8-8-2014

On study of lip segmentation in color space

Meng Li
Hong Kong Baptist University

Follow this and additional works at: https://repository.hkbu.edu.hk/etd_oa

Recommended Citation
Li, Meng, "On study of lip segmentation in color space" (2014). Open Access Theses and Dissertations. 42.
https://repository.hkbu.edu.hk/etd_oa/42

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at HKBU Institutional Repository. It has been accepted for inclusion in Open Access Theses and Dissertations by an authorized administrator of HKBU Institutional Repository. For more information, please contact repository@hkbu.edu.hk.
On Study of Lip Segmentation in Color Space

LI Meng

A thesis submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Principal Supervisor: Prof. CHEUNG Yiu Ming

Hong Kong Baptist University

August 2014
Declaration

I hereby declare that this thesis has been composed by myself under the guidance of my principal supervisor Prof. Cheung Yiu-ming and co-supervisor Dr. Chu Xiao-wen. The thesis has not previously included in any thesis, dissertation or report submitted to any institution for a degree, diploma or other qualification. All sources of information have been acknowledged by means of references to the relevant publications.

Signature:________________________

August 2014
Abstract

This thesis mainly addresses two issues: 1) to investigate how to perform the lip segmentation without knowing the true number of segments in advance, and 2) to investigate how to select the local optimal observation scale for each structure from the viewpoint of lip segmentation effectively.

Regarding the first issue, two number of predefined segments independent lip segmentation methods are proposed. In the first one, a multi-layer model is built up, in which each layer corresponds to one segment cluster. Subsequently, a Markov random field (MRF) derived from this model is obtained such that the segmentation problem is formulated as a labeling optimization problem under the maximum a posteriori-Markov random field (MAP-MRF) framework. Suppose the pre-assigned number of segments may over-estimate the ground truth, whereby leading to the over-segmentation. An iterative algorithm capable of performing segment clusters and over-segmentation elimination simultaneously is presented. Based upon this algorithm, a lip segmentation scheme is proposed, featuring the robust performance to the estimate of the number of segment clusters. In the second method, a fuzzy clustering objective function which is a variant of the partition entropy (PE) and implemented using Havrda-Charvat’s structural α-entropy is presented. This objective function features that the coincident cluster centroids in pattern space can be equivalently substituted by one centroid with the function value unchanged. The minimum of the proposed objective function can be reached provided that: (1) the number of positions occupied by cluster centroids in pattern space is equal to the truth cluster number, and (2) these positions are coincident with the optimal cluster...
centroids obtained under PE criterion. In the implementation, the clusters provided that the number of clusters is greater than or equal to the ground truth are randomly initialized. Then, an iterative algorithm is utilized to minimize the proposed objective function. The initial over-partition will be gradually faded out with the redundant centroids superposed over the convergence of the algorithm.

For the second issue, an MRF based method with taking local scale variation into account to deal with the lip segmentation problem is proposed. Supposing each pixel of the target image has an optimal local scale from the segmentation viewpoint, the lip segmentation problem can be treated as a combination of observation scale selection and observed data classification. Accordingly, a multi-scale MRF model is proposed to represent the membership map of each input pixel to a specific segment and local-scale map simultaneously. The optimal scale map and the corresponding segmentation result are obtained by minimizing the objective function via an iterative algorithm.

Finally, based upon the three proposed methods, some lip segmentation experiments are conducted, respectively. The results show the efficacy of the proposed methods in comparison with the existing counterparts.
Acknowledgements

This thesis would not have been possible without the advice and help of several individuals who contributed and extended their valuable assistances in the completion of this study.

First and foremost, I offer my sincerest gratitude to my principal supervisor Prof. Cheung Yiu-ming, for the patient instruction he has provided throughout my study period, and for his continual support on both academic and personal problems. Indeed, his guidance and advice have had a major positive impact on my development as a scientific researcher and as an individual.

I am very grateful to Dr. Zhang Hui, the CEO of Shenzhen CloudRiver Information Technologies Co. Ltd, for the invaluable suggestions, for the stimulating discussions, and for the sleepless nights we were working together. Thanks are also due to my fellow lab mates Dr. Liu Xin and Dr. Jia Hong for their insightful advices about my research work.

I thank my friends in Harbin Institute of Technology: Dr. Li Qiong, Dr. Huang Fu-rong, Dr. Han Qi, Dr. Jiao Yu-hua, Dr. Chen Yang and Dr. Li Jun. In particular, I appreciate Prof. Niu Xia-mu for enlightening me the first glance of research.

I am also deeply indebted to my friends Ms. Chai Yue and Ms. Shi Yan-bing for being the constant source of encouragement. It means a lot to me.

Last but not least, I express my special gratitude to my parents, Li Li-ping and Lv Min, for giving me all the dedication and understanding a child could want. I love you.
# Table of Contents

Declaration .................................................. i
Abstract ..................................................... ii
Acknowledgements ........................................... iv
Table of Contents ............................................. v
List of Tables ................................................ ix
List of Figures ............................................... xi

## Chapter 1 Introduction ..................................... 1
  1.1 Background and Motivation ............................ 1
  1.2 Problem Statement and Contributions ................. 5
    1.2.1 Segment Number Dependency Problem ............... 6
    1.2.2 Local Scale Determination Problem ................. 8
    1.2.3 Main Contributions of This Thesis ................. 10
  1.3 Organization of This Thesis .......................... 11

## Chapter 2 Review on Related Works, Used Models and Public Benchmark Databases .............................. 12
  2.1 Review on Lip Visual Feature Extraction ............ 12
    2.1.1 Color Space Selection ............................ 13
    2.1.2 Lip Segmentation ................................. 16
Chapter 3 MAP-MRF Based Lip Segmentation with Unknown Segment Number

3.1 Introduction .............................................. 32
3.2 Multi-layer Model based Image Segmentation ................. 33
   3.2.1 Multi-Layer Model ................................. 33
   3.2.2 Rival Penalty Iterative Algorithm ................. 37
3.3 Lip Boundary Extraction and Tracking .......................... 42
   3.3.1 Observation Space ............................... 42
   3.3.2 Segmentation and Binarization .................... 43
   3.3.3 Tracking ........................................ 44
   3.3.4 Postprocessing ................................. 46
3.4 Experimental Results .................................... 48
   3.4.1 Database and Initialization ...................... 48
   3.4.2 Experiment 1 ................................... 50
   3.4.3 Experiment 2 ................................... 51
   3.4.4 Experiment 3 ................................... 53
   3.4.5 Experiment 4 ................................... 56
   3.4.6 Discussion ..................................... 56
3.5 Summary .................................................. 59

Chapter 4 Fuzzy Clustering Based Lip Segmentation with Unknown Segment Number

4.1 Introduction .............................................. 60
4.2 Fuzzy Clustering With Unknown Cluster Number ............ 61
Chapter A 117
A.1 Derivation of Eq.(3.2.17) 117
A.2 Proof of Lemma 4.2.2 118
A.3 Validation of Lemma 4.2.2 123
A.4 Derivation of Eq.(5.2.7) 123

Chapter B 126
B.1 Publication Record 126

Curriculum Vitae 128
## List of Tables

3.1 Overlap rate of segmented lips over $m$ in the four groups of data, in which some cells are marked in “-” because the experiments with $\hat{m} < m^*$ are not available. .................................................. 51

3.2 Segment error of segmented lips over $\hat{m}$ in the four groups of data, in which some cells are marked in “-” because the experiments with $\hat{m} < m^*$ are not available. .................................................. 52

3.3 Average overlap and segmentation error for the images from AR, CVL and GTAV databases obtained by Wang07 and the proposed method, respectively. .......................................................... 56

3.4 Average overlap and segmentation error for the images from Vid-TIMIT databases obtained by Lievin04 and the proposed method, respectively. .......................................................... 57

4.1 Overlap of segmented lips with the ground truth. ......................... 75

4.2 Segmentation error of segmented lip. .......................................... 76

4.3 Average overlap and segmentation error for the images from AR and CVL databases obtained by Wang07 and the proposed method. .... 77

5.1 Average overlap and segmentation error for the images from AR, CVL, GTAV and VidTIMIT databases obtained by Lievin04, Liang06, Wang07 and the proposed method, respectively. ............... 90

5.2 Mean value and standard deviation of processing times for the images from AR, CVL, GTAV and VidTIMIT databases obtained by Lievin04, Liang06, Wang07 and the proposed method, respectively. . 90
A.1 MSE between the numerical simulation result $\hat{H}_{\eta m}^{\text{min}}$ and ideal value

$$H_{\eta m}^{\text{min}} = \eta_m \text{ over } m.$$
List of Figures

1.1 The main process blocks of audio only speech recognition (ASR), audio-visual speech recognition (AVSR), and visual only speech recognition (VSR) systems. ................................. 2

1.2 The detailed work flow of the “top-down” feature extraction in AVSR system. ................................................................. 4

1.3 The composition of ROI for lip segmentation problem. The legends posed in left denote the expected targets, while the right ones describes the possible noises. ........................................... 5

1.4 (a) is the source image. (b), (c), (d) illustrate the corresponding lip segmentation results obtained by the same MRF based segmentation with \( \hat{m} = m^* \), \( \hat{m} < m^* \) and \( \hat{m} > m^* \), respectively, where \( m^* \) denotes the true segment number and \( \hat{m} \) is the pre-assigned segment number. 6

1.5 The first row is the source image clips. The second row illustrates the corresponding segmentation results obtained by one MRF based method with the static pre-assigned segment number. The first image is over-segmented, the second image is segmented accurately, the third and forth images are under-segmented. The third row shows the segmentation results obtained by the same method with adaptive true segment numbers. ..................................................... 7

1.6 The flow chart of a typical clustering algorithm with adaptive number of clusters. The blocks in the dashed frame illustrate the number of clusters determination. ........................................ 8
1.7 A sample of the segmentation results obtained by the same MRF based method under different scales. (a) is the source image; (b) is the segmentation result obtained on source image; (c) is the segmentation result based on the sub-sampling images with \( \frac{1}{2} \) source size; (d) is the segmentation result based on the sub-sampling images with \( \frac{1}{4} \) source size. For the convenience of comparison, all sub-sampling images are zoomed to the size of source image.

2.1 Illustrations of the ROI in different color channels: (a) source image, (b) gray-level, (c) H channel in HSV, (d) S channel in HSV, (e) U channel in CIELUV, (f) V channel in CIELUV, (g) \( a^* \) channel in CIELAB, (h) \( b^* \) channel in CIELAB, (i) Q channel, (j) \( \hat{H} \) channel, (k) U channel in LUX, (l) X channel in LUX. The black dashed line corresponds to skin pixels and gray solid line to lip pixels.

2.2 Lip and skin pixel histograms in different color spaces: (a) gray-level, (b) H channel in HSV, (c) S channel in HSV, (d) U channel in CIELUV, (e) V channel in CIELUV, (f) \( a^* \) channel in CIELAB, (g) \( b^* \) channel in CIELAB, (h) Q channel, (i) \( \hat{H} \) channel, (j) U channel in LUX, (k) X channel in LUX. The black dashed line corresponds to skin pixels and gray solid line to lip pixels.

2.3 Some samples of the coarse segmentation results obtained by (b) the thresholding based method, (c) the fuzzy clustering method, (d) the gradient based method and (e) the statistical distribution method. (a) is the given source image.

2.4 Two lip localization samples via projection method based on the different coarse segmentation results obtained by (a) gradient and (b) thresholding.

2.5 Illustration of the lip model fitting result obtained by the (a) cubic polynomial curve based model, (b) order alterable polynomial curve based model and (c) B-spline curve based model.
2.6 The clique set with respect to regular lattice of sites and second-order neighborhood system, where the sites and neighboring relationships are shown as dots and joining lines, respectively.

2.7 Some sample images from AR (top row), CVL (middle row) and G-TAV (bottom row) face databases. Only clipped the lip region to show the details of the interesting parts. The three databases provide various lip shape states, some of them are even covered by the moustache. Moreover, the hues and color temperatures of images from different databases are dissimilar which is caused by the diverse complexions and illumination conditions.

2.8 Some sample frames from the VidTIMIT database. Only clipped the lip region to show the details of the interesting parts. The image sequence in each row corresponds a specific video clip.

3.1 The architecture of the proposed model, in which there are \( \hat{N} \) layers, i.e. \( \hat{N} \) segment clusters. Each layer is formed by a 2-D random field with the site set \( S \) on regular lattice of the image as shown in the bottom. The value \( f_i^{(j)} \) denotes the probability of pixel at site \( i \) belonging to segment \( j \).

3.2 Illustration of (a) the first-order neighborhood system, and (b) the corresponding cliques utilized in the proposed method, where the black and gray circles denote the pixel \( i \) and the neighborhood system, respectively. Please note that there are only two cliques because this is an isotropic system.
3.3 (a) The initial position of cluster centroids marked by "*"; (b) and (c) illustrate how the proposed algorithm work intuitively as given an input data marked by black circle, where the arrows with real line show the penalty direction and the length are proportional to the penalty strength. The arrows with the dotted line show the moving direction of the winner. (d) is the desired result in which the redundant centroids are driven far away from the input data set so that the over segmentation caused by them can be eliminated. 

3.4 There are two random fields laid on the different layers $j$ and $j'$, where each site is represented by a circle. In each layer, site $i$ is denoted by the black circle, and the neighboring site set of $i$, i.e. $\mathcal{N}_i$, is denoted by the gray circles. An inter-layer membership degree system with respect to site $i$, layer $j$ and $j'$, i.e. $\mathcal{F}_{i}^{(j,j')}$, is composed of the values in all non-white circles.

3.5 (a) is the source lip image. (b)-(h) illustrate the initial membership $f_i^{(j)}$ in the seven different layers, i.e. $\hat{m} = 7$. These memberships are calculated based on the initial $\mu$ and $\Sigma$.

3.6 The segmentation results in the seven different layers, while the initial membership is shown in Fig. 3.5. The white pixels in (a)-(g) represent the foreground, i.e. the site sets: $S^{(1)}$, $S^{(2)}$, ..., and $S^{(7)}$.

3.7 The first row is some samples in the input image sequence. Since it is hard to find the different between the nearby two frames, the 1st, 5th, 10th, 15th, 20th, 25th and 30th frames are selected to illustrate the tracking result. The second row represents the corresponding tracking results.
3.8 (a) The lip patch used to estimate the mean of $x_i$s fallen into the true lip region; (b) The illustration of $f_{lip}$; (c) The result of border connected noisy structures suppressing; (d) The binary version of (c); and (e) The result of morphological opening. In (f), the dark gray ellipse is defined by the eigenvectors and eigenvalues of the covariance matrix of $M$. The light gray macro block is noise which should be masked out. (g) The final extraction result.  

3.9 Sample images from AR database are in: (a) Group 1, (b) Group 2, (c) Group 3, and (d) Group 4.  

3.10 (a) is a sample of the input image for Wang07 in Experiment 3, (b) and (c) are the corresponding segmentation results obtained by Wang07 and the proposed method, respectively.  

3.11 (a) is a sample of the input image in Experiment 3, (b) and (c) are the corresponding segmentation results obtained by Wang07 and the proposed method, respectively.  

3.12 A Snapshot of lip segmentation results, in which the first four rows are the results obtained with the image from Group 1-4, respectively. The images of the last four rows are from CVL and GTAV databases.  

3.13 (a) and (b) are the frequency of specific OL and SE of the selected images from AR database, while (c) and (d) are from CVL, and (e) and (f) are from GTAV.  

3.14 A sample of lip segmentation result obtained by the proposed method, in which the outer and inner of lip can be extracted smoothly.  

3.15 Two samples of tongue misclassification: (a) and (c) are the source image; (b) and (d) are the corresponding segmentation result. In the two samples, the tongue regions are classified into the lip segment.
3.16 Some parts of the lip in source image: (a) Abnormal image caused by the light reflection; (b) The segmentation result obtained by the proposed method, in which the color abnormal regions are missed out; (c) is the segmentation result obtained by Wang07, in which the missed parts are relative small because the prior shape information of lip is considered and integrated into the clustering.

4.1 The relationship between position and centroid, where the plane with grid represents the pattern space. As shown in Sub-figure (a), there are 6 centroids (denoted by circles) with the coordinate (2, 4), (2, 4), (2, 4), (5, 2), (5, 2), (7, 5), but they only occupy 3 positions (2, 4), (5, 2), (7, 5) as shown in Sub-figure (b).

4.2 (a) The initial positions of centroids marked by “*”; (b) The clustering result by using the FCM, which falls into the local minimum; (c) The clustering result obtained from the proposed approach, where all redundant centroids are coincided.

4.3 The segmentation (clustering) results with $m = 9$, where Sub-figure (a) is the source image. The white pixels in Sub-figures (b)-(g) represent the pixels fall into cluster 1 - 9, respectively.

4.4 The lip segmentation result after redundant cluster centroids have been merged.

4.5 (a) is the lip patch which is used to estimate the mean of $x_i$'s falling into the true lip region, (b) is the result of border connected structures clearing, (c) is the result of morphological filter (closing with $5 \times 5$ structuring element and opening with $3 \times 3$ structuring element). In (d), the shape of gray ellipse is defined by the eigenvectors and eigenvalues of the covariance matrix of $P$. The continued objects on the outside of this ellipse are viewed as noises and masked out. (e) is the final extraction result obtained via quickhull algorithm.
4.6 The final centroid number after the clustering performed in Fisher’s iris database by the proposed method with different $m$ and initial $C$.  

4.7 This figure shows the histogram of the error between the cluster centroids obtained by our method and classical FCM (the cluster number is pre-defined by 3).  

4.8 (a) and (b) illustrate the histograms of $OL$ and $SE$ of the selected images from AR database. (c) and (d) illustrate the histograms of $OLs$ and $SEs$ of the selected images from CVL database.  

5.1 An illustration of the normalized 3-layer pyramid, comprising $x^{(0)}$ (i.e. the source image $x$), $x^{(1)}$, and $x^{(2)}$ from bottom to top.  

5.2 Illustration of the 3-D first-order neighborhood system of a given site, in which the black circle is the given site and gray circles denote the corresponding neighborhood system.  

5.3 The work flow of the proposed method.  

5.4 (a)-(d) are the raw image samples from AR, CVL, GTAV and VidTIMIT database, respectively. (e)-(h) are the corresponding source image samples of our experiment.  

5.5 A Snapshot of lip segmentation results obtained by the proposed method, the images in (a),(b),(c) and (d) are from AR, CVL, GTAV and VidTIMIT database, respectively. In each sub-figure, the first row is the source images, and the second row shows the corresponding results.  

5.6 (a) and (b) are the frequency of specific $OL$ and $SE$ of the selected images from AR database, while (c) and (d) are from CVL, (e) and (f) are from GTAV , (g) and (h) are from VidTIMIT.  

5.7 (a) and (c) are clips of 026 − MVC − 007F and 038 − MVC − 007F in CVL, respectively. (b) and (d) are the corresponding segmentation results obtained by Lievin04 with $\gamma = 165$.  

xvii
5.8  (a) is a clip of $m - 004 - 8$ image in AR, (b), (c), (d) are the segmentation results obtained by Liang06 with segment number 3, 2 and 4, respectively.  

5.9  (a) and (c) are the clips of $m - 004 - 17$ in AR and 046 – MVC – 007F in CVL, (b) and (d) are the corresponding segmentation results obtained by the proposed methods in which some parts of teethridge and tongue are misclassified.

A.1  The functional relationship $h_m$ between $\eta_m$ and $H^\text{min}_{\eta_m}$ with (a) $m = 2$, (b) $m = 3$, (c) $m = 4$, (d) $m = 5$, (e) $m = 6$, and (f) $m = 7$, respectively, where the horizontal axis represents the value of $\delta H_m$, and the vertical axis represents the value of $H^\text{min}_{\eta_m}$.

A.2  The curve of $H^\text{min}_{\eta_8} = h_8(\eta_8)$. When $\eta_8 \in (0, k^{-1} \frac{k-1}{k}]$ with $k = 1, 2, \ldots, 7$, the corresponding curve segments are coincident with $h_k(\eta_8)$.
Chapter 1

Introduction

1.1 Background and Motivation

In 1976, paper [1] firstly presented a perceptual phenomenon named “McGurk effect” which demonstrates the interaction between hearing and vision in speech perception. This perceptual phenomenon suggested that speech perception is multi-modal involving information from more than one sensory modality and demonstrated the intimate relation between the audio and visual sensory modality. According to this effect, the useful information on speech content can be obtained through analyzing the subtle cue conveyed by lip movement of speakers. It is considered as the foundation of lip movement analyzing via computer, which is formally known as automatic lip-reading [2].

As the development of computer technology, in 1984, the first automatic lip-reading system was presented by Petajan [3, 4] who used the current technology of dynamic time-warping with visual features derived from mouth opening. From then on, lip-reading has received considerable attention from the community because of its potential attractive applications such as audio-visual speech recognition (AVSR), speaker identification, virtual lip animation, human expression recognition and so forth [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. In particular, AVSR is a technique that uses image processing capabilities in lip reading to aid speech recognition systems in recognizing indeterministic phones. Compared to the traditional audio only speech
recognition (ASR) system which ignores the visual information, paper [15] presents the key reasons why vision benefits human speech perception: it helps speaker (audio source) localization, it contains information that supplements the audio, and it provides complimentary information about the place of articulation (tongue, teeth, and lips) which can help disambiguate. Papers [16, 17, 18, 19] further demonstrate that the visual cue, namely the movement or shape of lip can improve the performances of traditional ASR system in noisy environments significantly. Thus, if ASR is degraded due to environmental noise, the visual information plays an important role for speech understanding. On the other hand, visual only speech recognition (VSR) is another important application in which the speech content can be acquired by analyzing the visual cue only. Although VSR is not a mature technology at present, it still attracts considerable attentions [20, 21, 22, 23, 24] because of its promising prospect. A recent survey of AVSR and VSR is available in paper [25]. Meanwhile, AVSR and VSR introduce several novel and challenging tasks compared to traditional ASR. Fig. 1.1 [26] illustrates the basic frameworks of AVSR, ASR and VSR system. As shown in this figure, AVSR and VSR system have a common process step: the visual front end (face detection, lip localization and visual feature extraction).

