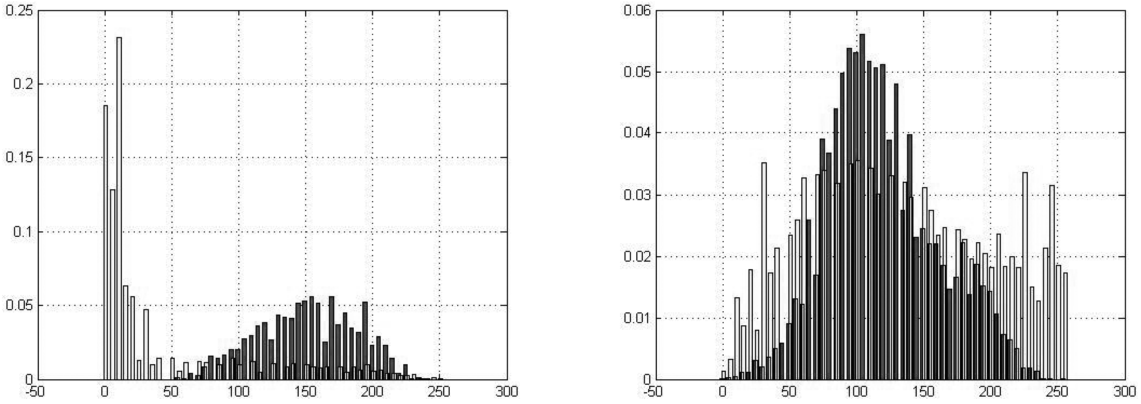
1. Introduction

Automatically segmenting out person's lip from face image is an active research area nowadays for its wide range of possible applications such as lip-reading for disabled people, audio-visual speech recognition in noisy environment, face detection, biometric person identification, lip synchronisation, facial expression recognition and so forth [1, 2, 3, 4, 5, 6].

Thus far, various segmentation techniques have been proposed. In general, these methods based on color space rather than gray level since color image can provide more useful clues. In [7, 8, 9, 10], they made an analysis of the original color space, and transformed its representation into a new space by intensity difference between lips and background. Clustering with color feature is an attractive preprocessing method as well. [11, 12, 13] utilize clustering based methods to conduct segmentation and achieve considerable high accuracy. Meanwhile, wavelet is another effective solution. [14] proposed a segmentation method which used wavelet multi-scale edge detection across the C_3 component of the discrete Hartley transform (DHT). Moreover, [10] employed Gaussian mixture model (GMM) to estimate the membership map of lips computed from the skin color distribution. Nevertheless, these methods only focus on color feature regardless of spatial characters which also convey important clues for segmentation. Thus, corresponding results seemed fragmented and so easily affected by noise, some of which are hard to overcome in postprocessing.

Nowadays, in image segmentation field, the assumption that "physical properties in a neighborhood of space present some coherence and generally do not change abruptly"[15] is utilized widely so as to overcome the segmentation errors arised from the intensity non-uniformity and local perturbations. In [16], a spatial fuzzy clustering algorithm was proposed. This method is able to take into account both the distributions of data in feature space (derived from 1976 CIELAB and 1976 CIELUV color space) and the spatial interactions between neighboring pixels during clustering. Furthermore, [5, 17, 18, 19] and so forth made advantage of the Markov random field (MRF) model to represent the spatial constraint.

Although empirical studies have shown their success, practical lip segmentation is a non-trivial task. The main difficulty lies in robustness and automation. In real world, the illumination condition and the complexion of speaker are multifarious. Thus, robustness is an important benchmark for a lip segmentation method. However, it is challenging for a color transform to achieve stable performance in different situations. It is because that the fluctuation range of hue of lip and skin region is considerable large under different illumination conditions, not to mention difficulties brought by testers (speakers) with totally different complexion. On the other hand, from the practical viewpoint, automation is also an important factor as well. Nevertheless, lots of common methods can not satisfy this requirement. For example, in the lip segmentation task, the fuzzy



(a)

(b)

Figure 1. The gray-level probability distribution of I_{a^*} (dark color) and I_U (light color) in (a) lip region, and (b) skin region. The x-axis represents the gray-level value, and the y-axis represents the probability.

clustering based method [16] requires pre-assignment of the number of clusters, and the MRF based method [5] heuristically fixes the cluster number as 3. However, in application, the existence of moustache, the visibility of teeth and tongue generally require that the number of clusters is selected adaptively. Unfortunately, the performances of these methods depend on the knowledge of cluster number significantly.