Figure 1.1: The main process blocks of audio only speech recognition (ASR), audio-visual speech recognition (AVSR), and visual only speech recognition (VSR) systems.
For the first two components of visual front end, namely the face detection and lip localization, there are a great variety of sophisticated methods have been proposed, which is, however, beyond the scope of this thesis. Interested readers can refer to the surveys [27, 28, 29] for more details.

Visual feature extraction is another component of visual front end. One kind of visual feature extraction methods widely used is “bottom-up” strategy [30, 31, 32], in which the image data is examined at a low level and features are estimated directly from the image after some form of filtering such as pixel intensities, edges or regions, which are assembled into groups in an attempt to identify objects of interest. However, the features obtained by “bottom-up” based methods are usually high dimension, high redundancy and difficult for the classifier to extract invariant features for the effects of scaling, rotation, translation and illumination [23]. Thus, without a global prior model of what to expect, the “bottom-up” based methods are prone to failure as long as it is employed solely [33].

In contrast, the other kind of methods named “top-down” feature extraction make use of a prior model of what is expected in the image. The methods fall into this category attempt to find the optimal matching model to the target object in a given image [33]. Having matched the model, the target object can be represented by a low dimension feature. Compared to “bottom-up” one, there are at least two advantages of “top-down” based methods: (1) low computation complexity, and (2) invariance of the varying factors mentioned above. The “top-down” visual feature extraction method can be divided into three steps: color space selection, segmentation and model fitting which are shown in Fig. 1.2).

Firstly, it is necessary to select a space in which lips and skin have discriminated feature representations. Since the human lip and skin are characterized more by chromatic information than by grey level [34], the relevant studies are usually conducted in the common color space such as HSV, YCrCb and CIELAB. Moreover, some novel color spaces derived from the classical ones have been proposed as well. A brief review of color selection in lip of segmentation problem is available in Section
2.1.1.

Subsequently, in most cases, segmentation is employed to extract the lip region roughly. Segmentation is a fundamental problem in image processing and computer vision. The task of image segmentation is the process of classifying the structure elements (e.g. pixel, region, boundary and so forth) of an interesting image into some segments which collectively cover the entire given image. Each of such segment should be uniform with respect to specific statistical characteristics such as intensity, color, texture, etc [35, 36, 37, 38]. However, since most interesting objects in real situation usually have complex shapes, boundaries and morphology, and are often disturbed by noise which is hard to eliminate, there is no all-purpose method nowadays. Thus, image segmentation is still an active research field. In the literature, a number of image segmentation methods have been proposed, e.g., thresholding based [39, 40, 41], clustering [42, 43, 44], fuzzy set based [45], neural network based [46, 47, 48] and so forth. Detailed surveys are available in [49, 50, 51, 52].

Some experiments show that robust and accurate lip region or boundary segmentation is of vital importance for the final accuracy. For example, paper [53] demonstrates that the visual-only recognition error rate of a VSR system in studio environment, i.e. ideal light condition without shadow, is 37.3%. In contrast, the error rate will reach 76.2% when the same system is utilized in real world. One main reason is that the imprecise segmentation of lip caused by the noises in the real world such as unbalanced illumination condition mostly leads to the degraded
performance. This implies that the accuracy of lip segmentation is one of the most important factors to determine the recognition accuracy. Thus, it is non-trivial to segment the lip from a face image precisely, robustly and automatically. Specifically, the task of lip segmentation is to extract the lip region of contour with avoiding the interferences caused by moustache, tooth and so forth. The composition of region of interest (ROI) for lip segmentation is shown in Fig. 1.3. A review of lip segmentation method can be found in Section 2.1.2.

Finally, based on the segmented lip region or contour which may be represented by binary image, probability map and so forth, the model fitting is employed to get the feature vector. In usual, the model fitting process is usually formulated as an optimization problem. A brief review of the model fitting methods in lip segmentation problem is provided in Section 2.1.3.

### 1.2 Problem Statement and Contributions

A number of state-of-the-art segmentation methods have been proposed to segment the lip region from a given face image in the past decades. Nevertheless, lip segmentation is still a challenging problem under which there are at least two major issues: (1) segment number dependency problem, and (2) local scale determination.
problem. The following sub-sections give the details.

### 1.2.1 Segment Number Dependency Problem

Maximum a posteriori - Markov random field (MAP-MRF) framework is widely used in image segmentation because they yield a local and economical texture description. However, the methods require pre-assigning the number of segments (also called number of clusters interchangeably) appropriately in advance. Otherwise, the performance may result in over- or under-segmentation. An example of the under- and over-segmentation caused by the inappropriate pre-assigned segment number are shown in Fig. 1.4 (c) and (d). Unfortunately, from the practical viewpoint, the segment number is hard to be determined exactly beforehand because the appearances (mustache, teeth, tongue and so forth) of testers are different which may induce the variation of segment number. An intuitive illustration is shown in Fig. 1.5. To the best of our knowledge, only the method proposed in [8] is somewhat independent to the segment number. However, the lip tracking scheme presented in Section VI-B of [8] suggests that the segment number should be fixed to 2.

![Figure 1.4:](image)

Figure 1.4: (a) is the source image. (b), (c), (d) illustrate the corresponding lip segmentation results obtained by the same MRF based segmentation with $\hat{m} = m^*$, $\hat{m} < m^*$ and $\hat{m} > m^*$, respectively, where $m^*$ denotes the true segment number and $\hat{m}$ is the pre-assigned segment number.

Similarly, the fuzzy clustering based lip segmentation methods may encounter the same problem as well. By far, there are some number of clusters determination methods such as VR [54] and index-I [55] have been proposed, the basic idea of
most of them can be summarized as following:

1. Set the initial number of clusters $\hat{m}$ a small value, e.g. 2.

2. Perform the clustering algorithm with $\hat{m}$ clusters.

3. Calculate a specific criterion value.

4. Repeat step 2 with $\hat{m}$ increased by 1 if current criterion value is superior than that of the previous iteration, otherwise stop.

5. The optimal number of clusters is set to $\hat{m} - 1$.

Figure 1.5: The first row is the source image clips. The second row illustrates the corresponding segmentation results obtained by one MRF based method with the static pre-assigned segment number. The first image is over-segmented, the second image is segmented accurately, the third and forth images are under-segmented. The third row shows the segmentation results obtained by the same method with adaptive true segment numbers.
Moreover, the flow chart of clustering algorithm with adaptive number of clusters is shown as Fig. 1.6. According to this illustration, the optimal number of clusters is obtained by local exhaustive search which means that there are redundant data traverses in this kind of methods.

Figure 1.6: The flow chart of a typical clustering algorithm with adaptive number of clusters. The blocks in the dashed frame illustrate the number of clusters determination.

1.2.2 Local Scale Determination Problem

In the traditional MRF based image segmentation methods, the entire given image is processed under a constant scale in usual. However, the ability to describe the variable scale behaviors is limited. That is, their performance may become deteriorated as long as the input image is composed of the structures with different scales.
which is common in real world [56]. Fig. 1.7 illustrates a sample of segmentation results obtained by one MRF based method proposed in [57] under different scales. Fig. 1.7 (b) shows the over-segmentation in the result obtained under the finest scale. Although this problem is overcome under coarser scales (Fig. 1.7 (c) and (d)), the extraction of lip boundary is imprecise. Since, over- and under-segmentation may arise synchronously under the same scale, this problem is hard to be overcome by simply adjusting the global observation scale.

![Figure 1.7](image_url)

**Figure 1.7:** A sample of the segmentation results obtained by the same MRF based method under different scales. (a) is the source image; (b) is the segmentation result obtained on source image; (c) is the segmentation result based on the sub-sampling images with \( \frac{1}{2} \) source size; (d) is the segmentation result based on the sub-sampling images with \( \frac{1}{4} \) source size. For the convenience of comparison, all sub-sampling images are zoomed to the size of source image.

This phenomenon implies that each structure (e.g. region, pixel and so forth) in a given image may have its own optimal scale from the viewpoint of segmentation. Thus, lip segmentation accuracy can be improved as long as the appropriate local scale map is selected. However, for most of the existing multi-scale MRF based segmentation methods, there are two main shortcomings:

1. the scale map fixing and segmentation procedures are separated, which may induce redundant computational complexity, and

2. the segment number must be pre-assigned in advance, otherwise the estimation bias of the segment number may cause over- or under-segmentation.
1.2.3 Main Contributions of This Thesis

According to the study of the problems stated in Section 1.2.1 and Section 1.2.2, some methods are proposed to overcome the corresponding awkward situations. The contributions of this thesis are as follows:

- **A segment number independent lip segmentation method based on MAP-MRF framework.** An MAP-MRF based multi-layer hierarchical model is proposed which can perform lip segmentation without knowing the true segment number in advance. To implement this method, the number of pre-assign segments is larger than or equal to the ground truth (it is feasible to estimate the bound of number of segments roughly). Then, a rival penalized algorithm is utilized to suppress the over-segmentation caused by the estimation bias of segment number iteratively. This method inherits the advantages of MRF model, e.g. local spatial restriction consideration, and provides a solution to the result dependency problem.

- **A segment number independent lip segmentation method based on fuzzy clustering.** An objective function of fuzzy clustering which can serve as segment number independent lip segmentation is proposed. The objective function will reach the minimum in case the number of positions occupied by cluster centroids in pattern space is equal to the truth number of clusters, and these positions are coincident with the optimal cluster centroids. Then, a cooperative algorithm is presented by which the redundant centroids can be driven close to minimize the objective function. Thus, the clustering and the number of clusters selection can be performed synchronously.

- **A lip segmentation method with taking local scale variation into account.** A multi-scale MRF based method is proposed. Assume that each specific structure (e.g. pixel, region, or feature) in a given image has its own optimal scale from the viewpoint of segmentation. Moreover, each scale value is considered as a parameter controlling the size of corresponding neighborhood system.
Then, the scale selection problem can be formulated under the MAP-MRF framework. The optimal local scale map and the corresponding segmentation result can be obtained by minimizing an objective function. Moreover, this method is integrated into the first proposed model to provide a practical lip segmentation scheme.

1.3 Organization of This Thesis

The organization of the thesis is as follows. Chapter 2 provides the surveys of the lip segmentation and some corresponding techniques. Moreover, to enable the reader to better understand the experiments conducted in this thesis, the utilized benchmark databases are described briefly as well. Chapter 3 and Chapter 4 address the lip segmentation with unknown true segment number mainly. In Chapter 3, a multi-layer MRF based model is proposed to perform the segment number independent lip segmentation. In Chapter 4, a minimum entropy based fuzzy clustering method is presented which can extract the lip region without known the segment number in advance as well. Chapter 5 focuses on the scale selection problem in lip segmentation and describes a method taking advantage of the local scales variation under the MAP-MRF framework. Finally, Chapter 6 concludes the thesis and gives the future work directions.
Chapter 2

Review on Related Works, Used Models and Public Benchmark Databases

This chapter will firstly review some lip segmentation and correlative methods to show the recent progress on the topics related to this thesis. Then, the background knowledge which the proposed methods described in this thesis based on will be introduced. Moreover, to evaluate the performance of the proposed methods in comparison with the existing counterparts fairly, some public benchmark databases are utilized in this thesis. Details on these databases will be discussed in this chapter as well.

2.1 Review on Lip Visual Feature Extraction

As stated in Section 1.1, the existing “top-down” lip feature extraction methods can be divided into three steps: color space selection, lip segmentation and model fitting. This section will review some typical methods fall into the three categories respectively.
2.1.1 Color Space Selection

Color space selection is a relatively uncomplicated but crucial task because it has a direct influence on the final segmentation accuracy. The optimal color space for lip segmentation in ideal case is a space in which lip pixels and non-lip pixels are represented by two compact and distinct groups with low intra-class variances and high inter class variances [58]. Based on this principle, some current color spaces are selected in lip segmentation studies.

\( YC_{b}C_{r} \) is a color space with the chromatic and luminance information represented in different channels. Paper [59] uses \( C_{r} \) and \( C_{r}/C_{b} \) as the observation values. However, some researchers believe that the \( YC_{b}C_{r} \) color space is not very suitable for lip segmentation [60, 58].

In addition, a family of color spaces with luminance and chrominance separation as well which involves \( HSI, HSL \) and \( HSV \) is used for lip segmentation usually since hue has a good discriminative power. A successful example of lip segmentation based on \( HSV \) color space is available in paper [60].

\( CIELUV \) also exhibits luminance and chrominance and relates better to human vision than \( HSV \). However there are only few works [61, 62] based on this color space solely since lips are often very close to skin in its subspace \( UV \) [63].

\( CIELAB \) is another color space used conventionally to describe all the colors visible to the human eye. This space utilizes a channel \( a^{*} \) to represent the changing from green to magenta which is appropriate for distinguish lip and non-lip regions. The lip segmentation methods introduced in papers [64, 65] are performed in this color space. Moreover, paper [66] uses the feature provided by \( CIELAB \) and \( CIELUV \) cooperatively and achieves fairly good result.

On the other hand, some novel color spaces and channels have been proposed to differentiate lip and non-lip regions. Although the \( RGB \) color space and its subspaces cannot be used directly for lip segmentation since the strong correlation between the light and color information of lip and non-lip regions, there are several variants of original \( RGB \) color space are proposed. For example, the channels \( Q = \)
\( R/G \) [67], \( \hat{H} = R/(R + G) \) [68] and the color space named \( LUX \) [8] are derived from \( RGB \) and seem to be very suitable for lip color characterization. The works presented in [69, 70, 71, 72, 73, 74, 75, 65] prove their success. The illustrations of the observation of ROI in different color channels are shown in Fig. 2.1, and the corresponding lip and skin pixel histograms are shown in Fig. 2.2.

Figure 2.1: Illustrations of the ROI in different color channels: (a) source image, (b) gray-level, (c) \( H \) channel in \( HSV \), (d) \( S \) channel in \( HSV \), (e) \( U \) channel in \( CIELUV \), (f) \( V \) channel in \( CIELUV \), (g) \( a^* \) channel in \( CIELAB \), (h) \( b^* \) channel in \( CIELAB \), (i) \( Q \) channel, (j) \( \hat{H} \) channel, (k) \( U \) channel in \( LUX \), (l) \( X \) channel in \( LUX \).
Figure 2.2: Lip and skin pixel histograms in different color spaces: (a) gray-level, (b) $H$ channel in HSV, (c) $S$ channel in HSV, (d) $U$ channel in CIELUV, (e) $V$ channel in CIELUV, (f) $a^*$ channel in CIELAB, (g) $b^*$ channel in CIELAB, (h) $Q$ channel, (i) $H$ channel, (j) $U$ channel in LUX, (k) $X$ channel in LUX. The black dashed line corresponds to skin pixels and gray solid line to lip pixels.
2.1.2 Lip Segmentation

The simplest lip segmentation is the thresholding based method. This kind of methods was popular before 2000. Paper [69] utilizes two threshold values in $Q$ channel to get the lip pixel candidates. Papers [67, 76, 60, 77] also employ the thresholding based methods to segment the coarse lip region in $Q$, $H$ and $S$ channels. Since the threshold values are determined empirically, these methods cannot be generalized as the complexion difference, the appearance of moustache, and the illumination variation. Recently, this kind of methods is reported rarely. But as an important image processing strategy, thresholding is utilized widely in lip segmentation as pre- or post-processing [8, 66].

Due to the efficacy in boundary detection, gradient based methods are utilized widely in lip segmentation as well. The gradient based lip contour extraction methods are mainly different in the representation of the interesting image. Papers

![Figure 2.3: Some samples of the coarse segmentation results obtained by (b) the thresholding based method, (c) the fuzzy clustering method, (d) the gradient based method and (e) the statistical distribution method. (a) is the given source image.](image-url)
[78, 79, 80, 81, 82] implement the lip contour extraction in intensity map. Papers [83, 84] utilize the gradients to extract the inner and outer lip boundary in color information map. However, the accuracy of this kind of methods is easy to be affected by the false boundary edges.

Paper [75] utilizes gradient method to extract the lip counter as well. However, the gradient is not computed on any color space or channel but on lip membership maps in which the value of each pixel represents the degree of the pixel belong to lip region. This paper therefore assumes that the color distribution of the picture in $G/R$ channel can be approached by a gaussian mixture model (GMM). The parameters of the GMM are established by an EM algorithm initialized by a K-means algorithm with clusters number preassigned. Moreover, papers [85, 86] also uses a GMM to represent the color distributions of lip and skin. The similar works which utilize the statistical models to express the difference between the lip and non-lip pixels involve [87, 65].

As a fundamental method used commonly in different fields, clustering based segmentation gives a promising way of lip segmentation as well. There are several related methods proposed in the pase decades. In paper [64], FCM clustering is applied to the $b^*$ channel in CIELAB color space. In papers [88, 89], the K-means clustering based methods are developed. However, the classical clustering methods only utilize the feature data but ignore the relation between spatial position of pixels and the corresponding feature values which may convey important information in image analysis. The conjecture that physical properties in a neighborhood of space or an interval of time present some coherence and generally do not change abruptly [90] provides a new way for lip segmentation. It is feasible that the segmentation performance will be improved by taking into account the observation coherence provided by spatial topology of pixels. Paper [91] therefore proposes a fuzzy clustering algorithm by taking spatial restriction into account. That is, this method considers both of the distributions of data in feature space and the spatial interactions between neighboring pixels during clustering. Another fuzzy clustering
based lip segmentation method proposed in paper [92] obtains the spatial continuity constraints by using a dissimilarity index that allows spatial interactions between image voxels. Similarly, papers [93, 94, 95] deal with the lip segmentation by using fuzzy clustering with spatial restriction as well.

Refer to spatial restriction, MRF is another feasible and competitive strategy. Recently, some architectural modifications on classical MRF have been presented [96, 97, 98, 99, 100, 101, 37, 102, 103]. Along this line, papers [71, 104, 8] provide a multi-layer MRF based lip segmentation to deal with the lip segmentation and tracking problem. The layers correspond to adjacent frames in video sequence, respectively. Since each MRF is constructed by color features of the corresponding frame, this model integrates both hue information and temporal changes. The lip segmentation proposed in [60] is also under the MRF model. In this method, the energy function is composed by not only the color but also the edge information.

Some samples of segmentation results obtained by the lip segmentation methods proposed in [69, 91, 73, 75] are shown in Fig. 2.3 (b), (c), (d) and (e), respectively.

2.1.3 Model Fitting

The lip models employed in the existing “top-down” feature extraction methods can be divided into two classes roughly: the key points based and the shape based. In the former one, the lip shape is discretely described by a set of key points posed in the contour. In the latter one, the lip shape is represented by a set of connected curves.

Key Points Based Model

Active shape model (ASM) is a statistical model in which the object shape is represented by some points posed in the contour line whose positions can be controlled by adjusting a set of parameters indirectly. The initial parameters are learned beforehand by using the prior knowledge (shape, texture, and so forth) provided by the training set in which the elements are similar to the target objects. When model
fitting is performed, the key points positions are adjusted to fit the shape of target object by minimizing a cost function with respective to the shape parameters. The basic idea of another model named active appearance model (AAM) is similar but take the appearance, e.g. the texture information, into account. The detailed introductions and some classical examples can be found in [105, 33, 106].

In application, some studies employ the original ASM or AAM directly to produce the feature vector [107, 108, 109]. On the other hand, some researches focus on improving the cost function so as to make the model fitting more efficient. Paper [110] proposes a region based cost function with the form of product of the probabilities of all pixel inside the landmarks belong to the lip and the probabilities of all pixel outside the landmarks belong to the lip non-lip region. Moreover, papers [111, 6, 112, 113] also provide some excellent ASM/AAM based instances.

Paper [114] presents a 16 points model to describe the lip contour. Given a lip membership calculated by the fuzzy c-means with shape function (FCMS) proposed in [93], a region-based cost function derived from [110] can be established. Then an iterative point-driven optimization procedure has been developed to fit the lip model to the probability map. In each iteration, the adjustment of the 16 key points is governed by three pieces of quadratic curves that constrain the points to form a physical lip shape.

Shape Based Model

Active contour model (snake model) [115] is one of the most famous models in object contour extraction and tracking. In the model fitting process, the initial curve (snake) is driven to fit the object boundary by minimizing an energy function iteratively which is composed by external and internal energy. The external energy is computed from the image features (e.g. the color information, the membership map, etc.). The external energy is supposed to be minimal when the snake is at the object boundary position. The internal energy is determined by the prior knowledge such as curve mechanical properties. The internal energy is supposed to be minimal
when the snake has a shape which is supposed to be relevant considering the shape of the sought object.

Specifically, the model fitting begins with the region of interest (ROI) extraction which can provide an acceptable initial snake. The projection methods are utilized widely to fulfill this task. For example, in papers [80, 116], the mouth is located by analyzing the horizontal and vertical projections of the image gradients, coarse segmentation result or the intensity image. Moreover, the projection is also employed in the crucial points localization in the function curve based model fitting method which will be described later. Some samples of lip localization based on projection method are shown in Fig. 2.4.

![Figure 2.4: Two lip localization samples via projection method based on the different coarse segmentation results obtained by (a) gradient and (b) thresholding.](image)

The design of energy function is an important issue of this kind of methods as well. Paper [81] evenly spaces the snake points in the curve to keep the model tension. For the same motivation, paper [79] forces the snake points to preserve the distance between them during the lip tracking. Papers [117, 79, 116] present some selection methods of rigidity coefficients $\alpha$ and $\beta$ to make the internal energy term suit the prior shape of lip. There are also some external energy functions have been proposed to serve the lip contour extraction. The external energy terms utilized in papers [81, 79] are based on the intensity gradient. Paper [118] introduce a famous
external energy named gradient vector flow (GVF) [119] into lip tracking problem and obtain the promising result. Furthermore, the external energy terms utilized in paper [67] are computed based on a block of neighboring binary pixels. In [88], it is formalized by the combination of hue and luminance.

Although the snake model is largely used in lip contour extraction and tracking, there are several inevitable drawbacks presented, e.g. the high dependency between the extraction accuracy and the initial position of snake curve. Moreover, the parameter selection is also a dilemma.

The function curve is another interesting tool to model the lip shape. The basic idea of most methods in this is to extract some crucial points based on the lip segmentation result, and then the optimal fitting curves can be determined. That is, the model fitting process can be seen as a curve interpolation or approximation problem. Paper [85] uses two parabolas to approximate the lip outer contour. Other studies propose to use two parabolas instead of one to fit the upper/lower lip contour [120] or to employ quartics instead of parabolas [78, 69]. However, these models are still limited in the case of asymmetric mouth shape [73]. Paper [84] utilizes four cubic curves to model the lip contours. Each curve is determined by five crucial points with the least squares approximation method. Paper [83] also employs four cubic curves to fit the lip contours. However, there are only six crucial points defined. The lacking constraint equations are brought by some assumptions such as the derivative values in specific crucial points are zero.

Paper [121] also presents a polynomial curve based model. This model is composed by three curves in which one represents the lower lip contour, and two represent the upper. However, its model fitting is not treated as the curve interpolation or approximation problem. Similar to [114], the model fitting is implemented by optimizing a region based cost function on a lip membership calculated by a fuzzy clustering based method proposed in [66].

Besides the polynomial curve, Bezier curve is also a frequently-used model in lip shape representation [122, 123, 124, 75]. Compared to the polynomial based
curve, Bezier curve can represent conic region more exactly and permit some more additional shape control. Moreover, the B-spline curve based methods [125, 126] also provide another alternative way. As an intuitive example, the function curve models proposed in [83], [121] and [126] are shown in Fig. 2.5 (a), (b) and (c), respectively.

![Figure 2.5: Illustration of the lip model fitting result obtained by the (a) cubic polynomial curve based model, (b) order alterable polynomial curve based model and (c) B-spline curve based model.](image)

### 2.2 Summary of Used Models

As stated in Section 1.2.3, this thesis investigates the lip segmentation problem under two frameworks: MAP-MRF and fuzzy clustering. To enable the readers better understanding the thesis, the corresponding preliminary knowledge will be introduced.

#### 2.2.1 MAP-MRF Based Segmentation

Strictly speaking, MRF is not a segmentation method but a statistical model which provides a convenient and consistent way of modeling context-dependent entities. In most objects in real world, their physical properties in a neighborhood of space present some coherence and generally do not change abruptly [90]. Thus, MRF is believed suitable to model digital image.
Suppose the image of interest has $r$ rows and $c$ columns in pixel. Let site set $S = \{1, 2, \ldots, s\}$ with $s = r \times c$ index the pixels on regular lattice of the image. That is, each pixel position is regarded as a site. Subsequently, define a random fields $\mathcal{X} = \{X_i | i \in S\}$ to represent the observation of image. Each $X_i$ is assigned a value $x_i$ with $x_i \in O$ where $O$ is the set of possible observation values. For example, in an 8-bit gray-scale image, $x_i$ is the gray level value of pixel $i$ with $O = \{0, 1, \ldots, 255\}$. Moreover, the image segmentation can be formulated as a labeling optimization problem which is to assign a label from the label set to each of the image sites [90]. Therefore an MRF $\mathcal{F} = \{F_i | i \in S\}$ is defined on $S$ to describe the label values distribution which can be viewed as the segmentation result. Similar to $\mathcal{X}$, each $F_i$ can take a value $f_i$ in the label set $L = \{l_1, l_2, \ldots, l_m\}$ or simply $L = \{1, 2, \ldots, m\}$. For the sake of the convenience of discussion, a abbreviation is utilized hereinafter. That is, toward arbitrary random field $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_s\}$ defined on $S$ with configuration $\{y_1, y_2, \ldots, y_s\}$, the notation $P(y_i)$ is employed to abbreviate the probability of event $Y_i = y_i$, and the joint probability $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_s = y_s)$ is denoted as $P(y)$.

Based on the definition of MRF, $\mathcal{F}$ satisfies the positivity and Markovianity, respectively, i.e.:

$$P(f) \geq 0,$$  \hspace{1cm} (2.2.1)

$$P(f_i|f_{S-\{i\}}) = P(f_i|f_{N_i}),$$  \hspace{1cm} (2.2.2)

where $S - \{i\}$ is the set difference, $f_{S-\{i\}}$ denotes the set of values at the sites in $S - \{i\}$, and

$$f_{N_i} = \{f_{i'} | i' \in \mathcal{N}_i\}.  \hspace{1cm} (2.2.3)$$

In Eq.(2.2.3), $\mathcal{N}_i = \{i' \in S | |i - i'| \leq rad, i \neq i'\}$ is the set of sites neighboring $i$ with $|i - i'|$ denoting the Euclidean distance between $i$ and $i'$, and $rad$ is an integer value.

An MRF is characterized by its local property (the Markovianity) whereas a GRF (Gibbs random field) is characterized by its global property (the Gibbs distribution). Paper [127] employed the Hammersley-Clifford theorem to establish the equivalence.
of these two types of properties so that the prior probability \( P(f) \) can be specified by

\[
P(f) = \frac{1}{Z} e^{-U(f)/\tilde{T}}, \tag{2.2.4}
\]

where \( Z \) is a normalizing constant, and \( U(f) \) is the prior energy function. \( \tilde{T} \) is a constant called the temperature, which will be simply set at 1 hereinafter.

Paper [127] also provides a further specified form of the prior energy function:

\[
U(f) = \sum_{c \in \mathcal{C}} V_c(f), \tag{2.2.5}
\]

which is the sum of clique potentials \( V_c(f) \) over the set \( \mathcal{C} \) of all possible cliques. Clique \( c \) is an arbitrary subset of \( S \) which consists of a single site or more. In clique \( c \), every site is the neighbor of every other sites [127]. The isotropy clique set with respect to regular lattice and second-order neighborhood system is shown in Fig. 2.6.

Consider the computational complexity, the cliques of size up to two are utilized in application usually. In this case, Eq. (2.2.5) can be written as

\[
U(f) = \sum_{i=1}^{s} V_1(f_i) + \sum_{i=1}^{s} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{2.2.6}
\]

A classical categorization of this pair-wise MRF model is called auto-models [127] when \( V_1(f_i) = f_i G_i(f_i) \) and \( V_2(f_i, f_{i'}) = \beta_{i,i'} f_i f_{i'} \) where \( G_i(\cdot) \) are arbitrary functions and \( \beta_{i,i'} \) is constant weighting representing the pair-wise interaction between \( i \) and
i'. If the label set \( L = \{0, 1\} \) or \( L = \{-1, 1\} \), this is said to be the auto-logistic model with

\[
U(f) = \sum_{i=1}^{s} \alpha_i f_i + \sum_{i'=1}^{s} \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}
\tag{2.2.7}
\]

where \( \alpha_i \) is constant weighting as well.

The generalization of auto-logistic model is called multilevel logistic (MLL) [128, 129, 130] also known as Potts model [131]. For a second-order MLL model, where only single-site and pair-site clique parameters (\( \alpha \) and \( \beta \)) are non-zero, the potential function for a pair-wise clique is

\[
v_2(f_i, f_{i'}) = \begin{cases} 
\zeta_c & \text{if states on clique } \{i, i'\} \text{ have the same label} \\
-\zeta_c & \text{otherwise}
\end{cases}
\tag{2.2.8}
\]

where \( \zeta_c \) is the weighting parameter for type \( c \) cliques and \( C_2 \) is set of pair-site cliques.

Advocated by paper [132] and others [133, 130, 134, 135], MAP approach is commonly used to estimate the optimal solution of MRF models. Within the MAP-MRF framework, the labeling problem based image segmentation can be formulated as an optimization problem reasonably based on Bayesian decision and estimation theory. The optimal labeling configuration \( f^* = \{f^*_i \mid i \in S\} \) can be obtained by solving the following equation:

\[
f^* = \arg \max_{f \in \Omega_L} P(f \mid x),
\tag{2.2.9}
\]

where the Cartesian product \( \Omega_L = L \times L \times \ldots \times L = L^s \) is the configuration space of the label MRF \( \mathcal{F} \) on the site set \( S \).

According to the Bayes rule, the posterior probability shown in Eq.(2.2.9) can be computed by

\[
f^* = \arg \max_{f \in \Omega_L} \frac{P(f)P(x \mid f)}{P(x)}
\tag{2.2.10}
\]

For a specific given image, \( P(x) \) is seen as a constant value. Thus, Eq.(2.2.10) can be rewritten as

\[
f^* = \arg \max_{f \in \Omega_L} P(f)P(x \mid f)
\tag{2.2.11}
\]
As for the likelihood term \( P(x \mid f) \), similar to Eq.(2.2.4), let

\[
P(x \mid f) = \frac{1}{Z'} e^{-U(x \mid f)}
\]  

(2.2.12)

provided that \( P(x \mid f) \) is Gaussian distributed, where \( Z' \) is a normalizing constant and \( U(x \mid f) \) denotes the likelihood energy function. As a result, Eq.(2.2.11) can be rewritten as:

\[
f^* = \arg \min_{f \in \Omega} \sum_{c \in \mathcal{C}} V_c(f) + \alpha U(x \mid f),
\]  

(2.2.13)

where the weighting parameter \( \alpha \) is introduced to determine the proportion of \( U(f) \) to \( U(x \mid f) \) in the entire energy. Thus far, a number of optimization methods such as simulated annealing (SA), iterated conditional modes (ICM), and so forth, can be utilized to solve Eq.(2.2.13).

### 2.2.2 Fuzzy Clustering Based Segmentation

Clustering is the process of assigning data elements into classes or clusters so that data in the same class are as similar as possible from the view of certain given measures of similarity. Since the comparability between the task of image segmentation and clustering, the image segmentation can be addressed as a clustering problem. Under the clustering framework, the specific property measured in feature space of each pixel is viewed as the data to be divided. Meanwhile, the image segments turn into data clusters.

Fuzzy clustering [136] is a class of algorithms for cluster analysis in which data elements can belong to more than one cluster, and associated with each element is a set of membership levels. To discuss the image segmentation problem under this framework, suppose that the image of interest has \( s \) pixels as well. For \( i \)th pixel, the feature vector which is utilized in clustering process is denoted by \( x_i \). Then, define \( m \) segments whose centroids consist a collection \( C = \{c_1, c_2, \ldots, c_m\} \). The purpose of fuzzy clustering algorithm is to give the optimal \( C \) and corresponding partition
matrix

\[
U = \begin{bmatrix}
  u_{11} & \cdots & u_{1m} \\
  \vdots & \ddots & \vdots \\
  u_{s1} & \cdots & u_{nm}
\end{bmatrix}
\]  

(2.2.14)

with

\[
\sum_{j=1}^{m} u_{ij} = 1, \quad (i = 1, 2, \ldots, n)
\]

(2.2.15)

where \(u_{ij} \in [0, 1]\) indicates the strength of the association between an input data \(x_i\) and cluster \(c_j\).

One of the most popular fuzzy clustering algorithms is the fuzzy C-means (FCM) algorithm [137]. In this algorithm, the optimal cluster centroids and partition can be achieved by minimizing the following objective function

\[
J = \sum_{i=1}^{s} \sum_{j=1}^{m} u_{ij} |x_i - c_j|^2.
\]

(2.2.16)

Moreover, there are lots of alternative fuzzy clustering methods. From the viewpoint of information theory, the information entropy can be viewed as a measure of the uncertainty. Moreover, the uncertainty of belonging of each input data is reduced during the clustering procedure. Thus, the relationship between clustering and entropy is naturally close. Mathematically, Shannon's entropy [138] of a random variable \(x\) with the probability \(p(x)\) is defined as

\[
H(x) = - \sum_{x} p(x) \log p(x).
\]

(2.2.17)

Based upon Shannon's entropy, Bezdek proposed a fuzzy clustering criterion named partition entropy (PE) [139, 137] to measure the fitness of a fuzzy partition which is shown as

\[
H(U, m) = - \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{m} u_{ij} \log u_{ij}.
\]

(2.2.18)

Paper [140] indicates that the partition matrix and number of clusters \((U^*, m^*)\) is optimal as long as

\[
(U^*, m^*) = \arg \min_{1 < m \leq m_{\text{max}}} \{ \arg \min_{U \in \Omega_m} \{ H(U, m) \} \},
\]

(2.2.19)
where $m_{max}$ denotes the maximum value of the number of clusters, and $\Omega_m$ is the collection of partition matrices with number of clusters $m$.

Further, paper [141] proposes another version of the minimum entropy criterion of fuzzy clustering, in which the membership degree of $x_i$ in cluster $c_j$ can be posed by conditional probability. Thus, Eq.(2.2.18) can be rewritten as

$$H(C|X) = -\frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{m} p(c_j|x_i) \log p(c_j|x_i). \tag{2.2.20}$$

Moreover, for the sake of analysis, paper [141] utilizes Havrda-Charvat’s structural $\alpha$-entropy [142]:

$$H^\alpha(x) = (2^{1-\alpha} - 1)^{-1} [\sum_x p^\alpha(x) - 1], \tag{2.2.21}$$
as a substitution of Shannon’s entropy, where $\alpha > 0$ and $\alpha \neq 1$. Evidently, different $\alpha$ can lead to different entropy measures.

Specifically, in this thesis, the following quadratic entropy with $\alpha = 2$ is selected, i.e.

$$H^2(x) = 1 - \sum_x p^2(x). \tag{2.2.22}$$

Thus, Eq.(2.2.20) can be further rewritten as

$$H(C|X) = 1 - \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{m} p^2(c_j|x_i). \tag{2.2.23}$$

To show the validity of this criterion, formulate the probability of clustering error as

$$P_e = P(C \neq C^*), \tag{2.2.24}$$

where $C^*$ denotes the optimal cluster centroid collection.

Based on Fano’s inequality [143], have

$$H(P_e) + P_e \log(m - 1) \geq H(C|X), \tag{2.2.25}$$

where $H(P_e)$ is the Shannon’s entropy of $P_e$. Since $H(P_e) \leq 1$ and $m \geq 2$, Eq.(2.2.25) can be further rewritten as

$$P_e \geq \frac{H(C|X) - 1}{\log(m - 1)}. \tag{2.2.26}$$
Eq.(2.2.26) indicates that $C^*$ can be estimated with a low error probability only if $H(C|X)$ is small. This implies that minimum $H(C|X)$ could be an appropriate choice for fuzzy clustering [141].

2.3 Testing Databases Used in This Thesis

In this thesis, AR [144], CVL [145, 146], GTAV [147] and VidTIMIT [148] databases are used for evaluation the proposed method and comparison against the existing counterparts.

Figure 2.7: Some sample images from AR (top row), CVL (middle row) and GTAV (bottom row) face databases. Only clipped the lip region to show the details of the interesting parts. The three databases provide various lip shape states, some of them are even covered by the moustache. Moreover, the hues and color temperatures of images from different databases are dissimilar which is caused by the diverse complexions and illumination conditions.

AR database contains over 4,000 color images corresponding to 126 people’s faces (70 men and 56 women). The images feature frontal view faces with different facial expressions. CVL database contains 798 color images corresponding to 114 people’s
Figure 2.8: Some sample frames from the VidTIMIT database. Only clipped the lip region to show the details of the interesting parts. The image sequence in each row corresponds a specific video clip.

faces (mostly male). For each person, the images were captured seven times with different facial expressions. GTAV database includes a total of more than 1000 color face images from 44 persons. For each person, the images were captured 27 times with different facial expressions. Since the lip shape state is highly related to the corresponding human expression, the three databases provide various lip shapes as the experiment samples to test the accuracy of lip segmentation methods. Moreover, since the complexions of the photographed people and the illumination conditions (natural light, strong light source from an angle of 45°, and finally an almost frontal mid-strong light source) of different databases are obviously dissimilar, the three databases are in the position of testing the robustness of the segmentation methods. Some sample images from these three databases are available in Fig. 2.7 in which only lip regions are clipped so as to show the details of interesting parts as much as possible.

VidTIMIT database is comprised of video and corresponding audio recordings of 43 people, reciting short sentences. The sentences were chosen from the test section of the TIMIT corpus. There are ten sentences per person. In this thesis, only the video information of this database is utilized to test the tracking ability of a
proposed lip segmentation method under the situation of people speaking. Sample video frame from this database is shown in Fig. 2.8 in which there are only lip regions are clipped as well.
Chapter 3

MAP-MRF Based Lip Segmentation with Unknown Segment Number

3.1 Introduction

This chapter will present an MAP-MRF based image segmentation method whose performance is independent of the estimation of the segment number. Firstly, a multi-layer model is built up, in which each layer is formed by 2-D random field defined on the pixel lattice of the interesting image. In this model, the number of layers is exactly the same as the number of predefined segments on purpose so that each layer corresponds to one segment cluster. Furthermore, the value configuration of a specific layer is composed by the membership degree of each pixel in the corresponding cluster. Then, an MRF is derived from the multi-layer model with an energy function, through which the segmentation problem is formulated as a labeling optimization problem under the MAP-MRF framework. Suppose the pre-assigned number of segment clusters may over-estimate the ground truth, i.e. it may lead to the over-segmentation. Analogous to the work in [149], a rival penalized iterative algorithm is therefore presented to optimize this energy function. Specifically, for
each observed data (also called *observation* or *input data* interchangeably), a winning segment cluster (simply called *winner*) is selected from all clusters, while the remaining clusters are regarded as rivals. Not only is the winner updated to adapt to the observed data, but also all rivals are penalized with the penalty strength proportional to the similarity between the rival and the winner. This rival penalized iterative algorithm is capable of gradually fading out the redundant segment clusters automatically, whereby the over-segmentation is smoothly eliminated. Based upon this algorithm, a lip segmentation and tracking scheme is proposed, featuring the robust performance to the estimate of the number of segment clusters. Experimental results have shown the efficacy of the proposed method in comparison with the existing counterparts.

The rest of this chapter is organized as follows. Section 3.2 describes the proposed method in detail. Section 3.3 presents the unsupervised lip segmentation and tracking scheme based upon the proposed method. Section 3.4 shows the experimental results and discussion. Finally, a summary is shown in Section 3.5.

### 3.2 Multi-layer Model based Image Segmentation

This section firstly introduces the framework of the multi-layer model. Then, a rival penalized iterative algorithm is presented to perform the optimal segmentation without true segment number, i.e., to fade out the over-segmentation caused by the redundant pre-defined segment(s).

#### 3.2.1 Multi-Layer Model

Given an image, suppose there are \( \hat{m} \) segment clusters in observation space and the pixels drawn from one cluster are independently and identically distributed. Each of such cluster should be uniform with respect to statistical properties characterized by the corresponding parameter from the collection \( \Theta = \{ \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(\hat{m})} \} \).

A multi-layer model is utilized to represent the soft segmentation, i.e. each pixel
may belong to multiple segments with the different degree of membership, in which each layer is implemented by 2-D random field with the sites on pixel lattice of the image. Moreover, these random fields share a continuous value set which can be regarded as a normalized fuzzy set $[0, 1]$. In this model, as shown in Fig. 3.1, the number of layers is the same as the pre-assigned segment number $m$ so that each layer can correspond to one segment cluster. In the random field laid on layer $j$, the value at site $i$ (denoted by $f^{(j)}_i$) can be considered as the probability of the event that the pixel at site $i$ belongs to segment $j$. Thus, each $f^{(j)}_i$ is assigned a value from $[0, 1]$ by

$$f^{(j)}_i = \frac{G(x_i; \theta^{(j)})}{\sum_{j'=1}^m G(x_i; \theta^{(j')})},$$

where $G$ denotes the probability density function (p.d.f.) of any distribution.