In this paper, we will present a new method for the robust automatic segmentation of lip images provided that the part between nostril and chin has been available. Firstly, the proposed method employs a color transform derived from the LUX and 1976 CIELAB color space to obtain a lip segment via the distinction between the lip and skin. The result can be used to 1) initialize the MRF label map, and 2) estimate the parameters of likelihood energy function in MRF model. Secondly, a MRF model is established on the 4-neighborhood system. Thirdly, a deterministic algorithm called iterated conditional modes (ICM) is performed to minimize the cost function and obtain a robust labeling of lip and background, respectively. Finally, given the prior knowledge of human mouth shape, noise suppressing is performed based on morphological operation and taken as postprocessing procedure, together with the convex hull algorithm. Experiments have shown the efficacy of the proposed approach in comparison with the existing lip segmentation methods.

The remainder of this paper is organized as follows. Section 2 describe the pre-processing (color transform), lip segmentation (MAP-MRF classification) and post-processing (lip boundary extraction) in turn. In Section 3, we will conduct the experiment to empirically compare our approaches with some existing methods. Finally, Section 4 draws a conclusion.

2. Method

2.1 Color Transform

It is desirable to work in a color space, in which the lip color (i.e. relative red) out of the others can be highlighted. Since the value of a^* channel in 1976 CIELAB color space can determine the color component between magenta and green, i.e. the small values indicate green while the large indicate magenta. We therefore transform the source image into 1976 CIELAB color space and employ the histogram equalization[20] to map the a^* component into the range of [0, 255], denoted as I_{a^*}. Furthermore,we utilize Eq.(1) as proposed in [5] to convert the source image to the range of $[0, 255]$ with equalization, denoted as U:

$$U = \begin{cases} 256 \times \frac{G}{R} \; ifR > G \\ \\ 255 \qquad otherwise. \end{cases} \tag{1}$$

Then, we also map U component into the range of [0, 255] denoted as I_U.

We select 100 lip images from 4 databases randomly, label the lip region manually, and obtain I_{a^*} and I_U from each image. Then, calculate the histogram (normalized into [0, 1]) of data set composed by the intensity of pixels belong to I_{a*} and I_U which fall into lip and skin region, respectively (see Fig. 1). We further calculate the mean values of the four distribution denoted as $\mu_{lip}^{a^*}$, $\mu_{skin}^{a^*}$, μ_{lip}^{U} and μ_{skin}^{U}. In this experiment, $\left|\mu_{lip}^{a^*} - \mu_{lip}^{U}\right| = 110.88$ is far more larger than $\left|\mu_{skin}^{a^*} - \mu_{skin}^{U}\right| = 15.64$.

Based on the analysis above, we believe that the difference between I_{a^*} and I_U in lip region is significant but inconspicuous in skin region. Thus, we can employ the following equation to get the lip segment roughly.

$$I_{sub} = I_{a^*} - I_U.\text{[1]} \tag{2}$$

Subsequently, we establish a Gaussian model for I_{sub} based on the gray-level value of each non-zero pixel with the mean $\hat{\mu}_{sub}$ and the standard deviation $\hat{\sigma}_{sub}$. The candidate lip segment can be obtained by

$$\tilde{I}^1_{candidate} = \begin{cases} 0 \ if \quad I_{sub} \leq \hat{\mu}_{sub} - 2\hat{\sigma}_{sub}, \\ \\ 1 \ otherwise. \end{cases} \tag{3}$$

Moreover, from a different aspect, we can calculate another representation of candidate lip segment. Firstly, the following equation is employed to get $U^{'}$:

$$U^{'} = \begin{cases} \frac{a^*}{G} \ if a^* > G \\ \\ 0 \quad otherwise. \end{cases} \tag{4}$$

where a^* denotes the equalized a^* component in 1976 CIELAB color space, and G denotes the equalized G component in RGB color space.

Then, a gray-level threshold selection method proposed in [21] is utilized to transform $U^{'}$ into a binary image denoted as $I^2_{candidate}$.

We assume the lip region is not connected to the border of input image. Thus, the morphological reconstruction based method proposed in [22] is performed to suppress noisy structures. For this operation, the mask is $\tilde{I}^1_{candidate}$, and the marker is an image which is all zero except along the border. The output image is denoted by $I^1_{candidate}$. Therefore, we use $I^1_{candidate} \cap I^2_{candidate}$ as lip segment, denoted by I_{seg}.

2.2 MAP-MRF Classification

In order to build a probability map of lip and skin region, each pixel s will be attributed a label l_s from the set Λ reflects its feature class. In this paper, $\Lambda = \{0, 1\}$. For s belong to lip class, $l_s = 1$, and $l_s = 0$ otherwise. An realization of a set of labels is considered a configuration defined on a 2-D rectangular regular lattice $\mathscr{S} = \{i | 1 \leq i \leq N\}$ where N is the number of pixels in the input image. An observed image in modified HSV color space $c = \{c_i | i \in \mathscr{S}\}$, and a configuration $l = \{l_i | i \in \mathscr{S}\}$ are instances of each random field. The form of c_i can refer to Eq. (5):

$$c_i = \{(H_i{\cdot}cos(2\pi{\cdot}S_i), H_i{\cdot}sin(2\pi{\cdot}S_i))^T | i \in \mathscr{S}\} \tag{5}$$

where H_i and S_i denote the H and S component value of pixel i.

A prior model should properly define the interactions between the labeled pixels. MRF are well suited for that purpose. Let us consider the spatial 4-neighborhood structure $\mathscr{N}_s$. The label field is supposed to verify the main MRF property related to that neighborhood, which means the label l_s of the current pixel s depends only on the labels l_r of its neighbors $r \in \mathscr{N}_s$. Assuming our scene is a piecewise constant surface and the model is spatial homogeneous, the prior probability can be written as follows with the independent assumption in terms of the MRF-Gibbs equivalence:

$$p(l) = \prod_{i \in \mathscr{S}} \frac{e^{-V(i)}}{\sum_{j \in \mathscr{S}} e^{-V(j)}} \tag{6}$$

Based on Potts model, the prior energy can be defined as:

$$V(l) = \sum_{i \in \mathscr{S}} V(i) = \sum_{i \in \mathscr{S}} \sum_{i' \in \mathscr{N}(i)} (1 - \delta(l_i - l_{i'})) \tag{7}$$

where $\delta(\cdot)$ is the Kronecker delta function.

To establish the likelihood energy function, we further assume that the intensity c_i for pixels with the same label follows the same bivariate Gaussian distribution. The likelihood probability can be written as:

$$p(c|l) = \prod_{i \in \mathscr{S}} \frac{1}{2\pi\sqrt{|\hat{\Sigma}^{l_i}|}} \cdot exp(-\frac{(c_i - \hat{\mu}^{l_i})(c_i - \hat{\mu}^{l_i})^T}{2\hat{\Sigma}^{l_i}}) \tag{8}$$

Given the label $l_i = \lambda \in \Lambda$, parameter estimation is made as follows:

$$\hat{\mu}^{\lambda} = \frac{\sum_{j=1}^{M^{\lambda}} c_j^{\lambda}}{M^{\lambda}}, \tag{9}$$

$$\hat{\Sigma}^{\lambda} = \frac{1}{M^{\lambda} - 1} \sum_{j=1}^{M^{\lambda}} (c_j^{\lambda} - \hat{\mu}^{\lambda})(c_j^{\lambda} - \hat{\mu}^{\lambda})^T, \tag{10}$$

[1] In this paper, all equations are employed in positive area. That is, as long as a result is negative, it will be set at 0 automatically.