Subsequently, an MRF is utilized to formulate the segmentation as a labeling optimization problem. The site set of the MRF is $S = \{1, 2, \cdots, s\}$, the value at site $i$ is denoted by $f_i$ which takes a value from $L = \{1, 2, \cdots, \hat{m}\}$. Recall the classical MRF based method introduced in Section 1.3.1, the objective function is:

$$\mathcal{E}(f; x) = U(f) + \alpha U(x|f) = \sum_{c \in \mathcal{C}} V_c(f) + \alpha \sum_{i=1}^s \sum_{f_i=1}^{\hat{m}} h(f_i) U(x_i|f_i)$$

with

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f)$$

and

$$U(x|f) = \sum_{i=1}^s \sum_{f_i=1}^{\hat{m}} h(f_i) U(x_i|f_i),$$

where $s$ denotes the pixel number in the given image and $h_i(\cdot)$ is an indicator function given by

$$h(f_i) = \begin{cases} 1, & f_i = \arg \max_j f^{(j)}_i, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, since the selection of $\Theta$ can also affect the value of Eq. (3.2.2), the notation $\mathcal{E}(f; x, \Theta)$ is utilized to replace $\mathcal{E}(f; x)$ hereinafter.
Figure 3.1: The architecture of the proposed model, in which there are $\hat{N}$ layers, i.e. $\hat{N}$ segment clusters. Each layer is formed by a 2-D random field with the site set $S$ on regular lattice of the image as shown in the bottom. The value $f_i^{(j)}$ denotes the probability of pixel at site $i$ belonging to segment $j$. 
Thus, the optimal segmentation can be obtained by solving the equation

$$f^* = \arg \min_{f \in \Omega} E(f; x, \Theta).$$

(3.2.6)

Figure 3.2: Illustration of (a) the first-order neighborhood system, and (b) the corresponding cliques utilized in the proposed method, where the black and gray circles denote the pixel $i$ and the neighborhood system, respectively. Please note that there are only two cliques because this is an isotropic system.

Moreover, since the first-order isotropy neighborhood system is utilized in this model as shown in Fig. 3.2, the prior energy term $U(f)$ in Eq. (3.2.3) can be further specified as

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) = \sum_{i=1}^{s} \sum_{i' \in \mathcal{N}_i} [1 - \delta(\sum_{f_i=1}^{\hat{m}} h(f_i) \cdot f_i - \sum_{f_{i'}=1}^{\hat{m}} h(f_{i'}) \cdot f_{i'})],$$

(3.2.7)

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

Suppose the p.d.f. of the observed data $x_i$, i.e. $\mathcal{G}(x_i; \theta^{(j)})$ in Eq. (3.2.1), is Gaussian. Thus, the parameter collection can be specified by $\theta^{(j)} = \{\mu^{(j)}, \Sigma^{(j)}\}$ where $\mu^{(j)}$ and $\Sigma^{(j)}$ denote the mean vector and covariance matrix of the segment cluster $j$, respectively. Subsequently, the likelihood energy term can be specified as

$$\sum_{i=1}^{s} \sum_{f_i=1}^{\hat{m}} h(f_i) U(x_i|f_i) = \sum_{i=1}^{s} \sum_{f_i=1}^{\hat{m}} h(f_i)[(x_i - \mu^{(f_i)})^T (\Sigma^{(f_i)})^{-1} (x_i - \mu^{(f_i)})],$$

(3.2.8)

where $T$ denotes the transpose operation of a matrix.
3.2.2 Rival Penalty Iterative Algorithm

In general, an upper bound of the true segment number can be easily estimated. For example, given a lower face image in lip segmentation problem, the possible segment number is around 5 from the viewpoint of the chromatism, i.e. lip, skin, moustache, teeth, and tongue. Thus, it is feasible to simply estimate this upper bound by a value significantly larger than 5, say 10, in this case. Subsequently, assign \( m \) a random integer as long as it is larger than or equal to the true segment number \( m^* \).

Evidently, over-segmentation will be occurred because of the redundant \( \hat{m} - m^* \) clusters. To eliminate the over-segmentation automatically, an iterative algorithm is therefore utilized to obtain the optimal segmentation by minimizing Eq. (3.2.6). The basic idea of this algorithm is that, for an input data, the winning segment cluster is determined, while the other clusters are regarded as rivals. As illustrated in Fig. 3.3, not only is the winner updated to adapt to the input data, but also all rivals are penalized with the penalty strength proportional to the similarity between the rival and the winner. As shown in [149], such a rival penalized scheme can make the redundant clusters gradually fade out, whereby the over-segmentation problem is solved. In the following of this section, this algorithm shall be introduced in detail.

For simplicity, the notation \( \mathfrak{N}_i^+ \) is utilized to represent site set \( \mathfrak{N}_i \cup i \). Let \( \{ f^{(j)}_{i'}|i' \in \mathfrak{N}_i^+ \} \) be regarded as a system as \( i \) and \( j \) are given. Eq. (3.2.7) can be therefore approximated by

\[
\hat{U}(f) = \sum_{i=1}^{s} \sum_{j=1}^{\hat{m}} \sum_{i' \in \mathfrak{N}_i^+} \frac{f^{(j)}_{i'}}{F_i^{(j)}} \ln \frac{f^{(j)}_{i'}}{F_i^{(j)}},
\]

where \( F_i^{(j)} \) denotes the normalizing constant with

\[
F_i^{(j)} = \sum_{i' \in \mathfrak{N}_i^+} f^{(j)}_{i'}, \quad i \in S, j \in L.
\]

It can be seen that, compared to Kronecker delta function, Eq. (3.2.9) utilizes the information entropy based disorder measurement in local object system to represent the spatial coherence restriction.
Figure 3.3: (a) The initial position of cluster centroids marked by ‘*’; (b) and (c) illustrate how the proposed algorithm work intuitively as given an input data marked by black circle, where the arrows with real line show the penalty direction and the length are proportional to the penalty strength. The arrows with the dotted line show the moving direction of the winner. (d) is the desired result in which the redundant centroids are driven far away from the input data set so that the over segmentation caused by them can be eliminated.
Moreover, an inter-layer membership degree system and a derived entropy based similarity measurement are introduced to implement the rival penalized algorithm. This system, denoted as $\mathcal{F}_{i}^{(j,j')}$, is a membership set composed by the specific values from $L$. As illustrated in Fig. 3.4, it can be mathematically represented as follows:

$$\mathcal{F}_{i}^{(j,j')} = \mathcal{F}_{i}^{(j)} \cup \mathcal{F}_{i}^{(j')} = \{f_{i}^{(j)}|i' \in \mathcal{M}_{i}^{+}, j \in L\} \cup \{f_{i}^{(j')}|i' \in \mathcal{M}_{i}^{+}, j' \in L\}, i \in S, j' \neq j.$$  

(3.2.11)

![Figure 3.4: There are two random fields laid on the different layers $j$ and $j'$, where each site is represented by a circle. In each layer, site $i$ is denoted by the black circle, and the neighboring site set of $i$, i.e. $\mathcal{M}_{i}$, is denoted by the gray circles. An inter-layer membership degree system with respect to site $i$, layer $j$ and $j'$, i.e. $\mathcal{F}_{i}^{(j,j')}$, is composed of the values in all non-white circles.](image)

Based upon this system, entropy difference is employed to measure the similarity between $\mathcal{F}_{i}^{(j)}$ and $\mathcal{F}_{i}^{(j')}$. Notation $E_{i}^{(j)}$ to represent the normalized entropy of system $\mathcal{F}_{i}^{(j)}$:

$$E_{i}^{(j)} = -\frac{1}{n} \sum_{i' \in \mathcal{M}_{i}^{+}} \frac{f_{i}^{(j)}}{F_{i}^{(j)}} \ln \frac{f_{i}^{(j)}}{F_{i}^{(j)}},$$  

(3.2.12)

where $F_{i}^{(j)}$ is the normalizing constant shown in Eq. (3.2.10), and $n$ is the number of elements in $\mathcal{F}_{i}^{(j)}$.

Similarly, if the system is extended to $\mathcal{F}_{i}^{(j,j')}$, the normalized entropy will become:

$$E_{i}^{(j,j')} = -\frac{1}{2n} \sum_{i' \in \mathcal{M}_{i}^{+}} \left[ \frac{f_{i}^{(j)}}{F_{i}^{(j,j')}} \ln \frac{f_{i}^{(j)}}{F_{i}^{(j,j')}} + \frac{f_{i}^{(j')}}{F_{i}^{(j,j')}} \ln \frac{f_{i}^{(j')}}{F_{i}^{(j,j')}} \right]$$  

(3.2.13)
with
\[ F_i^{(j,j')} = \sum_{i' \in \mathcal{N}_i^+} [f_i^{(j)} + f_i^{(j')}]. \] (3.2.14)

Hence, let
\[ \delta E_i^{(j,j')} = E_i^{(j)} - E_i^{(j')} \] (3.2.15)
denotes the entropy difference caused by every variation of system.

Intuitively, as long as \( \delta E_i^{(j,j')} \) is positive, it means that the incoming memberships reduce the uncertainty of the system. Under the circumstances, further consider that the distribution of incoming memberships is similar to the primary’s, and vice versa. Thus, \( \delta E_i^{(j,j')} \) can be regarded as a measurement of membership distribution similarity between \( \tilde{\mathcal{H}}_i^{(j)} \) and \( \tilde{\mathcal{H}}_i^{(j')} \). The similarity can be considered as a menace to the uniqueness of the winning cluster, from which pixel \( i \) is drawn. It is reasonable that the rival with more similarity against the winner should get more fierce penalty. Thus, the larger \( \delta E_i^{(jw,jr)} \) is, the stronger the penalty force is. On the other hand, as \( \delta E_i^{(jw,jr)} \) is negative, it means that the rival is not similar to winner and unnecessary to be penalized. Thus, \( \delta E_i^{(jw,jr)} \cdot 1_{[0, +\infty)}(\delta E_i^{(jw,jr)}) \) is utilized as the penalty strength, where \( 1_{[0, +\infty)}(\delta E_i^{(jw,jr)}) \) denotes the step function given by
\[
1_{[0, +\infty)}(\delta E_i^{(jw,jr)}) = \begin{cases} 
1, & \delta E_i^{(jw,jr)} \in [0, +\infty), \\
0, & \text{otherwise}. 
\end{cases}
\] (3.2.16)

Subsequently, the segmentation result \( f^* \) can be obtained by solving the optimization problem shown in Eq. (3.2.6). Initialize \( \Theta \), and calculate \( f_i^{(j)} \). Then, for each observed data \( x_i \), the rival penalized method is utilized to update \( \Theta \) in order to minimize the cost function while fading out the redundant segment cluster(s). Analogous to the Expectation-Maximization (EM) algorithm, the proposed one consists of the three steps: E-Step (Expectation), A-Step (Adjustment) and M-Step (Minimization). The detailed implementation is shown as follows:

1. **E-Step.** Fixing \( \Theta \), calculate \( f_i^{(j)} \) by Eq. (3.2.1).
2. **A-Step.** Considering the prior energy term Eq. (3.2.7) approximated by Eq. (3.2.9), adjust \( f_i^{(j)} \) via ICM to make the local systems \( \mathfrak{N}_i^+ \) reaching the most uncertainty, which provides the spatial restriction, i.e.

\[
\hat{f}_i^{(j)} = \exp(-\sum_{i' \in \mathfrak{N}_i} \hat{f}_i^{(j)} \ln \hat{f}_{i'}^{(j)}),
\]

(3.2.17)

whose derivation can be found in Appendix A.1.

Let \( f_i^{(j)} = \hat{f}_i^{(j)} \).

3. **M-Step.** Fixing \( \hat{f}_i^{(j)} \), update \( \Theta \) for each \( x_i \) via

\[
\hat{\theta}^{(jw)} = \theta^{(jw)} - \eta \frac{d \mathcal{E}(f; x, \Theta)}{d \Theta} \bigg|_{\theta^{(jw)}},
\]

(3.2.18)

and

\[
\hat{\theta}^{(jr)} = \theta^{(jr)} + \eta \cdot \delta E_i^{(jw, jr)} \cdot 1_{[0, +\infty)}(\delta E_i^{(jw, jr)}) \cdot \frac{d \mathcal{E}(f; x, \Theta)}{d \Theta} \bigg|_{\theta^{(jr)}},
\]

(3.2.19)

where \( j_w = \arg \max_j f_i^{(j)} \), \( j_r \in L \) but \( j_r \neq j_w \), and \( \eta \) is the positive learning rate. Let \( \theta^{(jw)} = \hat{\theta}^{(jw)} \) and \( \theta^{(jr)} = \hat{\theta}^{(jr)} \).

In implementation, given site \( i \) and \( j_w \), \( \frac{d \mathcal{E}(f; x, \Theta)}{d \Theta} \bigg|_{\theta^{(jw)}} \) can be written by \( \{ \frac{\partial \mathcal{E}(f; x, \Theta)}{\partial \mu} \bigg|_{\theta^{(jw)}} \}, \frac{\partial \mathcal{E}(f; x, \Theta)}{\partial \Sigma} \bigg|_{\theta^{(jw)}} \} \) with

\[
\frac{\partial \mathcal{E}(f; x, \Theta)}{\partial \mu} \bigg|_{\theta^{(jw)}} = 2 \cdot \Sigma^{(jw)} \cdot (\mu^{(jw)} - x_i)
\]

(3.2.20)

and

\[
\frac{\partial \mathcal{E}(f; x, \Theta)}{\partial \Sigma} \bigg|_{\theta^{(jw)}} = (\mu^{(jw)} - x_i) \cdot (\mu^{(jw)} - x_i)^T.
\]

(3.2.21)

Thus, Eq. (3.2.18) can be expressed as

\[
\hat{\mu}^{(jw)} = \mu^{(jw)} - 2\eta \cdot \Sigma^{(jw)} \cdot (\mu^{(jw)} - x_i)
\]

(3.2.22)

and

\[
\hat{\Sigma}^{(jw)} = \Sigma^{(jw)} - \eta \cdot (\mu^{(jw)} - x_i) \cdot (\mu^{(jw)} - x_i)^T.
\]

(3.2.23)

with \( \hat{\theta}^{(jw)} = \{ \hat{\mu}^{(jw)}, \hat{\Sigma}^{(jw)} \} \).
Similarly, given \( i \) and corresponding \( j_r \), Eq. (3.2.19) can be expressed as

\[
\hat{\mu}^{(j_r)} = \mu^{(j_r)} + 2\eta \cdot \delta E_i^{(j_w,j_r)} \cdot 1_{[0,\infty)}(\delta E_i^{(j_w,j_r)}) \cdot \Sigma^{(j_r)} \cdot (\mu^{(j_r)} - x_i) \tag{3.2.24}
\]

and

\[
\hat{\Sigma}^{(j_r)} = \Sigma^{(j_r)} + \eta \cdot \delta E_i^{(j_w,j_r)} \cdot 1_{[0,\infty)}(\delta E_i^{(j_w,j_r)}) \cdot (\mu^{(j_r)} - x_i) \cdot (\mu^{(j_r)} - x_i)^T \tag{3.2.25}
\]

with \( \hat{\theta}^{(j_r)} = \{\hat{\mu}^{(j_r)}, \hat{\Sigma}^{(j_r)}\} \).

The E-Step, A-Step and M-Step are implemented iteratively until \( f \) converges.

Please note that Step 3 not only updates the associated parameters of the winning segment cluster, i.e., the \( j_w \) one, to adapt to the input, but the others are also penalized with the force strength proportional to \( \delta E_i^{(j_w,j_r)} \), respectively.

### 3.3 Lip Boundary Extraction and Tracking

This section applies the proposed method in Section 3.2 to the unsupervised lip segmentation and tracking problem. The task is to extract the lip boundary from a color image consisting of a part of face between nostril and chin. A sample of original image is shown in Fig. 3.5 (a).

#### 3.3.1 Observation Space

In general, the image segmentation methods are based on color space rather than gray level because color image can provide more useful information for segmentation. Furthermore, since HSV color space is similar to the way human being perceives [150], a modified HSV color space is utilized as our observation space.

Transform the original image into the HSV color space. The hue, saturation, and value component for site \( i \) are denoted by \( H_i \), \( S_i \) and \( V_i \), respectively. The \( V \) component only conveys the luminance information, which may be disturbed by illumination. Moreover, some regions with relative deep color regions are hard to be discriminated in \( V \) channel (e.g. the hues of lip and moustache are obvious different, but the luminance values are similar). Thus, only the \( H \) and \( S \) components are
Figure 3.5: (a) is the source lip image. (b)–(h) illustrate the initial membership $f_i^{(j)}$ in the seven different layers, i.e. $\hat{m} = 7$. These memberships are calculated based on the initial $\mu$ and $\Sigma$.

utilized. Since the following calculation is based on the Euclidean distance, transform $H$ and $S$ components from Polar Coordinate system to Cartesian Coordinate system. That is, for each site, let the observed data be:

$$x_i = [H_i \cdot \cos(2\pi \cdot S_i), H_i \cdot \sin(2\pi \cdot S_i)]^T, \quad i \in S.$$  \hspace{1cm} (3.3.26)

### 3.3.2 Segmentation and Binarization

After obtaining $f^*$ and $f_i^{(j)}$ via the proposed method in Section 3.2, the following equation is utilized to obtain the hard segmentation result, i.e. each pixel belongs to one segment only:

$$S^{(j)} = \{i \mid j = \arg \max_j f_i^{(j)}, i \in S\},$$  \hspace{1cm} (3.3.27)

where $S^{(j)}$ denotes the set of sites fallen into layer $j$. A sample of $S^{(j)}$s with $\hat{m} = 7$ is shown in Fig. 3.6. The corresponding initial membership map $f_i^{(j)}$ is shown in Fig. 3.5. Obviously, almost no site is assigned into $S^{(1)}$, $S^{(3)}$, $S^{(4)}$ and $S^{(5)}$, that
Figure 3.6: The segmentation results in the seven different layers, while the
initial membership is shown in Fig. 3.5. The white pixels in (a)-(g) represent
the foreground, i.e. the site sets: $S^{(1)}, S^{(2)}, \ldots, \text{and } S^{(7)}$.

is, the corresponding redundant segments have been faded out during the learning
process.

### 3.3.3 Tracking

For the $k^{th}$ frame in a video, the observation value of site $i$ is denoted by $x_{i, [k]}$. The
difference can be written by

$$\delta x_{i, [k], [k+1]} = |x_{i, [k]} - x_{i, [k+1]}|. \quad (3.3.28)$$

Based on the observation difference values, the “inequable” regions between two
adjacent frames are obtained by

$$S_{[k], [k+1]} = \{i \mid \delta x_{i, [k], [k+1]} > t_{[k], [k+1]}, \quad i \in S\}. \quad (3.3.29)$$

where $t_{[k], [k+1]}$ is a threshold value selected by the method proposed in [151] based
on $\{\delta x_{i, [k], [k+1]} \mid i \in S\}$.

Given the segmentation result of the $k^{th}$ frame in label map form denoted by $f^*_k$.
Suppose that the segmentation result of next frame, that is $f^*_k$, can be derived
from \( f^*_i \). Specifically, in the \( k + 1 \)th frame, for the sites not belong to “inequable” regions \( S_{[k],[k+1]} \), believe that their labels are inherited from the \( f^*_i \) by

\[
f^*_i,_{[k+1]} = f^*_i,_{[k]}, \quad i \notin S_{[k],[k+1]}, \tag{3.3.30}
\]

where \( f^*_i,_{[k]} \) and \( f^*_i,_{[k+1]} \) denote the final label of the site \( i \) in the \( k \)th and \( k + 1 \)th frame, respectively.

Moreover, for the \( k \)th frame, define a label MRF \( \mathcal{F}_{[k]} = \{ F_i,_{[k]} \mid i \in S_{[k],[k+1]} \} \) and an observation value random field \( X_{[k]} = \{ X_i,_{[k]} \mid i \in S_{[k],[k+1]} \} \) on the site set \( S_{[k],[k+1]} \). Each \( F_i,_{[k]} \) is assigned a value \( f_i,_{[k]} \) with \( f_i,_{[k]} \in L \), each \( X_i,_{[k]} \) can take a value \( x_i,_{[k]} \) in \( O \), the corresponding configuration can be denoted by \( f,_{[k]} \) and \( x,_{[k]} \), respectively. Thus, for the sites belong to \( S_{[k],[k+1]} \), the segmentation results should be reestimated by the classical MAP-MRF method:

\[
f^*_{[k+1]} = \arg \min_{f,_{[k+1]} \in \Omega} \sum_{c \in \mathcal{C}} V_c(f) + \alpha U(x,_{[k+1]} \mid f,_{[k+1]}) \tag{3.3.31}
\]

with

\[
\sum_{c \in \mathcal{C}} V_c(f) = \sum_{i \in S_{[k],[k+1]}} \sum_{v \in \mathbb{R}_i} [1 - \delta(f_i,_{[k+1]} - f_{i',_{[k+1]}})], \tag{3.3.32}
\]

and

\[
U(x,_{[k+1]} \mid f,_{[k+1]}) = \sum_{i \in S_{[k],[k+1]}} \sum_{f_i,_{[k+1]} \in L} \left[ (x_i,_{[k+1]} - \mu_i^{(f_i,_{[k+1]})})^T (\Sigma_i^{(f_i,_{[k+1]})})^{-1} (x_i,_{[k+1]} - \mu_i^{(f_i,_{[k+1]})}) \right] \tag{3.3.33}
\]

where \( \mu_i^{(j)} \) and \( \Sigma_i^{(j)} \) denote the statistical parameters for segment \( j \), frame \( k \).

The detailed implementation of obtaining \( f^*_{[k+1]} \) is shown as follows:

1. Preparing \( f^*_{[k]}, x_{[k]}, x_{[k+1]}, \mu_{[k]} = \{ \mu_{[k]}^{(1)}, \ldots, \mu_{[k]}^{(m)} \} \) and \( \Sigma_{[k]} = \{ \Sigma_{[k]}^{(1)}, \ldots, \Sigma_{[k]}^{(m)} \} \) as inputs.

2. Finding the site set \( S_{[k],[k+1]} \) by Eq. (3.3.28) and (3.3.29).

3. Minimizing the objective function Eq. (3.3.31) by ICM so as to update the label value of sites located in \( S_{[k],[k+1]} \). Moreover, the label values of the other sites are inherited from the \( f^*_{[k]} \).
4. Let $k = k + 1$, go to step 1.

In application, the initial reference label map $f^*_{[i]}$ is obtained by the method described in Section 3.2. A sample of tracking result is shown in Fig. 3.7.

![Tracking Result](image_url)

Figure 3.7: The first row is some samples in the input image sequence. Since it is hard to find the difference between the nearby two frames, the 1st, 5th, 10th, 15th, 20th, 25th and 30th frames are selected to illustrate the tracking result. The second row represents the corresponding tracking results.

### 3.3.4 Postprocessing

By using the methods stated above, the rough lip segmentation and tracking results can be obtained. Then, employ a postprocessing to refine the results based upon the prior knowledge of the lip shape.

Let the index of lip segment layer be denoted by $j^{lip}$, and the lip membership set be:

$$f^{lip} = \{f_{i}^{(j^{lip})} | i \in S\}. \quad (3.3.34)$$

For the convenience of description, the sites set $S$ is mapped into a 2-D coordinate set $\{(p, q) | 1 < p \leq c, 1 < q \leq r\}$ by $i = (q - 1) \cdot r + p$. Thus, $f^{lip}$, as illustrated in Fig. 3.8(b), can be rewritten as

$$f^{lip} = \{f_{(p,q)}^{(j^{lip})} | 1 < p \leq c, 1 < q \leq r\}. \quad (3.3.35)$$

Suppose that the lip region is not connected to the border of image. The morphological reconstruction based method in [152] is therefore employed to suppress border connected noisy structures as shown in Fig. 3.8(c). For the positive elements in reconstruction result, a threshold selection method proposed in [151] is utilized.
Figure 3.8: (a) The lip patch used to estimate the mean of $x_i$'s fallen into the true lip region; (b) The illustration of $f^{lip}$; (c) The result of border connected noisy structures suppressing; (d) The binary version of (c); and (e) The result of morphological opening. In (f), the dark gray ellipse is defined by the eigenvectors and eigenvalues of the covariance matrix of $M$. The light gray macro block is noise which should be masked out. (g) The final extraction result.

to bring a binary version of $f^{lip}$ as shown in Fig. 3.8(d). Furthermore, utilize the morphological opening with $3 \times 3$ structuring element and the result is denoted by $b^{lip}$ as illustrated in Fig. 3.8(e) with

$$b^{lip} = \{b_{(p,q)}^{(j^{lip})} \mid 1 < p \leq c, 1 < q \leq r \}. \quad (3.3.36)$$

For the foreground elements in $b^{lip}$, the corresponding positions

$$\{(p, q) \mid b_{(p,q)}^{(j^{lip})} = 1, 1 < p \leq c, 1 < q \leq r \} \quad (3.3.37)$$

are recorded and compose a matrix $M$ as follows:
\[
M = \begin{bmatrix}
q_1 & p_1 \\
q_2 & p_2 \\
\vdots & \vdots \\
q_l & p_l 
\end{bmatrix}
\] (3.3.38)

where \( l \) is the number of foreground elements in \( b_{iwp} \).

Then, the eigenvectors and eigenvalues of the covariance matrix of \( M \) can be calculated. An ellipse can be obtained whose position and inclination are defined by the eigenvectors, and the length of major/minor axis are defined by the 1.5 times square root of eigenvalues, respectively. Consequently, two horizontal lines crossing the highest and lowest points of the ellipse are obtained. The continued objects on the outside of the two lines are masked out as shown in Fig. 3.8(f).

Finally, given the prior knowledge of human mouth shape, the quickhull algorithm proposed in [153] is employed to draw the boundary of lip as illustrated in Fig. 3.8(g).

### 3.4 Experimental Results

#### 3.4.1 Database and Initialization

To show the performance of the proposed approach under different capture environments, i.e. illumination condition, color temperature, and skin color of testers, three databases were utilized:

1. AR face database,
2. CVL face database,
3. GTAV face database, and
4. VidTIMIT face database.

For each image, the part of face between nostril and chin is clipped by a \( 128 \times 128 \) (AR and CVL) or \( 64 \times 64 \) (GTAV) pixels window as the source of segmentation.
Figure 3.9: Sample images from AR database are in: (a) Group 1, (b) Group 2, (c) Group 3, and (d) Group 4.

experiment. For the images from VidTIMIT database, a $128 \times 80$ pixels window is utilized to clip the interesting region whose top left corner is posed in the coordinate $(240, 180)$ in source image (the origin is the top left corner).

In the following experiments, $\hat{m}$ pixels in source image were selected randomly and utilized their observation values to initialize mean vectors, i.e.

$$\mu^{(j)} = x_{\text{rand}(S)}, \quad j \in L,$$

where $\text{rand}(S)$ denotes a number which is selected from the set $S$ randomly.

Meanwhile, the covariance matrixes are initialized at

$$\Sigma^{(j)} = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}, \quad j \in L,$$

Moreover, the two parameters $\alpha$ and $\eta$ are set at 2 and 0.001, respectively.
3.4.2 Experiment 1

To demonstrate the robustness of the proposed method to true number of clusters \( m^* \) and the pre-defined number of clusters \( \hat{m} \), all source images from AR database were divided into four groups based on the different appearances. The details are:

- **Group 1.** People have no evident mustache with the mouth closed as shown in Fig. 3.9(a).

- **Group 2.** People have evident mustache with the mouth closed as illustrated in Fig. 3.9(b).

- **Group 3.** People have no evident mustache with the mouth opened as shown in Fig. 3.9(c).

- **Group 4.** People have evident mustache with the mouth opened as shown in Fig. 3.9(d).

From the viewpoint of clustering in color space, it is believed that the true number of clusters \( m^* \) of the images fall into same group are uniform. For example, in **Group 1**, there are skin and lip regions only because the face image is captured with mouth closed and no mustache. Hence, the segment number should be 2. The experiments were conducted on each group of data, respectively. For each group, 20 images were randomly selected as the input, and manually segmented the lip to serve as the ground truth. Two measures defined in [91] are used to evaluate the performance of the algorithms. The first measure

\[
OL = \frac{2(A_1 \cap A_2)}{A_1 + A_2} \times 100\%
\]  

(3.4.41)

determines the percentage of overlap between the segmented lip region \( A_1 \) and the ground truth \( A_2 \). The second measure is the segmentation error (SE) defined as

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

(3.4.42)

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 3.1: Overlap rate of segmented lips over $m$ in the four groups of data, in which some cells are marked in "-" because the experiments with $\hat{m} < m^*$ are not available.

The experiments were repeated with $\hat{m} = 2, 3, \ldots, 12$. Table 3.1 and Table 3.2 list the average $OL$ and $SE$ on the different image groups and $m$. It can be seen that the segmentation performance of the proposed approach is robust to $m^*$ and $\hat{m}$ in this experiment.

### 3.4.3 Experiment 2

To evaluate the performance of the proposed method under the different capture environments, i.e. illumination condition, color temperature, and complexion, 50 images were selected from AR, CVL and GTAV database, respectively. Specifically, the images from GTAV database was captured under three different illuminations, i.e. environment or natural light, strong light source from an angle of 45°, and finally an almost frontal mid-strong light source. Moreover, for each image, $m$ was randomly
<table>
<thead>
<tr>
<th>Group</th>
<th>ı̂m</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9.64%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>7.40%</td>
<td>8.31%</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>6.36%</td>
<td>8.30%</td>
<td>8.98%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>7.74%</td>
<td>8.29%</td>
<td>9.03%</td>
<td>12.27%</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>7.53%</td>
<td>8.50%</td>
<td>10.10%</td>
<td>13.56%</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>7.68%</td>
<td>8.02%</td>
<td>9.35%</td>
<td>10.27%</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>8.13%</td>
<td>8.47%</td>
<td>25.35%</td>
<td>13.23%</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>7.60%</td>
<td>8.70%</td>
<td>10.05%</td>
<td>13.50%</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>8.35%</td>
<td>8.51%</td>
<td>9.31%</td>
<td>13.59%</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>8.13%</td>
<td>8.71%</td>
<td>10.21%</td>
<td>12.42%</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>8.02%</td>
<td>11.05%</td>
<td>9.14%</td>
<td>11.90%</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.2: Segment error of segmented lips over ı̂m in the four groups of data, in which some cells are marked in ‘-’ because the experiments with ı̂m < m* are not available.

Figure 3.10: (a) is a sample of the input image for Wang07 in Experiment 3, (b) and (c) are the corresponding segmentation results obtained by Wang07 and the proposed method, respectively.
assigned an integer from the interval $[6, 24]$ to conduct lip segmentation. For images from either database, the average $OL$ and $SE$ were calculated. A snapshot of lip segmentation results is shown in Fig. 3.12. The right two columns in Table 3.3 lists the average $OL$ and $SE$ values obtained by the proposed method. The detail results are available in Fig. 3.13. It can be seen that the proposed approach is robust to the capture environment.

### 3.4.4 Experiment 3

This experiment compared the proposed approach with a existing counterpart, denoted as Wang07 [95] which focuses on lip segmentation in color space with taking the spatial restriction into account and can work with unknown true number of clusters as well. Implementing this algorithm, the same data set was utilized as Experiment 2 and randomly assigned $m$ an integer from the interval $[6, 24]$ to conduct lip segmentation as well. Table 3.3 shows the experimental results. Compared to Wang07, the proposed method has the much smaller segmentation error, while having the competitive advantage on AR and GTAV databases in terms of $OL$. Further, when implementing Wang07 in comparative studies, the size of image clips was utilized as suggested in [95] rather than the other size, e.g. $128 \times 128$ or $64 \times 64$. In this case, as shown in Fig. 3.10, Wang07 works quite well as well as the proposed method. Nevertheless, our empirical studies have also found that the performance of
Figure 3.12: A Snapshot of lip segmentation results, in which the first four rows are the results obtained with the image from Group 1-4, respectively. The images of the last four rows are from CVL and GTAV databases.
Figure 3.13: (a) and (b) are the frequency of specific $OL$ and $SE$ of the selected images from AR database, while (c) and (d) are from CVL, and (e) and (f) are from GTAV.
<table>
<thead>
<tr>
<th>Database</th>
<th>Measure</th>
<th>Wang07 OL</th>
<th>Wang07 SE</th>
<th>Proposed Method OL</th>
<th>Proposed Method SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>OL</td>
<td>87.35%</td>
<td>23.13%</td>
<td>90.30%</td>
<td>7.88%</td>
</tr>
<tr>
<td>CVL</td>
<td>OL</td>
<td>93.53%</td>
<td>10.05%</td>
<td>91.30%</td>
<td>7.26%</td>
</tr>
<tr>
<td>GTAV</td>
<td>OL</td>
<td>79.30%</td>
<td>25.18%</td>
<td>91.05%</td>
<td>7.38%</td>
</tr>
</tbody>
</table>

Table 3.3: Average overlap and segmentation error for the images from AR, CVL, and GTAV databases obtained by Wang07 and the proposed method, respectively.

Wang07 somewhat depends on the size of image clips. For example, if the $128 \times 128$ was utilized for AR and CVL, or $64 \times 64$ for GTAV as the inputs of Wang07, the segmentation results given by Wang07 become deteriorate as shown in Fig. 3.11, where Fig. 3.11(a) is the same image, which is ID02_030 in GTAV, as Fig. 3.10(a). In contrast, the proposed method gives the correct result as shown in Fig. 3.11(c). This implies that the proposed algorithm has more robust performance.

### 3.4.5 Experiment 4

A comparison study was conducted with an existing method denoted by Lievin04 [8]. This counterpart focus on lip segmentation and tracking in color space under MAP-MRF framework as well. We randomly select 10 video from the VidTIMIT database and conduct the lip segmentation experiment on each video, respectively.

Tab. 3.4 shows the average OL and SE obtained by Lievin04 and the proposed method. It can be seen that the proposed method outperforms Lievin04 in all cases we have tried so far.

### 3.4.6 Discussion

Although the proposed method shows the promising result, there are also some limitations which are discussed in this subsection.
Table 3.4: Average overlap and segmentation error for the images from VidTIMIT databases obtained by Lievin04 and the proposed method, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Lievin04</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>OL</td>
<td>87.21%</td>
<td>91.13%</td>
</tr>
<tr>
<td>SE</td>
<td>21.03%</td>
<td>7.75%</td>
</tr>
</tbody>
</table>

Figure 3.14: A sample of lip segmentation result obtained by the proposed method, in which the outer and inner of lip can be extracted smoothly.

**Tongue Misclassification**

In most cases, the proposed method can obtain both outer and inner lip boundary, e.g. see Fig. 3.14. Nevertheless, if the tongue has the similar color to lip, the tongue may be classified into the lip cluster as shown in Fig. 3.15. Actually, the similar problems are also occurred in Wang07 and Levein04. From the theoretical viewpoint, this kind of problem is hard to circumvent by a color analysis based method. In the future work, we shall explore to integrate lip modeling with the proposed method to deal with this problem.

**Global Features**

The proposed method only consider the local feature, i.e. spatial restriction in neighboring pixels, but ignore the global feature, e.g. the shape of lip. Thus, in case a lip has abnormal color, which maybe caused by illumination, some parts of lip may be missed out by the proposed segmentation method. This problem is
Figure 3.15: Two samples of tongue misclassification: (a) and (c) are the source image; (b) and (d) are the corresponding segmentation result. In the two samples, the tongue regions are classified into the lip segment.

Figure 3.16: Some parts of the lip in source image: (a) Abnormal image caused by the light reflection; (b) The segmentation result obtained by the proposed method, in which the color abnormal regions are missed out; (c) is the segmentation result obtained by Wang07, in which the missed parts are relative small because the prior shape information of lip is considered and integrated into the clustering.
observed in less than 3% of the test set. An example is shown in Fig. 3.16 (a) and (b). Although the post-processing can partially make up this problem, it is still one issue the proposed method can be further improved. Actually, Wang07 and [94] have shown a promising way to reduce this error by using the global shape feature (e.g. see Fig. 3.16(c)), which provides a direction for our future work.

3.5 Summary

Formulating the segmentation problem into a labeling optimization problem under MAP-MRF framework, this chapter has presented a rival penalized iterative algorithm to perform the segment clusters and over-segmentation elimination simultaneously. Based upon this algorithm, a lip segmentation and tracking scheme have been proposed, featuring the robust performance to the pre-assigned number of segment clusters. Experimental results have shown the efficacy of the proposed method in comparison with the existing two counterparts.
Chapter 4

Fuzzy Clustering Based Lip Segmentation with Unknown Segment Number

4.1 Introduction

This chapter will present a fuzzy clustering based segmentation method with objective function derived from the classical PE and implemented using Havrda-Charvat’s structural $\alpha$-entropy. This objective function features that the coincident cluster centroids in pattern space (also called input space interchangeably) can be equivalently substituted by one centroid with the function value unchanged. It is shown that the minimum of the proposed objective function can be obtained provided that: (1) the number of positions occupied by centroids (the intuitive relationship between position and centroid is shown in Fig. 4.1) in pattern space is the same as the truth number of clusters; and (2) these positions are coincident with the optimal cluster centroids obtained under the PE criterion. Thus, the optimal partition can be acquired by minimizing the proposed objective function regardless of whatever the pre-assigned cluster number is as long as it is greater than or equal to the ground truth. From the practical viewpoint, it is generally feasible to estimate...
the range of cluster number. In implementation, some cluster centroids are defined (i.e. the learnable data points in the input space towards the cluster centers), whose number is greater than or equal to the ground truth, and initialize them randomly. Subsequently, an iterative algorithm is utilized to minimize the proposed objective function. For each time step, not only is the centroid with the maximum membership degree, so called “winner”, updated to adapt to the corresponding input data, but also the other centroids are adjusted with a specific cooperation strength so that they are each closer to the winner. Subsequently, some neighboring centroids will be gradually merged into one so that the over-partition caused by redundant centroids can be eventually faded out. That is, the clustering performance of the proposed algorithm is robust against the pre-assigned cluster number. Based upon the proposed algorithm, a lip segmentation scheme is presented which is robust against the visibility of moustache, teeth and tongue. Experiments have shown the efficiency of the proposed approach.

The remainder of this chapter is organized as follows: Section 4.2 describes our method in detail. Section 4.3 presents the unsupervised lip segmentation scheme based upon the proposed method. Section 4.4 shows the experimental results. Lastly, this method is summarized in Section 4.5.

### 4.2 Fuzzy Clustering With Unknown Cluster Number

As stated in the Section 1.3.2, $H(C|X)$ is a classical criterion for fuzzy clustering, which, however, depends on the number of centroids. Although the optimal partition can be achieved under this criterion, the over- or under-segmentation almost always occurs if the number of centroids is not pre-assigned appropriately. This section proposes a variant of $H(C|X)$ which depends on the number of positions occupied by centroids instead of the number of centroids. Moreover, the proposed one inherits the property of $H(C|X)$. That is, when the number of positions occupied by centroids
Figure 4.1: The relationship between position and centroid, where the plane with grid represents the pattern space. As shown in Sub-figure (a), there are 6 centroids (denoted by circles) with the coordinate (2, 4), (2, 4), (2, 4), (5, 2), (5, 2), (7, 5), but they only occupy 3 positions (2, 4), (5, 2), (7, 5) as shown in Sub-figure (b).
is equal to the true cluster number, the proposed objective function reaches the minimum value. In the following, this method will be presented in detail.

4.2.1 The Proposed Objective Function for Fuzzy Clustering

Given an observation data \(x_i\) and the centroid collection \(C\), by adjusting the order of the elements in \(C\), we can obtain

\[
\tilde{C}_i = \{\tilde{c}_1^i, \tilde{c}_2^i, \ldots, \tilde{c}_m^i\}
\]  

satisfying \(p(\tilde{c}_k|x_i) \leq p(\tilde{c}_j|x_i)\) if and only if \(k < j\) \((j, k = 1, 2, \ldots, m)\).

Derived from Eq. (2.2.23), the objective function can be written as:

\[
\delta H(C \mid X) = 1 - \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{m} \left[ \frac{p(\tilde{c}_j|x_i) - p(\tilde{c}_{j-1}|x_i)}{p(\tilde{c}_m|x_i)} \right]^2,
\]  

(4.2.2)

where \(p(\tilde{c}_0|x_i) = 0\).

Since \(p(\tilde{c}_j|x_i)\) can be calculated by

\[
p(\tilde{c}_j|x_i) = \frac{1}{\sum_{k=1}^{m} \frac{\|x_i - \tilde{c}_k^i\|^2}{\|x_i - \tilde{c}_k^i\|^2}}^2,
\]  

(4.2.3)

Eq. (4.2.2) can be further rewritten as

\[
\delta H(C \mid X) = 1 - \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{m} \left[ \frac{\|x_i - \tilde{c}_m^i\|^2}{\|x_i - \tilde{c}_j^i\|^2} - \frac{\|x_i - \tilde{c}_j^i - \tilde{c}_{j-1}^i\|^2}{\|x_i - \tilde{c}_j^i - \tilde{c}_{j-1}^i\|^2} \right]^2.
\]  

(4.2.4)

The basic property of this objective function is shown as follows:

Theorem 4.2.1. Given a centroid collection \(C = \{c_1, c_2, \ldots, c_m\}\), a new centroid collection obtained by adding an element into \(C\) is denoted as \(C' = \{c_1, c_2, \ldots, c_m, c'\}\). We have \(\delta H(C \mid X) = \delta H(C' \mid X)\) if there exists \(c_j \in C\) \((j \in [1, m])\) satisfying \(c_j = c'\).

Proof. For specific \(x_i\), \(C\) and \(C'\) can be written as the ordered forms (see Eq. 4.2.1), i.e. \(\tilde{C}_i = \{\tilde{c}_1^i, \ldots, \tilde{c}_m^i\}\) and \(\tilde{C}'_i = \{\tilde{c}_1^i, \ldots, \tilde{c}_{m+1}^i\}\), respectively.
In $\tilde{C}_i'$, assume the corresponding element of $c_j$ is $\tilde{c}_k^i$ ($k \in [1, m]$). Since $c_j = c'$, the corresponding element of $c'$ can be written by $\tilde{c}_{k+1}^i$.