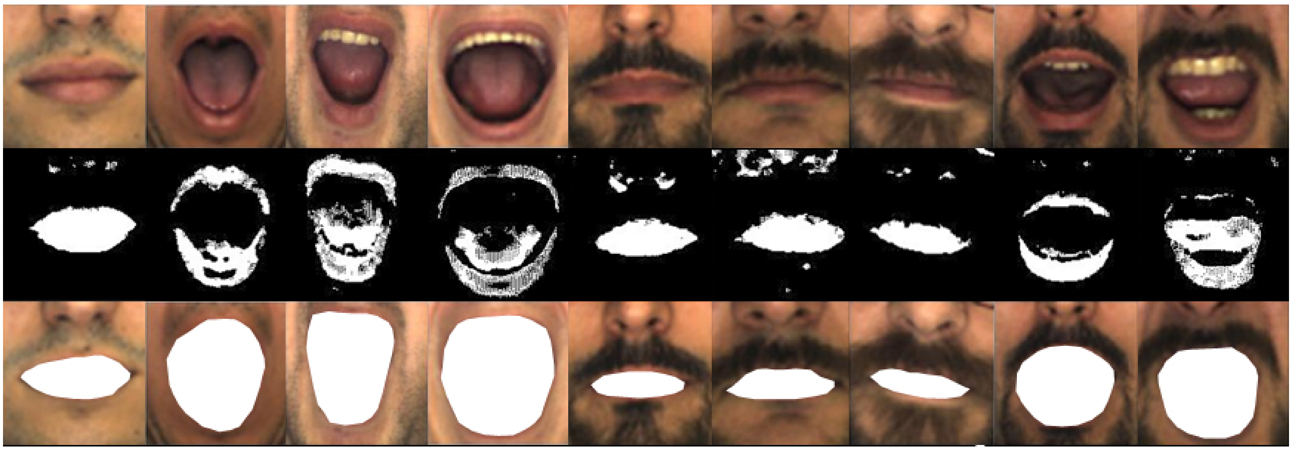


Figure 2. Top row: some examples of source input images, in which several representative samples – normal situation, mouth opening (teeth existed), mustache existed, etc., are involved. Middle row: corresponding segmentation results obtained by ICM. Bottom row: final segmentation results.

where c_j^λ is obtained by Equ. (5) at the jth pixel with label λ in the input image, and M^λ denotes the number of these pixels.

Thus, the likelihood energy can be defined as:

$$V(c|l) = \sum_{i \in \mathscr{S}} (c_i - \hat{\mu}^{l_i}) {\Sigma^{l_i}}^{-1} (c_i - \hat{\mu}^{l_i})^T \qquad (11)$$

Based on MAP framework, the label can be selected for each pixel through the following optimal function:

$$l = \arg \max_{l_i \in \Lambda} p(l) p(c|l) = \arg \min_{l_i \in \Lambda} (\alpha V(l) + V(c|l)) \qquad (12)$$

where α is a positive weight which can be used to balance the dimensions of the two terms, say, $V(l)$ and $V(c|l)$. And the segmentation result won't be sensitive to this value. In our experiments, α was set to 2.

We choose the iterative deterministic algorithm ICM (Iterated Conditional Modes) to compute the minimum energy at each site for sake of its low computation cost. But the problem is it may converge towards local minima. In our experiments, however, a stable solution is always found in practice after a few iterations on the field (less than 5). The initial label set is derived from I_{seg}. The relative variation of global energy is used as termination condition: $\Delta E(l)/E(l) < \epsilon$ (typically, $\epsilon = 0.05$) where $E(l) = \alpha V(l) + V(c|l)$. Some segmentation results obtained by ICM can be found in middle row, Figure. 2.

2.3 Boundary Extraction

The result obtained by ICM can be considered as a binary image (the pixels with label 1 are foreground, and the others are background). Then, we suppress the boundary connected structures [22] in it and denoted as B_{RT}. The biggest continued foreground block is marked by B_{lip_1}. In the case of mouth closing, B_{lip_1} can represent the whole lip region accurately. However, in most cases of mouth opening, the blocks corresponding to upper and lower lips are usually separate. It is hard to extract the whole lip region via selecting the biggest connected block. Thus, some refinements are needed.

Considering the primary reason for disconnection between upper and lower lip is that the teeth and tongue are eliminated in the above steps. Hence, we utilize the following equation

$$I_{TTM} = I_U - I_{a^*} \qquad (13)$$

to obtain the region covering the teeth, tongue and some parts of oral cavity approximately.