Thus,

$$\delta H(C' \mid X) = 1 - \frac{1}{s} \sum_{i=1}^{s} \left\{ \left[ \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_1^i \|^2} \right]^2 + \cdots + \left[ \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_{k+1}^i \|^2} - \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_k^i \|^2} \right]^2 + \cdots + \left[ \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_{m+1}^i \|^2} - \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_m^i \|^2} \right]^2 \right\} \right\} \right\}$$

As $\tilde{c}_k^i = \tilde{c}_{k+1}^i$, we have

$$\left[ \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_{k+1}^i \|^2} - \frac{\| x_i - \tilde{c}_{m+1}^i \|^2}{\| x_i - \tilde{c}_k^i \|^2} \right]^2 = 0.$$  \hspace{1cm} (4.2.6)

Moreover, as $\tilde{C}_i' \setminus \tilde{C}_i = \tilde{c}_{k+1}^i$, we have

$$\tilde{c}_l^i = \begin{cases} \tilde{c}_l^i, & l = 1, 2, \ldots, j, \\ \tilde{c}_{l-1}^i, & l = k + 2, k + 3, \ldots, m + 1. \end{cases} \hspace{1cm} (4.2.7)$$

Thus, Eq.(4.2.5) can be written as

$$\delta H(C' \mid X) = 1 - \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{m} \left[ \frac{\| x_i - \tilde{c}_{m}^i \|^2}{\| x_i - \tilde{c}_j^i \|^2} - \frac{\| x_i - \tilde{c}_{m}^i \|^2}{\| x_i - \tilde{c}_{j-1}^i \|^2} \right]^2$$

$$= \delta H(C \mid X). \hspace{1cm} (4.2.8)$$

According to Theorem 4.2.1, the value of Eq.(4.2.2) depends on the number of the positions of centroids in pattern space but not the number $m$ of centroids. For the sake of description, two functions named $PNum(C)$ and $ENum(C)$ are defined. The former one returns the number of positions of the centroids in $C$, the latter one returns the number of positions of the centroids in $C'$. The following section will discuss the properties of these functions.
and the latter one returns the number of centroids in \( C \). Moreover, \( \text{Pos}(C) \) is employed to obtain a collection composed by the positions of the centroids in \( C \), where \( \text{Pos}(C) = \{p_1, p_2, \cdots, p_{\text{PNum}(C)}\} \). Further, let
\[
H(C \mid x_i) = 1 - \sum_{j=1}^{m} p^2(c_j \mid x_i),
\]
(4.2.9)
and
\[
\delta H(C \mid x_i) = 1 - \sum_{j=1}^{m} \left[ \frac{p(\tilde{c}_j \mid x_i) - p(\tilde{c}_{j-1} \mid x_i)}{p(\tilde{c}_m \mid x_i)} \right]^2.
\]
(4.2.10)
Then, we have the following lemma:

**Lemma 4.2.2.** Given an input \( x \) and a constant \( \eta_m \in (0, \frac{m-1}{m}] \), the minimum of \( H(C \mid x) \) is approximately equal to \( \eta_m \) subject to \( \delta H(C \mid x) = \eta_m \), where \( \text{PNum}(C) = \text{ Enum}(C) = m \).

The detailed proof of Lemma 4.2.2 is given in Appendix I, and the experimental justification is shown in Appendix II.

Based on Lemma 4.2.2, the result can be written as follows:

**Theorem 4.2.3.** Given an input \( x \), there exist two centroid collections \( C_1 \) and \( C_2 \) satisfying \( C_1 = \text{Pos}(C_2) \) such that \( H(C_1 \mid x) \) reaches the minimum value approximately when \( \delta H(C_2 \mid x) \) reaches minimum as well.

**Proof.** Utilize the notation \( C \) to represent the solution space of \( \delta H(C \mid x) = \vartheta \), where \( \vartheta \) denotes the global minimum value of \( \delta H(C \mid x) \).

According to Lemma 4.2.2, the minimum value of \( H(C \mid x) \) with \( C \in C \) is approximately equal to \( \vartheta \) as well. The corresponding centroid collection is denoted as \( C_1 \) with \( \text{ Enum}(C_1) = \text{PNum}(C_1) \).

Suppose \( H(C \mid x) \) can reach the global minimum value \( \theta \) when \( C = C_1' \). Based on Lemma 4.2.2, we have \( \delta H(C_1' \mid x) \approx \theta \), i.e. \( \delta H(C_1' \mid x) < \vartheta \).

However, \( \vartheta \) is the global minimum value of \( \delta H(C \mid x) \). Hence, the supposition is false and \( \vartheta \) can be viewed as the global minimum value of \( H(C \mid x) \), which is reached as \( C = C_1 \).

According to Theorem 4.2.1, there exists the centroid collection \( C_2 \) with \( C_1 = \text{Pos}(C_2) \) such that \( \delta H(C_2 \mid x) = \delta H(C_1 \mid x) \).
Recall the property of PE presented in paper [140], i.e. see Eq.(2.2.19), the proposed objective function $\delta H(C \mid X)$ will reach the minimum value provided that:

1. the number centroid positions in pattern space is equal to the true cluster number, i.e. $PNum(C) = ENum(C^*)$, and
2. the positions are coincident with the optimal centroids under the PE criterion, i.e. $Pos(C) = C^*$,

where $C^*$ is the centroid collection obtained by minimizing Eq. (2.2.20) with $m = m^*$.

### 4.2.2 The Iterative Algorithm

This sub-section will present an iterative algorithm to perform fuzzy clustering by minimizing the proposed objective function shown in Eq.(4.2.2). At each iterative step, based upon the idea of “cooperative learning” initially proposed in paper [154], this algorithm not only updates the winner centroid to adapt to the corresponding input data, but also the other centroids are adjusted with a specific cooperation strength so that they are each close to the winner. Subsequently, the initial over-partition will be gradually faded out with the redundant centroids superposed over the convergence of the algorithm.

Specifically, the segment number $m$ is pre-assigned a value which is greater than or equal to the ground truth, and initialize the centroid collection $C$ randomly. Then, the subsequent implementation is given as follows:

**Step 1.** Fixing $C$, calculate $p(c_j \mid x_i)$ and obtain the collection $\tilde{C}$ by Eq.(4.2.3) and Eq. (4.2.10) for each $x_i$.

**Step 2.** For each input data $x_i$, update $C$ via

$$c^{(new)}_j = c^{(old)}_j - \eta_w \cdot \frac{\partial \delta H(C \mid x_i)}{\partial c_j} c^{(old)}_j ,$$

(4.2.11)
and
\[ c_{ji}^{(new)} = c_{ji}^{(old)} - \eta_r \cdot \frac{\partial H_m(C|x_i)}{\partial c_{ji}} \bigg|_{c_{ji}^{(old)}} \] (4.2.12)

where \( c_{ji} \) is the “winner” centroid with \( j_i^w = \arg \max_j \frac{p(\tilde{c}_j|x_i) - p(\tilde{c}_{j-1}|x_i)}{p(\tilde{c}_m|x_i)} \), \( j_i^r = 1, 2, \ldots, m \) but \( j_i^r \neq j_i^w \), \( \eta_w, \eta_r \) are the positive learning rates.

The above two steps are processed iteratively until \( C \) converges. An example of the clustering results with redundant centroids obtained by the FCM (i.e. use Eq.(4.2.11) only in Step 2) and the proposed method is shown in Fig. (4.2).

![Figure 4.2](image)

(a) The initial positions of centroids marked by ‘‘∗’’; (b) The clustering result by using the FCM, which falls into the local minimum; (c) The clustering result obtained from the proposed approach, where all redundant centroids are coincided.
4.3 Lip Segmentation and Postprocessing

This section applies the proposed method in Section 4.2 to the unsupervised lip segmentation problem. The task is to extract the lip boundary from a color image consisting of a part of face between nostril and chin. A sample of original image is shown in Fig. 4.3 (a).

The pattern space which this method performs in is also the modified \(H\)SV color space. The definition of \(x_i\) can be found in Eq. (3.3.26).

Subsequently, the centroid collection \(C\) is calculated via the proposed method as introduced in Section 4.2. Utilize the following equation to obtain the hard
Figure 4.4: The lip segmentation result after redundant cluster centroids have been merged.

segmentation result:

\[ S^{(j)} = \{ i \mid j = \arg \max_j \frac{p(\hat{c}_j \mid x_i) - p(\hat{c}_{j-1} \mid x_i)}{p(\hat{c}_m \mid x_i)}, 1 \leq i \leq n, 1 \leq j \leq m \}, \]  

(4.3.13)

where \( S^{(j)} \) denotes the set of data falling into cluster \( j \).

A sample of \( S^{(j)} \)'s with \( m = 9 \) is shown in Fig. 4.3 (b) - (j). Obviously, the site sets: \( \{ S^{(1)}, S^{(2)}, S^{(3)} \} \), \( \{ S^{(4)}, S^{(5)}, S^{(6)} \} \), and \( \{ S^{(7)}, S^{(8)} \} \) are similar because the corresponding redundant centroids are coincident in pattern space.

Then, for any two centroids \( c_j \) and \( c_k \) in pattern space, if

\[ \|c_j - c_k\| \leq \varepsilon, \]  

(4.3.14)

where \( \varepsilon \) is a small threshold value, they can be replaced by a new centroid \( c_l \)

\[ c_l = \frac{c_j + c_k}{2}. \]  

(4.3.15)

Thus, in the example shown in Fig. 4.3, the number of centroids is reduced from 9 to 4. The new clustering result after centroids merged is shown in Fig. 4.4.

Then, Utilize the method proposed in [65] to extract a patch of lip region. Then, the mean of \( x_i \)'s restricted by the patch is calculated and denoted by \( \hat{\mu} \). It is regarded as an estimate of the mean of \( x_i \)'s which fall into the true lip region. To save space, interested readers may refer to [65] for more details about this method. The major steps of this method are summarized as follows:

**Step 1:** Transform source lip image into 1976 CIELAB color space. \( a^* \) component
Figure 4.5: (a) is the lip patch which is used to estimate the mean of $x_i$'s falling into the true lip region, (b) is the result of border connected structures clearing, (c) is the result of morphological filter (closing with $5 \times 5$ structuring element and opening with $3 \times 3$ structuring element). In (d), the shape of gray ellipse is defined by the eigenvectors and eigenvalues of the covariance matrix of $P$. The continued objects on the outside of this ellipse are viewed as noises and masked out. (e) is the final extraction result obtained via quickhull algorithm.

for site $i$ is mapped into the range of $[0, 255]$ via histogram equalization and denoted by $a^*_i$. Meanwhile, utilize the equation

$$U_i = \begin{cases} 
256 \times \frac{G_i}{R_i}, & R_i > G_i, \\
255, & \text{otherwise},
\end{cases} \quad (4.3.16)$$

proposed in [8] to calculate $U$ component for each pixel, where $R_i$ and $G_i$ denote the red and green component for site $i$ in a source lip image, respectively.

**Step 2:** Let $B_i = a^*_i - U_i$, a Gaussian model can be established for the positive $B_i$s with the mean $\mu_B$ and the standard deviation $\sigma_B$. The following
The equation is employed to binarize the source lip image:

$$\hat{B}_i = \begin{cases} 
0, & B_i \leq \mu_B - 2\sigma_B, \\
1, & \text{otherwise}.
\end{cases} \quad (4.3.17)$$

**Step 3:** Consider the site set $$\hat{S} = \{i \mid \hat{B}_i = 1, 1 \leq i \leq n\}$$ as a lip patch (see Fig. 4.5 (a)), $$\hat{\mu}$$ can be calculated by

$$\hat{\mu} = \frac{1}{s} \sum_{i \in \hat{S}} x_i, \quad (4.3.18)$$

where $$s$$ denotes the number of elements in the set $$\hat{S}$$.

Thus, the index of lip segment layer can be determined by

$$j^{lip} = \arg \min_j \|c_j - \hat{\mu}\|, \quad 1 \leq j \leq m. \quad (4.3.19)$$

The site set corresponding to lip segment is denoted as $$S^{(j^{lip})}$$.

Finally, post-processing is employed to extract the lip region accurately. The detail can be found in Section 3.3.4.

### 4.4 Experimental Results

#### 4.4.1 Database and Initialization

To show the performance of the proposed approach, 3 databases were utilized:

1. Fisher’s iris database [155]. This data set contains 3 classes of 50 instances each, where each instance has 4 attributes.

2. AR face database, and

3. CVL face database.
For each image in AR and CVL database, the part of face between nostril and chin is clipped by a $128 \times 128$ pixels window as the source of segmentation experiment.

Moreover, in the following experiments, this equation was utilized to initialize the centroids

$$c_j = x_{rand(i)}, \quad 1 \leq j \leq m,$$

(4.4.20)

where $rand(i)$ denotes a number selected from the range of $i$, i.e. $\{1, 2, \cdots n\}$, randomly. Further, let $\varepsilon = 0.5$, $\eta_w = 0.01$, and $\eta_r = 0.001$.

4.4.2 Experiment 1

In this experiment, the method described in Section 4.2 was employed to perform the fuzzy clustering in Fisher’s iris database. This experiment was conducted with $\hat{m} = 5, 6, \cdots, 15$, respectively. Moreover, for each specific $\hat{m}$, the experiments were repeated 5 times with different initial $C$. After the redundant centroids merged (Eq. (4.3.14) and (4.3.15)), the histogram of final centroid number is shown in Fig. 4.6.

![Figure 4.6](image)

Figure 4.6: The final centroid number after the clustering performed in Fisher’s iris database by the proposed method with different $m$ and initial $C$.

Moreover, for these results with 3 centroids remained (as shown in Fig. 4.6, in our experiment, 47 out of 55 results have met this criterion), the estimation error
was calculated by using the following equation:

\[
error = \sqrt{\frac{\sum_{j=1}^{3} \|\hat{c}_j - c_j^*\|^2}{3}}
\]  

(4.4.21)

where \(\hat{c}_j\) is the final centroid obtained by our method, \(c_j^*\) is the corresponding centroid obtained by classical FCM method with the cluster number pre-defined by 3. The histogram or error is shown in Fig. 4.7.

![Histogram of error between cluster centroids](image)

Figure 4.7: This figure shows the histogram of the error between the cluster centroids obtained by our method and classical FCM (the cluster number is pre-defined by 3).

### 4.4.3 Experiment 2

To demonstrate the accuracy and robustness of the proposed method, as with Section 3.4.2, the source images from AR database were separated into the four groups as well. The experiment was conducted on each group, respectively. For each group, 20 images were randomly selected as the input, and manually segmented the lip to serve as the ground truth.

The experiments were repeated with \(\hat{m} = 5, 6, \ldots, 15\). Table 4.1 and Table 4.2 list the average \(OL\) and \(SE\) on the different image groups and \(m\). It can be seen that the segmentation performance of the proposed approach is robust to all \(\hat{m}\) utilized in the experiments.
Figure 4.8: (a) and (b) illustrate the histograms of $OL$ and $SE$ of the selected images from AR database. (c) and (d) illustrate the histograms of $OL$ and $SE$ of the selected images from CVL database.
To evaluate the performance of the proposed method under the different capture environments, 50 images were selected from AR and CVL database, respectively. Moreover, for each image, randomly assigned an integer from the set \( \{5, 6, \ldots, 15\} \) to conduct lip segmentation. For images from either AR or CVL database, the average \( OL \) and \( SE \) were calculated. The rightmost two columns in Table 4.3 list the average \( OL \) and \( SE \) values obtained by the proposed method. Fig. 4.8 shows the histograms of \( OL \) and \( SE \) for images from each database. Once again, it can be seen that the proposed approach is robust against the pre-assigned number of clusters.

### Experiment 3

Table 4.1: Overlap of segmented lips with the ground truth.

<table>
<thead>
<tr>
<th>( \hat{m} )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>91.5%</td>
<td>93.5%</td>
<td>85.4%</td>
<td>88.3%</td>
</tr>
<tr>
<td>6</td>
<td>92.3%</td>
<td>93.3%</td>
<td>87.6%</td>
<td>90.6%</td>
</tr>
<tr>
<td>7</td>
<td>90.7%</td>
<td>94.6%</td>
<td>85.0%</td>
<td>86.5%</td>
</tr>
<tr>
<td>8</td>
<td>94.5%</td>
<td>89.7%</td>
<td>88.2%</td>
<td>90.3%</td>
</tr>
<tr>
<td>9</td>
<td>91.0%</td>
<td>92.7%</td>
<td>90.3%</td>
<td>87.6%</td>
</tr>
<tr>
<td>10</td>
<td>91.3%</td>
<td>95.3%</td>
<td>91.5%</td>
<td>89.4%</td>
</tr>
<tr>
<td>11</td>
<td>93.0%</td>
<td>92.5%</td>
<td>88.4%</td>
<td>90.4%</td>
</tr>
<tr>
<td>12</td>
<td>90.7%</td>
<td>93.6%</td>
<td>87.9%</td>
<td>87.8%</td>
</tr>
<tr>
<td>13</td>
<td>92.8%</td>
<td>93.0%</td>
<td>88.1%</td>
<td>88.3%</td>
</tr>
<tr>
<td>14</td>
<td>92.3%</td>
<td>92.9%</td>
<td>94.2%</td>
<td>88.2%</td>
</tr>
<tr>
<td>15</td>
<td>90.4%</td>
<td>92.4%</td>
<td>88.3%</td>
<td>89.8%</td>
</tr>
</tbody>
</table>

75
4.4.5 Experiment 4

To demonstrate the performance of the proposed approach in comparison with the existing method denoted as Wang07 proposed in [95], which focuses on lip segmentation in color space by fuzzy clustering with unknown cluster number as well. The Wang07 algorithm was implemented on the same images utilized in Experiment 3. The experimental result is shown in Table 4.3.

4.5 Summary

This chapter has proposed a fuzzy clustering method, through which the fuzzy clustering can be performed without knowing the true cluster number in advance. This method features that the overlapped (or close) cluster centroids in pattern space can be merged into one from the viewpoint of objective function value. Then, an iterative algorithm is utilized to minimize the proposed objective function by
superposing the redundant centroids. At each time step, not only are the “winner” centroids updated to adapt to an input data, but also the other centroids are driven closed. Therefore, the clustering performance is robust against the pre-assigned number of clusters. Finally, based upon this method, a lip segmentation scheme has been presented, whose efficiency has been shown in the experiments.
Chapter 5

MAP-MRF Based Lip Segmentation with Automatic Scale Selection

5.1 Introduction

This chapter will propose an MRF based segmentation method with scale variation considered. Firstly, assuming that each pixel of a given image has its own optimal observation scale from the viewpoint of segmentation. Then, the scale values are seen as an MRF defined on the regular pixel lattice. Thus, the scale selection problem and segmentation problem can be formulated under the maximum an MAP-MRF framework jointly. To this end, we propose an MAP-MRF based objective function with respect to two variables: the label configuration and the scale configuration. In addition, this method is incorporated into the model presented in Chapter 3 under which the lip segmentation can be implemented without true segment number. The optimal scale and the corresponding segmentation result can be obtained by minimizing the objective function. Experimental results have shown the efficacy of the proposed method in comparison with the existing counterparts.

The remainder of this chapter is organized as follows: Section 5.2 describes the
proposed method in detail. Section 5.3 shows the experimental results. Lastly, we draw a summary in Section 5.4.

5.2 MRF Based Image Segmentation with Various Scales

In this section, the basic statistical model with various scales under the MAP-MRF framework is proposed firstly. Then, to perform the segmentation without true segment number which is an important issue in implementation, a modified model is presented based on our previous work. Finally, an iterative algorithm to optimize the proposed objective function is introduced.

5.2.1 Multi-Scale Model

As stated in Chapter 3, a given image can be represented by $x$, which denotes a configuration of the observation random field $\mathcal{X} = \{X_i = x_i | i \in S, x_i \in O\}$. For convenience of description, the sites set $S$ is mapped into a 2-D pixel coordinate $\{(px, py) | 1 < px \leq c, 1 < py \leq r\}$ temporarily. Thus, we have $x_i = x_{px,py}$ if $i = (py - 1) \cdot r + px$.

Based on the scale-space representation [156], an image pyramid can be obtained by:

$$\tilde{x}_{px,py}^{(k)} = g_k(px,py) \ast x_{px,py}$$

with

$$g_k(px,py) = \frac{1}{2\pi \cdot 2^k} \cdot exp(-\frac{p_x^2 + p_y^2}{2^{k+1}}),$$

where $\tilde{x}_{px,py}^{(k)}$ denotes the corresponding observed data of $x_{px,py}$ in layer $k$ of the pyramid.

We utilize the notation $\tilde{x}^{(k)}$ to represent the image posed in layer $k$ of the pyramid. Since the size of each $\tilde{x}^{(k)}$ is different, we further employ interpolation to normalize them to have the same size as the finest layer. The normalized configuration is denoted by $\hat{x}_{i}^{(k)} = \{x_{i}^{(k)} | i \in S\}$. Specifically, let $x^{(0)}$ be equal to $x$. An
illustration of $x^{(k)}(k = 0, 1, 2)$ is shown in Fig. 5.1. Since the values in neighboring interpolation nodes are not changed abruptly, the differences among the results obtained by various interpolation methods are small. In this research, the linear interpolation is utilized because of its low computational complexity.

Thus, the traditional MRF-based image representation can be extended into multi-scale space by using a hierarchical model. For arbitrary observation scale $k$ (i.e. the layer $k$ of the image pyramid), there is an observation random field $\mathcal{X}^{(k)} = \{X_i^{(k)} = x_i^{(k)} | i \in S, x_i^{(k)} \in O \}$. For each $\mathcal{X}^{(k)}$, there is a corresponding hidden label MRF named $\mathcal{F}^{(k)} = \{F_i^{(k)} = f_i^{(k)} | i \in S, f_i^{(k)} \in L \}$ to represent the segmentation result.

Suppose each pixel of a given image has an optimal local scale from the segmentation viewpoint. Given an image, we further assume that the local scale values compose an MRF

$$\mathcal{T} = \{T_i = t_i | i \in S, t_i \in D \}, \quad (5.2.3)$$

where $D = \{0, 1, ..., d - 1 \}$ is the scale value set, in which 0 refers to the finest scale while $d - 1$ means the coarsest scale. $t = \{t_i | i \in S, t_i \in D \}$ is the configuration of
Thus, the observation random filed and the corresponding hidden label MRF in the hierarchical model can be further rewritten as

\[ \mathcal{X} = \{ x_i^{(t_i)} \mid i \in S, t_i \in D, x_i^{(t_i)} \in O \} \]  

\[ \mathcal{F} = \{ f_i^{(t_i)} \mid i \in S, t_i \in D, f_i^{(t_i)} \in L \}, \]

respectively.

Accordingly, the optimization problem shown in Eq.(2.2.9) can be extended by adding the scale configuration as another variable directly:

\[ \{ f^*, t^* \} = \arg \max_{f \in \Omega_L, t \in \Omega_D} P(f, t \mid x), \]

where \( \Omega^D = D \times D \times ... \times D = D^s \) is the configuration space of the scale MRF \( \mathcal{T} \) on the site set \( S \), \( t \) denotes the configuration of scale, and \( t^* \) denotes the optimal scale selection result.

The corresponding energy function can be denoted by

\[ E(f, t; x) = U(t) + \beta U(f^{(t)}) + \gamma U(x^{(t)} \mid f^{(t)}) \]

where \( \beta \) and \( \gamma \) are weighting parameters. The details of the derivation process are given in Appendix I.

We suppose that the observed data, which belong to segment \( j \) and layer \( k \), follow a specific Gaussian distribution with the parameter \( \theta^{(k)}[j] = \{ \mu^{(k)}[j], \Sigma^{(k)}[j] \} \), where \( \mu^{(k)}[j] \) and \( \Sigma^{(k)}[j] \) denote the mean vector and covariance matrix, respectively. Since the content changes usually take place continuously in a video stream, we assume that the difference between \( \theta^{(0)}[j], \theta^{(1)}[j], ..., \theta^{(d-1)}[j] \) is neglectable. Thus, \( \theta^{(k)}[j] \) can degenerate to \( \theta[j] = \{ \mu[j], \Sigma[j] \} \), which is shared by all observed data that fall into the segmentation class \( j \). Thus, the notation of the likelihood energy term, \( U(x^{(t)} \mid f^{(t)}) \), can be replaced by \( U(x^{(t)} \mid f^{(t)}; \Theta) \), where \( \Theta = \{ \theta[j] \mid j \in L \} \).

Eq.(5.2.7) can be rewritten as

\[ E(f, t; x, \Theta) = U(t) + \beta U(f^{(t)}) + \gamma U(x^{(t)} \mid f^{(t)}; \Theta). \]  

81
Hereinafter, Eq.(5.2.8) is called the objective function. Since $T$ is an MRF, the prior energy term $U(t)$ in Eq.(5.2.8) can be specified as

$$U(t) = \sum_{c \in C} V_c(t) = \sum_{i \in S} \sum_{i' \in N_i} [1 - \delta(t_i - t_{i'})]$$

(5.2.9)

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

![Figure 5.2: Illustration of the 3-D first-order neighborhood system of a given site, in which the black circle is the given site and gray circles denote the corresponding neighborhood system.](image)

To specify the second term of Eq.(5.2.8), we introduce a 3-D neighborhood system. Recall the definition of the 2-D neighborhood system $N_i = \{i' \in S \mid |i - i'| \leq \text{rad}, i \neq i'\}$, which represents the spatial restriction between the neighboring sites in a 2-D MRF. Since the change of label in specific pixel caused by the minor observation scale variation cannot be abrupt or frequent, it is feasible to assume that the label MRFs on different pyramid layers $\{F^{(k)} \mid k \in D\}$ can be treated as a 3-D MRF defined on $S$. In this MRF, the value at site $i$ is not only dependent on the values at sites belonging to $N_i$, but also the values at neighboring sites in the next scale layers. The values in the neighboring sites of $i$ is $\{f^{(k)}_{i'} \mid i' \in N_i, k \in D, k' = k, k+1, k-1\}$. 

82
Fig. 5.2 depicts a 3-D neighborhood system.

Thus, the second term in Eq.(5.2.7) can be expressed by

\[
U(f^{(l)}) = \sum_{c \in C} V_c(f) = \sum_{i \in S} \sum_{i' \in \{i-1,i,i+1\}} \sum_{t' \in \{t^{(l)}_i-1,t^{(l)}_i,t^{(l)}_i+1\}} \left[ 1 - \delta(f^{(l)}_i - f^{(l)}_{i'}) \right].
\]

(5.2.10)

The likelihood term is specified as

\[
U(x^{(l)} | f^{(l)}; \Theta) = \sum_{i \in S} \left[ (x^{(l)}_i - \mu[f^{(l)}_i])^T (\Sigma[f^{(l)}_i])^{-1} (x^{(l)}_i - \mu[f^{(l)}_i]) \right],
\]

(5.2.11)

where \(T\) denotes the matrix transpose operation.

According to MAP theory, the optimal segment label map \(f^*\) and optimal local-scale map \(t^*\) can be obtained by solving the following equation:

\[
\{f^*, t^*\} = \arg \min_{f \in \Omega, t \in \Omega} \mathcal{E}(f, t; x; \Theta).
\]

(5.2.12)

### 5.2.2 Rival Penalty Iterative Algorithm

From the practical viewpoint, the true number of segments \(m^*\) in lip segmentation usually varies due to the appearance (e.g. moustache, teeth, and tongue) of speakers, and the estimation bias of \(m^*\) may induce segmentation error. Therefore, we present an algorithm featuring automatic selection of the number of segments to perform lip segmentation within the proposed model.

Suppose each pixel belongs to multiple segments with a different degree of membership when the corresponding observation scale is determined, the membership can be formulated as

\[
f^{(k)}_i[j] = \frac{G(x^{(k)}_i; \theta[j])}{\sum_{j'=0}^{m-1} G(x^{(k)}_i; \theta[j'])}, \quad i \in S, \quad j \in L, \quad k \in D
\]

(5.2.13)

where \(G(\cdot)\) denotes the Gaussian probability density function (p.d.f.), and \(k\) denotes the scale of the given observation. Given \(k\), there are \(m\) membership sets named \(\{f^{(k)}_i[0] | i \in S\}, \{f^{(k)}_i[1] | i \in S\}, \ldots, \{f^{(k)}_i[m-1] | i \in S\}\). Moreover, as \(k \in \{0, 1, \ldots, d-1\}\), the number of all possible membership sets is \(m \times d\). When the observation scale at site \(i\), denoted as \(t_i\), is determined, there are only \(m\) membership
sets left and denoted by

\[
\hat{f}_{i}^{(t_{i})}[j] = \frac{G(x_{i}^{(t_{i})}; \theta[j])}{\sum_{j' = 0}^{m-1} G(x_{i}^{(t_{i})}; \theta[j'])}, \quad i \in S, \ j \in L.
\]  

(5.2.14)

Figure 5.3: The work flow of the proposed method.

When each membership set is viewed as a random field defined on \( S \), they compose a hierarchical model. There are \( m \) layers in this model, and the random field posed in the \( j \)th layer can be represented by \( \{f_{i}^{(t_{i})}[j] \mid i \in S\} \). Then, the value of label MRF \( F \) is used for representing the segmentation result, which can be specified by

\[
f_{i} = \arg \max_{j \in L} f_{i}^{(t_{i})}[j].
\]  

(5.2.15)

According to Eq.(5.2.15), \( f_{i} \) can be regarded as a function with respect to \( f_{i}^{(t_{i})}[j] \). Thus, Eq.(5.2.10) can be regarded as a function with respect to the configuration.
\{ f_i^{(t_i)}[j] \mid i \in S, j \in L \}. For simplicity, we utilize the notation \( N_i^+ \) to represent the site set \( N_i \cup i \). We employ the following equation to approximate the map between \( U(f(t)) \) and \( \{ f_i^{(t_i)}[j] \mid i \in S, j \in L \} \):

\[
\hat{U}(f(t)) = \sum_{i \in S} \sum_{j \in L} \sum_{\nu' \in N_i^+} \frac{f_{i'}^{(t_{i'})}[j]}{\sum_{\nu'' \in N_i^+} f_{\nu''}^{(t_{\nu''})}[j]} \ln \frac{f_{i'}^{(t_{i'})}[j]}{\sum_{\nu'' \in N_i^+} f_{\nu''}^{(t_{\nu''})}[j]}.
\] (5.2.16)

Unlike the Kronecker delta function, Eq.(5.2.16) utilizes the information entropy-based disorder measurement in the local object system to represent the spatial coherence restriction.

We first initialize the parameter set \( \Theta = \{ \theta[j] \mid j \in L \} \), the configuration of label MRF \( f = \{ f_i \mid i \in S, f_i \in L \} \), and the configuration of observation scale MRF \( t = \{ t_i \mid i \in S, t_i \in D \} \). For each observed data \( x_i^{(t_i)} \), we utilize an iterative method to update \( \Theta \) in order to minimize the objective function shown in Eq.(5.2.7) while eliminating the redundant segment cluster(s). The proposed method is summarized in Fig. 5.3, and the detailed implementation is given as follows:

1. The proposed objective function shown in Eq.(5.2.7) becomes:

\[
\mathcal{E}(t; f, x, \Theta) = U(t) + \beta U(f(t)) + \gamma U(x(t) \mid f(t); \Theta).
\] (5.2.17)

when \( \Theta \) and \( f \) are fixed.

We find the scale configuration \( \hat{t} = \{ \hat{t}_i \mid i \in S \} \) which makes Eq.(5.2.17) minimum by ICM. Let \( t = \hat{t} \).

2. Fixing \( \Theta \) and \( t \), calculate \( \tilde{f}_i^{(t_i)}[j] \) for each site via Eq.(5.2.14).

3. Considering the prior energy term approximated by Eq.(5.2.16), we adjust \( \tilde{f}_i^{(t_i)}[j] \) to make the local systems \( \{ \tilde{f}_{i'}^{(t_{i'})}[j] \mid i' \in N_i^+ \} \) reach the highest uncertainty, which provides the spatial restriction, i.e.

\[
\tilde{f}_i^{(t_i)}[j] = \exp\left(-\frac{\sum_{\nu' \in N_i} \tilde{f}_{\nu'}^{(t_{\nu'})}[j] \ln \tilde{f}_{\nu'}^{(t_{\nu'})}[j]}{\sum_{\nu' \in N_i} f_{\nu'}^{(t_{\nu'})}[j]}\right).
\] (5.2.18)

Let \( f_i^{(t_i)}[j] = \tilde{f}_i^{(t_i)}[j] \).
4. Update $f_i$ for each site $i$ via Eq. (5.2.15).

5. Fixing $f$ and $t$, we rewrite Eq. (5.2.7) as

$$
E(\Theta ; f, t, x) = U(t) + \beta U(f^{(t)}) + \gamma U(\Theta ; x^{(t)}, f^{(t)}).
$$ (5.2.19)

Then, we update $\Theta$ for each $x_i$ via

$$
\dot{\theta}[j_w] = \theta[j_w] - \eta_w \frac{dE(\Theta ; f, t, x)}{d\theta} |_{\theta[j_w]},
$$ (5.2.20)

and

$$
\dot{\theta}[j_r] = \theta[j_r] + \eta_r \cdot \delta E_i^{(t)}[j_w, j_r] \cdot 1_{[0, +\infty)}(\delta E_i^{(t)}[j_w, j_r]) \cdot \frac{dE(\Theta ; f, t, x)}{d\theta} |_{\theta[j_r]},
$$ (5.2.21)

where $j_w = \text{arg max}_j f_i^{(t)}[j]$, $j_r \in L$ but $j_r \neq j_w$, $\eta_w$ and $\eta_r$ are two positive numbers which denote the learning rate and penalty rate, respectively. Let $\theta[j_w] = \dot{\theta}[j_w]$, $\theta[j_r] = \dot{\theta}[j_r]$, and $\delta E_i^{(t)}[j_w, j_r]$ represent the penalty strength, which is an entropy difference-based measurement of the similarity defined by:

$$
\delta E_i^{(t)}[j, j'] = E_i^{(t)}[j] - E_i^{(t)}[j, j']
$$ (5.2.22)

where

$$
E_i^{(t)}[j] = -\frac{1}{n} \sum_{v' \in N_i^+} \frac{f_v^{(t, r)}[j]}{\sum_{v'' \in N_i^+} f_{v''}^{(t, r)}[j]} \ln \frac{f_v^{(t, r)}[j]}{\sum_{v'' \in N_i^+} f_{v''}^{(t, r)}[j]},
$$

$$
i \in S, j \in L, t_i \in D,
$$ (5.2.23)

and

$$
E_i^{(t)}[j, j'] =
$$

$$
-\frac{1}{2n} \sum_{v' \in N_i^+} \left\{ \frac{f_v^{(t, r)}[j]}{F_i^{(t, r)}[j, j']} \ln \frac{f_v^{(t, r)}[j]}{F_i^{(t, r)}[j, j']} + \frac{f_v^{(t, r)}[j']}{\sum_{v'' \in N_i^+} f_{v''}^{(t, r)}[j] + f_{v''}^{(t, r)}[j']} \times \ln \frac{f_v^{(t, r)}[j']}{\sum_{v'' \in N_i^+} f_{v''}^{(t, r)}[j] + f_{v''}^{(t, r)}[j']} \right\},
$$

$$
i \in S, j \in L, j' \in L, t_i \in D \text{ and } j \neq j'.
$$ (5.2.24)
Moreover, \( 1_{[0,\infty)}(\delta E^{(t)}_{i}[j_w, j_r]) \) denotes the step function given by

\[
1_{[0,\infty)}(\delta E^{(t)}_{i}[j_w, j_r]) = \begin{cases} 
1, & \delta E^{(t)}_{i}[j_w, j_r] \in [0, +\infty), \\
0, & \text{otherwise}.
\end{cases}
\]

The above steps 1-5 are implemented iteratively until \( f \) and \( t \) converge.

Finally, post-processing is employed to extract the lip region accurately. The detail can be found in Section 3.3.4.

## 5.3 Experiments

### 5.3.1 Database and Initialization

To show the performance of the proposed approach under different capture environments, we utilized four databases:

1. AR face database,
2. CVL face database,
3. GTAV face database, and
4. VidTIMIT face database.

For each image, the part of face between nostril and chin is clipped by a 128 × 128 (AR, CVL and VidTIMIT) or 64 × 64 (GTAV) pixels window as the source of segmentation experiment. Fig. 5.4 shows some samples of both the raw and the clipped source image of our experiment. The images from different databases have different illumination condition, color temperature, and the skin color and moustache are also various. Thus, these images can be seen as a fair testing set.

In the following experiments, \( \hat{m} \) was assigned 10 which can provide enough margin for the lip segmentation. In this paper, we only focus on the scale selection problem, the effect on segmentation performance caused by the variation of \( \hat{m} \) was not discussed. The experiment results regarding segment number selection can be
found in our previous work. Then, we randomly selected \( \hat{m} \) pixels in source image and utilized their observation values to initialize mean vectors, i.e.

\[
\mu^{(j)} = x_{\text{rand}(S)},
\]

where \( \text{rand}(S) \) denotes a number which is selected from the set \( S \) randomly.

Meanwhile, the covariance matrixes are initialized at

\[
\Sigma^{(j)} = \begin{bmatrix}
100 & 0 \\
0 & 100
\end{bmatrix}, \quad 0 \leq j \in L.
\]

Moreover, the three parameters \( \beta, \gamma, \eta_w \) and \( \eta_r \) are set at 2, 2, 0.01 and 0.001 respectively.

The configuration of the MRF \( K \) is pre-assigned by \( \{ K_i = 0 \mid i \in S \} \). That is, all pixel are observed under the finest scale in the initial.

The observation space \( X \) is also the modified \( HSV \) color space defined by Eq. (3.3.26).
5.3.2 Counterparts and Benchmark

In the experiments conducted below, we compared the proposed approach with three existing counterparts denoted as Lievin04 [8], Liang06 [157] and Wang07 [95], respectively.

Lievin04 is an MRF based lip segmentation and tracking method without true segment number in advance. This method is designed to deal with the continuous image sequence. Thus, it was utilized directly in the VidTIMIT database. However, for the databases (AR, CVL and GTAV) composed by separate images, it was employed with sight modifications. We believe that segmenting lip from one separate image can be viewed as the special case of tracking lip from two adjacent frames. Thus, when Lievin04 was employed in the separate image databases, a given image was assigned to the \( t \) and \( t - 1 \) frames in Lievin04’s model simultaneously.

Liang06 is a multi-scale MRF based image segmentation with automatic segment number selection. Actually, this method is designed for texture image segmentation rather than lip segmentation specifically. However, since the similar characters (multi-MRF based, cluster number independent) between Liang06 and the proposed method, it was utilized in our comparative study.

Wang07 is a spatial fuzzy clustering (SFC) based lip segmentation with automatic cluster number selection. Although the mathematical basis is different from the proposed one, they are all fall into the spatial restriction considered category. Another reason why it was utilized in our comparative study is that lots of variants of Wang07 have been proposed, that is, Wang07 can be viewed as a representative lip segmentation method.

We utilize the measures \( OL \) and \( SE \) once again to evaluate the performance of the algorithms. For each image, the ground truth stated above is obtained by manual segmentation. Due to the resolution and illustration condition, the lip region near boundary is hard to be classified even by human. To make the experiment results precise, we asked two volunteers to segment the ground truth independently and each experiment is evaluated by using the two corresponding ground truths.
5.3.3 Result

To evaluate the performance of the proposed method under the different capture environments, i.e. illumination condition, color temperature, and complexion, we selected 50 images from AR, CVL, GTAV and VidTIMIT database, respectively. For images from either database, the average $OL$ and $SE$ were calculated. A snapshot of lip segmentation results is shown in Fig. 5.5. Table 5.1 lists the average $OL$ and $SE$ values obtained by the four methods. The detail results are available in Fig. 5.6.

Another important factor for algorithm evaluation is the processing time. We
Figure 5.5: A Snapshot of lip segmentation results obtained by the proposed method, the images in (a), (b), (c) and (d) are from AR, CVL, GTAV and VidTIMIT database, respectively. In each sub-figure, the first row is the source images, and the second row shows the corresponding results.
Figure 5.6: (a) and (b) are the frequency of specific $OL$ and $SE$ of the selected images from AR database, while (c) and (d) are from CVL, (e) and (f) are from GTAV, (g) and (h) are from VidTIMIT.
selected 50 images from each database to test the time consumption. The mean values and standard deviations are shown in Table 5.2.

5.3.4 Discussion

In the original model of Lievin04, the threshold parameter $\gamma$ is a constant value. However, when we implemented Lievin04, the parameter $\gamma$ was selected adaptively to suit different images. Otherwise there were lots of over- or under-segmentations appeared. For example, for the clip of 026 $- MVC - 007F$ in CVL which is shown in Fig 5.7 (a), the lip region can be extracted accurately by Lievin04 with $\gamma = 165$ (Fig 5.7 (b)). However, the clip of 038 $- MVC - 007F$ in CVL (Fig 5.7 (c)) is over-segmented with the same parameter (Fig 5.7 (d)). In contrast, the proposed method is not sensitive to the parameter selection. There are four parameters utilized in our method: $\beta$, $\gamma$, $\eta_w$ and $\eta_r$. The values of them are selected by rule of thumb. It is no need to adjust them for different images. Moreover, the proposed method is somewhat robust against these parameters as long as they are not changed too much.

Figure 5.7: (a) and (c) are clips of 026 $- MVC - 007F$ and 038 $- MVC - 007F$ in CVL, respectively. (b) and (d) are the corresponding segmentation results obtained by Lievin04 with $\gamma = 165$.

Liang06 utilizes the $VR$ criterion based method to estimate the segment number online. However, this method is not work smoothly all the times in lip segmentation. Consider Liang06 is not designed for lip segmentation specifically, the segment number selection mistakes were corrected manually to make the comparative study
fair in the comparative study. However, this problem shows that Liang06 is highly
depended on the segment number. An example is shown in Fig. 5.8. Obviously, the
lip can be segmented accurately when segment number is select correctly as shown
in Fig. 5.8 (b). Nevertheless, the under- and over-segmentation are appeared caused
by the bias of segment number estimation (Fig. 5.8 (b) and (c)). In our method,
this problem is solved by a rival-penalized algorithm. Result of the third experiment
shows that the segmentation accuracy is independent on the selection of segment
number as long as \( \hat{m} \) is pre-assigned a value larger or equal to \( m^* \).

![Figure 5.8](image)

Figure 5.8: (a) is a clip of \( m - 004 - 8 \) image in AR, (b), (c), (d) are
the segmentation results obtained by Liang06 with segment number 3, 2 and 4,
respectively.

The phenomena stated above imply that the proposed method has more robust
performance.

Although the proposed method shows the promising result in the most cases, the
teethridge and tongue may be classified into the lip segment as shown in Fig. 5.9
since the visible tongue or teethridge have the similar value to lip in the observation
space. From the theoretical viewpoint, this kind of problem is hard to circumvent
by a color analysis based segmentation method. In the future work, we shall explore
to integrate the prior shape modeling with the proposed method to deal with this
problem.
Figure 5.9: (a) and (c) are the clips of $m - 004 - 17$ in AR and $046 - MVC - 007F$ in CVL, (b) and (d) are the corresponding segmentation results obtained by the proposed methods in which some parts of teethridge and tongue are misclassified.

5.4 Summary

This chapter has proposed a local scale dependent segmentation model formulating into the MAP-MRF framework. In this model, the classical energy function in MAP-MRF based segmentation has been rewritten by adding an observation scale variable for each site. Then, an automatic lip segmentation method has been presented. By using this method, the optimal scale configuration and the corresponding optimal segmentation result can be obtained synchronously. Experimental results have shown the efficacy of the proposed method in comparison with the existing three counterparts.
Chapter 6

Conclusion

6.1 Conclusion

The primary objective of this thesis is to overcome the segment number dependency and local scale selection problems in lip segmentations. On one hand, the thesis puts efforts on developing a novel model to perform the lip segmentation without knowing the true segment number. On the other hand, it studies how to exploit the multi-scale image representation for lip segmentation in an effective way.