We further transform I_{TTM} into a binary image denoted as B_{TTM} by the threshold selection method. Then, the morphological closing is employed to $B_{RT} \cup B_{TTM}$ by performing a 5×5 structuring element operation. We select the biggest foreground block denoted as B_{lip_2} in the closing operation result. Hence, the binary image $B_{lip_1} \cup B_{lip_2}$ can represent the whole lip region even in the case of mouth opening. Furthermore, we utilize the morphological opening with 3×3 structuring element so as to make the edge more smooth. The result is denoted as B_{lip}.

Finally, the quickhull algorithm proposed in [23] is employed to draw the contour of lip (e.g. see bottom row, Figure 2).

The proposed segmentation method can be summarized as follows:

```
input the RGB source image;

compute the binary $I_{seg}$ as the initial segmentation;

initialize the number of iteration $t = 0$;

update $l^t = \{l_i|i \in \mathscr{S}\}$;

do
{
        update the variables: $\hat{\mu}^\lambda, \hat{\Sigma}^\lambda$;

        get a label set which can reach the minimum of
        Equ.(12) via ICM;

        perform the morphological filters and convex
        hull method to refine the label set;

        update $l^{t+1} = \{l_i|i \in \mathscr{S}\}$ by the label set;

        $t = t + 1$;

}until($\frac{d(l^t,l^{t-1})}{size(l^t)} \leq \xi$ )

output the final result;
```

where $d(l^t, l^{t-1})$ denotes the Hamming distance between the label set obtained by the tth and $(t-1)$th iteration, respectively; $size(l)$ denotes the number of elements belong to set l; ξ is termination condition which can get 0.005 heuristically.

3. Experiment Results

To demonstrate the performance of the proposed approach in comparison with the existing methods denoted as: Liew03 proposed in [16], and Lievin04 in [5]. We utilized four databases to test the accuracy and robustness in different capture environments: (1) AR face database (126 people with 26 images for each) [24], (2) CVL face database (114 persons with 7 images for each) [25], (3) GTAV face database (44 persons with 27 images for each), (4) a database established by ourselves, including 19 persons (10 male and 9 female) with 15 pictures per person corresponding to different mouth shapes. We randomly selected 800 images in total (400 images from AR database, 200 images from CVL database, 100 images from GTAV database, 100 images from our database) and manually segmented the lip to serve as the ground truth. Moreover, in AR database, the images with the feature number 11, 12, 13, 24, 25, 26 (wearing scarf which covers the whole mouth) were not used for this experiment.

Algorithm	Liew03	Lievin04	Proposed
average OL, %	78.73	87.46	**92.48**
average SE, %	35.15	25.01	**7.10**

Table 1. The segmentation results across the four databases.

Two measures defined in [16] are used to evaluate the performance of the algorithms. The first measure determines the percentage of overlap (OL) between the segmented lip region A_1 and the ground truth A_2:

$$OL = \frac{2(A_1 \cap A_2)}{A_1 + A_2} \times 100\%. \qquad (14)$$

The second measure is the segmentation error (SE) defined as

$$SE = \frac{OLE + ILE}{2 \times TL} \times 100\%, \qquad (15)$$

where OLE is the number of non-lip pixels classified as lip pixels (i.e. outer lip error), ILE is the number of lip-pixels classified as non-lip ones (inner lip error), and TL denotes the number of lip-pixels in the ground truth.

Table 1 shows the segmentation results on the four different databases. It can be seen that the proposed method outperforms the Liew03 and Lievin04 in both of the two measurements. Specifically, for the embarrassed problem in lip segmentation – the moustache existed cases, say, the tester in AR database with number 4, 5, 18, 26, 31, 38 and so forth, the OL and SE our method proposed is 91.15% and 10.14%, respectively.

4. Conclusion

In this paper, we have proposed a new approach to automatic lip segmentation via color transform and the MAP-MRF framwork. This approach features the high accuracy of lip segmentation and robust performance against diverse capture environment and different skin color (white and yellow). Experiments have shown the promising result of the proposed approach in comparison with the existing methods.

References

[1] T. Chen and R.R. Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–851, 1998.

[2] I. Matthews, T.F. Cootes, and J.A. Bangham. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:198–213, 2002.