Chapter 3 presents an MAP-MRF based image segmentation method whose performance is independent of the estimation of the segment number. In this model, a multi-layer model is employed in which each layer is formed by 2-D random field defined on the pixel lattice of the given image. Then, an MRF is derived under which the segmentation problem is formulated as a labeling problem based on the MAP rule. Suppose the pre-assigned number of layers may over-estimate the ground truth, i.e. it may lead to the over-segmentation. We therefore present a rival penalized algorithm to minimize this energy function which is capable of gradually fading out the redundant segment clusters automatically, whereby the over-segmentation is smoothly eliminated. Based upon this algorithm, we propose a lip segmentation and tracking scheme, featuring the robust performance to the estimate of the number of segment clusters.

Chapter 4 presents a fuzzy clustering based segmentation method with objective
function derived from the classical PE and implemented using Havrda-Charvat’s structural $\alpha$-entropy. This objective function features that the coincident cluster centroids in pattern space can be equivalently substituted by one centroid with the function value unchanged. Thus, the optimal partition can be acquired by minimizing the proposed objective function regardless of whatever the pre-assigned cluster number is as long as it is greater than or equal to the ground truth. From the practical viewpoint, it is generally feasible to estimate the range of cluster number. In implementation, we firstly define some cluster centroids whose number is greater than or equal to the ground truth as well. Subsequently, a cooperative algorithm is utilized to minimize the proposed objective function iteratively. Some neighboring centroids will be gradually merged into one so that the over-partition caused by redundant centroids can be eventually faded out. That is, the clustering performance of the proposed algorithm is robust against the pre-assigned cluster number. Based upon the proposed algorithm, a lip segmentation scheme is presented which is robust against the visibility of moustache and teeth.

Chapter 5 proposes an MRF based segmentation method with scale variation considered. We assume that each pixel of a given image has its own optimal scale from the viewpoint of segmentation. Then, the scale map is seen as an MRF defined on the regular pixel lattice. Thus, the scale selection problem and segmentation problem can be formulated under the maximum an MAP-MRF framework jointly. To this end, we propose an MAP based objective function with respect to two variables: the label configuration and the scale configuration. In addition, this method is incorporated into the hierarchical model presented in Chapter 3 under which the lip segmentation can be implemented without true segment number. The optimal scale and the corresponding segmentation result can be obtained by minimizing the objective function.
6.2 Future Work

Although the results presented here have demonstrated the effectiveness, it could be further developed in a number of ways. In the future research, some potential possibilities along the following research directions will be explored:

1. Building a geometric model which can represent the shape of human lip and introducing the model into the proposed methods to implement the lip segmentation jointly.

2. Developing an illumination equalization method to deal with the lip segmentation problem under unsymmetrical illumination environment by improving the proposed multi-scale method and proposing a non-linear color space selection approach.

3. Exploring the visual only speech recognition method by using the features provided by the proposed methods.
Bibliography


106


[99] G. Poggi, G. Scarpa, and J. Zerubia. Supervised segmentation of remote
sensing images based on a tree-structured MRF model. *IEEE Transactions

[100] Z. Kato and T.C. Pong. A multi-layer MRF model for video object segmen-
tation. In *Proceedings of Asian Conference on Computer Vision*, pages 953–962,
Hyderabad, India, 2006.

object-motion detection in unregistered airborne image-pairs. In *Proceedings
of IEEE International Conference on Image Processing*, pages 141–144, San

based on tree-structured MRF model in wavelet domain. *IEEE Geoscience and

[103] Y. Liu, A. Wong, and P. Fieguth. Synthesis of remote sensing label fields using
a tree-structured hierarchical model. *IEEE Transactions on Geoscience and

[104] M. Liévin and F. Luthon. Unsupervised lip segmentation under natural condi-
tions. In *Proceedings of IEEE International Conference on Acoustics, Speech,

their training and application. *Computer Vision and Image Understanding*,

appearance models. *IEEE Transactions on Pattern Analysis and Machine

tive shape models and hidden markov models. In *Proceedings of IEEE Interna-


113


[154] Y.M. Cheung. A competitive and cooperative learning approach to robust
data clustering. In Proceedings of the IASTED International Conference of
Neural Networks and Computational Intelligence, pages 131–136, Grindelwald,


[157] K. Liang and T. Tjahjadi. Adaptive scale fixing for multiscale texture seg-
Appendix A

A.1 Derivation of Eq.(3.2.17)

In layer $j$ of the model proposed in Section 3.2.1, given the site $i$, the corresponding neighboring sites set $\mathcal{N}_i$, and the memberships $\hat{f}_{i'}^{(j)}$ ($i' \in \mathcal{N}_i$), the entropy of $\hat{f}_k^{(j)}$ ($k \in \mathcal{N}_i^+$) can be written as the function with respect to $\hat{f}_i^{(j)}$ as follows:

$$E(\hat{f}_i^{(j)}) = -\sum_{i' \in \mathcal{N}_i} \frac{\hat{f}_{i'}^{(j)}}{K_i + \hat{f}_i^{(j)}} \ln \frac{\hat{f}_{i'}^{(j)}}{K_i + \hat{f}_i^{(j)}} - \frac{\hat{f}_i^{(j)}}{K_i + \hat{f}_i^{(j)}} \ln \frac{\hat{f}_i^{(j)}}{K_i + \hat{f}_i^{(j)}},$$  \hspace{1cm} (A.1.1)

with

$$K_i = \sum_{i' \in \mathcal{N}_i} \hat{f}_{i'}^{(j)}. \hspace{1cm} (A.1.2)$$

To implement the spatial restriction, $\hat{f}_i^{(j)}$ should make $E(\hat{f}_i^{(j)})$ reach the maximum, i.e.

$$\frac{dE(\hat{f}_i^{(j)})}{d\hat{f}_i^{(j)}} = 0. \hspace{1cm} (A.1.3)$$

Solving Eq. (A.1.3), we can obtain:

$$\hat{f}_i^{(j)} = \exp\left(-\frac{\sum_{i' \in \mathcal{N}_i} \hat{f}_{i'}^{(j)} \ln \hat{f}_{i'}^{(j)}}{\sum_{i' \in \mathcal{N}_i} \hat{f}_{i'}^{(j)}}\right). \hspace{1cm} (A.1.4)$$
A.2 Proof of Lemma 4.2.2

We can obtain a mapping \( h_m : \eta_m \rightarrow H_{\eta_m}^{\text{min}} \) by solving the following optimization problem, where \( H_{\eta_m}^{\text{min}} \) is the minimum of \( H(C \mid x_i) \) subject to \( \delta H(C \mid x_i) = \eta_m \):

\[
\begin{align*}
\min & : \quad 1 - \sum_{j=1}^{m} [p(\tilde{c}_j^i \mid x_i)]^2, \\
\text{s.t.} & : \\
& \quad \left\{ \begin{array}{l}
1 - \sum_{j=1}^{m} \left[ \frac{p(\tilde{c}_j^i \mid x_i) - p(\tilde{c}_{j-1}^i \mid x_i)}{p(\tilde{c}_m^i \mid x_i)} \right]^2 = \eta_m, \\
\sum_{j=1}^{m} p(\tilde{c}_j^i \mid x_i) = 1, \\
p(\tilde{c}_j^i \mid x_i) \geq 0, \\
p(\tilde{c}_j^i \mid x_i) \leq p(\tilde{c}_k^i \mid x_i) \text{ if } j < k.
\end{array} \right. \quad (A.2.5)
\end{align*}
\]

Using the substitution

\[
p(\tilde{c}_j^i \mid x_i) = \sum_{k=1}^{j} a_k^2, \quad a_k \in \mathbb{R},
\]

the optimization problem can therefore be simplified as

\[
\begin{align*}
\min & : \quad 1 - \sum_{j=1}^{m} \left( \sum_{k=1}^{j} a_k^2 \right)^2, \\
\text{s.t.} & : \\
& \quad \left\{ \begin{array}{l}
\sum_{k=1}^{m} a_k^4 + (\eta_m - 1)(\sum_{k=1}^{m} a_k^2)^2 = 0 \\
\sum_{k=1}^{m} (m + 1 - k)a_k^2 = 1.
\end{array} \right. \quad (A.2.7)
\end{align*}
\]

Subsequently, the corresponding Lagrange function is

\[
\Lambda_m(a_1, \ldots, a_n, \alpha, \beta) = 1 - \sum_{j=1}^{m} \left( \sum_{k=1}^{j} a_k^2 \right)^2 + \alpha \left[ \sum_{k=1}^{m} a_k^4 + (\eta_m - 1)(\sum_{k=1}^{m} a_k^2)^2 \right] + \beta \sum_{k=1}^{m} (m + 1 - k)a_k^2 - 1, \quad (A.2.8)
\]
where $\alpha$ and $\beta$ are Lagrange multipliers.

Thus, the constrained extrema of Eq.(A.2.7) are the extreme points of Eq.(A.2.8), which can be obtained by solving the following equations:

$$\nabla_{a_l, \alpha, \beta} \Lambda_m = 0 \ (l = 1, 2, \cdots, m), \quad (A.2.9)$$

which can be further expressed as

$$\begin{cases}
    a_1 \sum_{j=1}^{m} (\sum_{k=1}^{\tilde{j}} a_k^2) - \alpha a_1 [a_1^2 + (\eta_m - 1) \sum_{k=1}^{m} (a_k^2)] = \frac{m \beta a_1}{2} \\
    a_2 \sum_{j=2}^{m} (\sum_{k=1}^{\tilde{j}} a_k^2) - \alpha a_2 [a_2^2 + (\eta_m - 1) \sum_{k=1}^{m} (a_k^2)] = \frac{(m-1) \beta a_2}{2} \\
    \vdots \\
    a_l \sum_{j=l}^{m} (\sum_{k=1}^{\tilde{j}} a_k^2) - \alpha a_l [a_l^2 + (\eta_m - 1) \sum_{k=1}^{m} (a_k^2)] = \frac{(m-l+1) \beta a_l}{2} \\
    \vdots \\
    a_m \sum_{k=1}^{m} a_k^2 - \alpha a_m [a_m^2 + (\eta_m - 1) \sum_{k=1}^{m} (a_k^2)] = \frac{\beta a_m}{2}
\end{cases} \quad (A.2.10)$$

For any of the first $m$ equations in Eq.(A.2.10), i.e. $\nabla_{a_l} \Lambda_m = 0$, we fix $a_{\tilde{l}}$ ($\tilde{l} = 1, \ldots, l-1, l+1, \ldots, m$) and $\alpha$. Therefore, $a_l^2$ can be represented as a linear function with respect to $\beta$, i.e.

$$(m - l + 1 - \alpha \eta_m)a_l^2 + \frac{l-m-1}{2} \beta + \sum_{j=l}^{m} (\sum_{k=1, k \neq l}^{j} a_k^2) - \alpha (\eta_m - 1) \sum_{k=1, k \neq l}^{m} (a_k^2) = 0. \quad (A.2.11)$$

Subsequently, we can eliminate $a_l$ and obtain a quadratic polynomial of $\beta$ via substituting Eq.(A.2.11) into $\nabla_{\alpha} \Lambda_m = 0$.

On the other hand, $\beta$ can be calculated by solving Eq.(A.2.10). The result is:

$$\beta = 2 \sum_{j=1}^{m} (\sum_{k=1}^{j} a_k^2)^2 = 2(1 - H_{\eta_m}^{sta}), \quad (A.2.12)$$

119
where $H_{\eta m}^{sta}$ can be calculated by the possible stationary point of Eq.(A.2.8). Substituting Eq. (A.2.12) into the quadratic polynomial determined by Eq.(A.2.11) and $\nabla_\alpha \Lambda_m = 0$, we can obtain a quadratic polynomial with respect to $H_{\eta m}^{sta}$. That is, for Eq.(A.2.8), the number of stationary points is 0, 1, or 2. Based on the extreme value theorem, this number can be further fixed to 2, corresponding to global maximum and minimum, respectively.

Suppose the minimum of Lagrange function $\Lambda_{m-1}$ is obtained at the point $(a_1, a_2, \ldots, a_{m-1}, a_1, \alpha, \beta)$. According to Eq.(A.2.10), $\Lambda_m$ has the stationary point at $(0, a_1, a_2, \ldots, a_{m-1}, a_1, \alpha, \beta)$ as long as $H_{\eta m}^{sta} = H_{\eta m-1}^{sta}$. Let the Hessian matrix of $\Lambda_{m-1}$ at $(a_1, a_2, \ldots, a_{m-1}, a_1, \alpha, \beta)$ be $\mathcal{H}_{m-1}$. Then, the Hessian matrix of $\Lambda_m$ at $(0, a_1, a_2, \ldots, a_{m-1}, a_1, \alpha, \beta)$ can be represented recursively as

$$
\mathcal{H}_m = \begin{bmatrix}
4 \alpha a_1^2 + 2 \beta & A \\
B & \mathcal{H}_{m-1}
\end{bmatrix}
$$

(A.2.13)

where $A = [0, 0, \ldots, 0]$ and $B = [0, 0, \ldots, 0]^T$.

As we know, the entropy value of a random variable will tend to zero as the variable becomes certainty. Thus, we suppose that $p(\tilde{c}_i^1 \mid x_i) = a_1^2 \rightarrow 0$ and $p(\tilde{c}_m^1 \mid x_i) = \sum_{j=1}^{m} a_j^2 \rightarrow 1$ when the constrained minimum in Eq.(A.2.7) is obtained. Under this situation, $\mathcal{H}_m$ is a positive definite matrix. Moreover, as stated above, since there is only one minimum stationary point in $\Lambda_m$ as given a specific $\eta_m$, $(0, a_1, a_2, \ldots, a_{m-1}, a_1, \alpha, \beta)$ must be the global minimum of Eq.(A.2.8). Thus, $h_m$ can be represented by the following recursion approximatively:

$$
h_m(\eta_{m-1}) \approx h_{m-1}(\eta_{m-1}),
$$

(A.2.14)

as illustrated in Fig. A.1.

When $\eta_{m-1} \in (0, \frac{k-1}{k}]$ with $k = 2, 3, \ldots, m - 1$, the curves of $h_m(\eta_{m-1})$ and $h_k(\eta_{m-1})$ are coincident (e.g. see Fig. A.2). Then, Eq.(A.2.14) can be further
Figure A.1: The functional relationship $h_m$ between $\eta_m$ and $H_{\eta m}^{min}$ with (a) $m = 2$, (b) $m = 3$, (c) $m = 4$, (d) $m = 5$, (e) $m = 6$, and (f) $m = 7$, respectively, where the horizontal axis represents the value of $\delta H_m$, and the vertical axis represents the value of $H_{\eta m}^{min}$.
Figure A.2: The curve of $H_{\eta_k}^{\text{min}} = h_k(\eta_k)$. When $\eta_k \in (0, \frac{k-1}{k}]$ with $k = 1, 2, \ldots, 7$, the corresponding curve segments are coincident with $h_k(\eta_k)$.

formulated as

$$h_m(\eta_{m-1}) = \begin{cases} h_{m-1}(\eta_{m-1}), & \eta_{m-1} \in (0, \frac{m-2}{m-1}], \\ h_{m-2}(\eta_{m-1}), & \eta_{m-1} \in (0, \frac{m-3}{m-2}], \\ \ldots, & \\ h_2(\eta_{m-1}), & \eta_{m-1} \in (0, \frac{1}{2}], \\ 0, & \eta_{m-1} = 0. \end{cases} \tag{A.2.15}$$

Subsequently, substituting Eq.(A.2.12) and $\nabla \Lambda_m = 0$ into $\nabla a_1 \Lambda_m = 0$, we can obtain

$$H_{\eta_m}^{\text{sat}} = \frac{\alpha \sum_{k=1}^{m} a_k^2}{m} \eta_m + \frac{\alpha a_1^2 - \alpha \sum_{k=1}^{m} a_k^2 + m - 1}{m}. \tag{A.2.16}$$

When the minimum of Eq.(A.2.16) is achieved, and $m \to +\infty$, we have

$$H_{\eta^{\text{min}}_{\eta_m}} = \eta_{+\infty}. \tag{A.2.17}$$

Based on Eq.(A.2.15), the relationship between $\eta_m$ and $H_{\eta_m}^{\text{min}}$ can be written approximatively as

$$H_{\eta_m}^{\text{min}} \approx \eta_m. \tag{A.2.18}$$
Table A.1: MSE between the numerical simulation result $\hat{H}^\text{min}_{\eta_m}$ and ideal value $H^\text{min}_{\eta_m} = \eta_m$ over $m$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.0016</td>
<td>0.0025</td>
<td>0.0028</td>
<td>0.0028</td>
<td>0.0028</td>
<td>0.0028</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

A.3 Validation of Lemma 4.2.2

We conduct an experiment to show the validity of Theorem 4.2.3. Firstly, we select an input $x_i$ in pattern space randomly, and calculate the corresponding $H^\text{min}_{\eta_m}$ for different $\eta_m \in (0, \frac{m-1}{m}]$ by interior point method. Then, we utilize the mean square errors (MSE) to evaluate the bias between the numerical simulation result, denoted as $\hat{H}^\text{min}_{\eta_m}$, and the desired value, i.e. $H^\text{min}_{\eta_m} = \eta_m$. Moreover, this experiment is repeated with $m = 2, 3, \ldots, 8$. For each $m$, we select 5 different $x_i$s. Table A.1 lists the average MSE with a specific value of $m$. It can be seen that the error is relative small. This implies that Theorem 4.2.3 is indeed valid in all cases we have tried so far.

A.4 Derivation of Eq.(5.2.7)

Based on the Bayesian rule, Eq.(5.2.6) can be further expressed as

$$\{f^*, t^*\} = \arg \max_{f \in \Omega} P(t) \cdot P(f | t) \cdot P(x | f, t).$$  \hspace{1cm} (A.4.19)

Under the proposed hierarchical model, we have $x_i = x_i^{(t_i)}$ and $f_i = f_i^{(t_i)}$ when $t = \{t_i \mid i \in S, t_i \in D\}$ is determined. The conditional terms can be rewritten as:

$$P(f | t) = P(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}),$$  \hspace{1cm} (A.4.20)
and

\[ P(x \mid f, t) = P(\mathcal{X}_1 = x_1^{(t_1)}, \mathcal{X}_2 = x_2^{(t_2)}, \cdots, \]
\[ \mathcal{X}_s = x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \]
\[ \mathcal{F}_s = f_s^{(t_s)}). \]  

(A.4.21)

Thus, Eq.(A.4.19) can be described as:

\[ \{f^*, t^*\} = \arg \max_{\{f^{(t_i)\mid i \in S}\} \in \Omega^L} P(\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2, \cdots, \mathcal{T}_s = t_s). \]
\[ P(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \mathcal{F}_s = f_s^{(t_s)}), \]
\[ P(\mathcal{X}_1 = x_1^{(t_1)}, \mathcal{X}_2 = x_2^{(t_2)}, \cdots, \]
\[ \mathcal{X}_s = x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \]
\[ \mathcal{F}_s = f_s^{(t_s)}). \]  

(A.4.22)

According to the Hammersley-Clifford theorem, Eq.(A.4.22) can be written in the form of energy function \( U(\cdot) \):

\[ \{f^*, t^*\} = \arg \min_{\{f^{(t_i)\mid i \in S}\} \in \Omega^L} U(\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2, \cdots, \mathcal{T}_s = t_s) + \]
\[ \beta U(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \mathcal{F}_s = f_s^{(t_s)}) + \]
\[ \gamma U(\mathcal{X}_1^{(t_1)} = x_1^{(t_1)}, \mathcal{X}_2^{(t_2)} = x_2^{(t_2)}, \cdots, \]
\[ \mathcal{X}_s^{(t_s)} = x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \]
\[ \mathcal{F}_s = f_s^{(t_s)}). \]  

(A.4.23)

where \( \beta \) and \( \gamma \) are the weighting parameters, and the equation

\[ \mathcal{E}(f, t) = U(\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2, \cdots, \mathcal{T}_s = t_s) + \]
\[ \beta U(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \mathcal{F}_s = f_s^{(t_s)}) + \]
\[ \gamma U(\mathcal{X}_1^{(t_1)} = x_1^{(t_1)}, \mathcal{X}_2^{(t_2)} = x_2^{(t_2)}, \cdots, \]
\[ \mathcal{X}_s^{(t_s)} = x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \cdots, \]
\[ \mathcal{F}_s = f_s^{(t_s)}). \]  

(A.4.24)

124
can be denoted by

\begin{equation}
\mathcal{E}(f, t) = U(t) + \beta U(f^{(t)}) + \gamma U(x^{(t)} | f^{(t)}).
\end{equation}

When $x$ is regarded as a parameter, we can obtain Eq.(5.2.7).
Appendix B

B.1 Publication Record

Publications during the period of study for Ph.D. which directly relevant to this thesis are listed as follows.

Journal Papers


Conference Papers

126


Curriculum Vitae

Academic qualifications of the thesis author, Mr. LI Meng:

• Received the degree of Bachelor of Engineering from Harbin Institute of Technology, July 2004.

• Received the degree of Master of Engineering from Harbin Institute of Technology, July 2007.

August 2014