[3] W. Gao, Y. Chen, R. Wang, S. Shang, and D. Jiang. Learning and synthesizing mpeg-4 compatible 3-d face animation from video sequence. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(11):1119–1128, 2003.

[4] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[5] M. Lievin and F. Luthon. Nonlinear color space and spatiotemporal mrf for hierarchical segmentation of face features in video. *IEEE Transactions on Image Processing*, 13(1):63–71, 2004.

[6] H.E. Cetingul, Y. Yemez, E. Erzin, and A.M. Tekalp. Discriminative analysis of lip motion features for speaker identification and speech-reading. *IEEE Transaction on Image Processing*, 15(10):2879–2891, 2006.

[7] N. Eveno, A. Caplier, and P.Y. Coulon. A new color transformation for lips segmentation. In *Proceedings of the 4th IEEE Workshop on Multimedia Signal Processing*, pages 3–8, Cannes, France, 2000.

[8] T. Wark, S. Sridharan, and V. Chandran. An approach to statistical lip modelling for speaker identification via chromatic feature extraction. In *Proceedings of IEEE International Conference on Pattern Recognition*, pages 123–125, Brisbane, Australia, 1998.

[9] N. Eveno, A. Caplier, and P.-Y. Coulon. A parametric model for realistic lip segmentation. In *Proceedings of International Conference on Control, Automation, Robotics and Vision*, pages 1426–1431, Singapore, 2002.

[10] C. Bouvier, P.-Y. Coulon, and X. Maldague. Unsupervised lips segmentation based on roi optimisation and parametric model. In *Proceedings of IEEE International Conference on Image Processing*, pages 301–304, San Antonio, USA, 2007.

[11] M. SADEGHI, J. KITTLER, and K. MESSER. Spatial clustering of pixels in the mouth area of face images. In *Proceedings of IEEE Internatioal Conference on Image Analysis and Processing*, pages 36–41, Palermo, Italy, 2001.

[12] S.L. Wang, W.H. Lau, S.H. Leung, and A.W.C. Liew. Lip segmentation with the presence of beards. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 529–532, 2004.

[13] W.C. Liew S.H. Leung S.L. Wang, W.H. Lau. Robust lip region segmentation for lip images with complex background. *Pattern Recognition*, 40(12):3481–3491, 2007.

[14] Y.P. Guan. Automatic extraction of lips based on multi-scale wavelet edge detection. *Computer Vision, IET*, 2(1):23–33, 2008.

[15] S.Z. Li. *Markov Random Field Modeling in Image Analysis (Third Edition)*. Springer, 2009.

[16] Alan W.C. Liew, S.H. Leung, and W.H. Lau. Segmentation of color lip images by spatial fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 11(4):542–549, 2003.

[17] Albert C.S. Chung. A segmentation model using compound markov random fields based on a boundary model. *IEEE Transactions on Image Processing*, (1):241–252, 2007.

[18] H. Gribben, P. Miller, G.G Hanna, K.J Carson, and A.R. Hounsell. Map-mrf segmentation of lung tumours in pet/ct images. In *Proceedings of 6th International Symposium on Biomedical Imaging: From Nano to Macro*, pages 290 – 293, Boston, USA, 2009.

[19] B. Scherrer, F. Forbes, C. Garbay, and M. Dojat. Distributed local mrf models for tissue and structure brain segmentation. *IEEE Transactions on Medical Imaging*, (8):1278–1295, 2009.

[20] C. Gonzalez. *Digital Image Processing (Third Edition)*. Pearson Eduation Eduaction, Inc., 2008.

[21] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[22] P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer, 1999.

[23] C.B. Barber, D.P. Dobkin, and H.T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[24] A.M. Martinez and R. Benavente. The ar face database. *CVC Technical Report No.24*, June 1998.

[25] F. Solina, P. Peer, B. Batagelj, S. Juvan, and J. Kovac. Color-based face detection in the '15 seconds of fame' art installation. In *Proceedings of Conference on Computer Vision / Computer Graphics Collaboration for Model-based Imaging, Rendering, Image Analysis and Graphical Special Effects*, pages 38–47, Versailles, France, 2003.