Markov Random Fields Based Image and Video Processing

LIU, Ming

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Information Engineering

The Chinese University of Hong Kong
June 2010

Abstract

Many problems in computer vision involve assigning each pixel a label, which represents some spatially varying quantity such as image intensity in image denoising or object index label in image segmentation. In general, such quantities in image processing tend to be spatially piecewise smooth, since they vary smoothly in the object surface and change dramatically at object boundaries, while in video processing, additional temporal smoothness is satisfied as the corresponding pixels in different frames should have similar labels. Markov random field (MRF) models provide a robust and unified framework for many image and video applications. The framework can be elegantly expressed as an MRF-based energy minimization problem, where two penalty terms are defined with different forms. Many approaches have been proposed to solve the MRF-based energy optimization problem, such as simulated annealing, iterated conditional modes, graph cuts, and belief propagation.

In this dissertation, we propose three methods to solve the problems of interactive image segmentation, video completion, and image denoising, which are all formulated as MRF-based energy minimization problems. In our algorithms, different MRF-based energy functions with particular techniques according to the characteristics of different tasks are designed to well fit the problems. With the energy functions, different optimization schemes are proposed to find the optimal results in these applications. In interactive image segmentation, an iterative optimization based framework is proposed, where in

each iteration an MRF-based energy function incorporating an estimated initial probabilistic map of the image is optimized with a relaxed global optimal solution. In video completion, a well-defined MRF energy function involving both spatial and temporal coherence relationship is constructed based on the local motions calculated in the first step of the algorithm. A hierarchical belief propagation optimization scheme is proposed to efficiently solve the problem. In image denoising, label relaxation based optimization on a Gaussian MRF energy is used to achieve the global optimal closed form solution.

Promising results obtained by the proposed algorithms, with both quantitative and qualitative comparisons to the state-of-the-art methods, demonstrate the effectiveness of our algorithms in these image and video processing applications.

摘要

計算機視覺領域中很多問題涉及到為每個像素分配一個標註，這個標註代表着在空間上變化的一個參量，比如在圖像去噪中代表圖像灰度，在圖像分割中代表物體的索引編號。一般情況下，這類參量在圖像處理中趨於空間上的分段平滑，因為他們在圖像物體的表面變化平緩而在物體的邊緣變化劇烈。同時在視頻處理中，時間域上的平滑性也同樣滿足，因為視頻中不同幀上對應的像素應該有相似的標註值。馬爾可夫場模型為許多圖像和視頻處理問題提供了一個統一而魯棒的框架。這個框架能夠被表述為一個基於馬爾可夫場的能量最小化問題，其中能量函數中的兩個懲罰項能夠被定義為不同的形式。已經有許多方法被提出來去解決基於馬爾可夫場的能量優化問題，比如模擬退火法，條件模式迭代法，圖切割算法和置信傳播算法。

本論文提出了三個算法分別解決了交互式圖像分割問題，視頻修補問題和圖像去噪問題。三個算法都被表述為基於馬爾可夫場的能量最小化問題。在各個的算法中，根據不同任務的特點，用不同的技巧設計出了不同的馬爾可夫場能量函數來更好地適應問題本身。有了能量函數，在這些應用中，不同的優化方案被提出來尋找最優的結果。在交互式圖像分割問題中，我們提出了一個基於迭代優化的算法框架。在每一次迭代中，結合了被估計出的圖像概率圖譜的馬爾可夫場能量函數被優化從而得到一個寬鬆的全局最優解。在視頻修復問題中，基於在算法第一步中計算出的局部運動，我們建立了一個包括空間域和時間域一致性關係的馬爾可夫場能量函數。為了有效地解決這個優化問題，我們提出了多級置信傳播方案。在圖像去噪問題中，我們通過標註鬆弛優化得到了高斯馬爾可夫場能量優化問題的全局最優封閉解。

本文提出的算法所得到的實驗結果以及與最新方法定性定量的比較都證明了我們的算法對於這些圖像和視頻處理應用的有效性。

Acknowledgement

First of all I would like to express my sincere thanks to my supervisors Prof. Tang Xiaoou and Prof. Liu Jianzhuang. Prof. Tang has always been an knowledgeable adviser, providing us deep insight and significant big-picture perspectives. He hardly criticizes students, instead, encourages us to think actively and independently. He shares his valuable life experience with us to guide us on the right way. The conversations with Prof. Tang are pleasing and instructive. After working with Prof. Liu in the past four years, I understand the real meaning of being hardworking, conscientious, energetic, and persistent. He is always being there to discuss the ideas with us, to help us think through the problems. His kind help makes me grow up step by step in research.

Four years spent in MMLab are memorable. I knows many nice guys here. Weige, a trustworthy friend, always stands by me at any time with his big support; Shifeng and ChenMo, being good friends I am proud of, generously help me a lot in my research and life; and Huang Ting, Xiaodai, Chunjing, Yueming, Yiwen, I am really grateful to you guys for everything we have in these four years. Moreover, I also extend my thanks to Xiaowei, Boqing, Duhao, Zhimin, LiuKe, Tianfan, Xiaotian, Yichen, Dahua, Yingze, Huanzi, Pengfei, Qiaoyu, Zhaofeng, Zhenguo, Deli, LiYun, Chenyu, JiaKui, and Kaiming. I will remember the time with you.

Besides, I wish to thank all my friends everywhere for your kind care about my life. Although we are not in the same place or work in the same area, the connections with you guys always warm my heart.

Last but not least, I give my deepest gratitude to my family, mom and dad, for giving my life in the first place, for educating, supporting, and comforting me in any situations throughout my life; my girlfriend Huijun, for her support and encouragement to get me through the hard days, for her understanding to everything in our love at distance, and for her deep love to me making my heart peaceful. I love you all!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many vision problems require estimating some spatially varying quantities from noisy measurements. Therefore, label assignment is the essential part in a range of image and video processing tasks. In image denoising, the label of each pixel in the image is the noise-free intensity, which we need to estimate from the observed noisy image. In the context of interactive foreground object extraction problem (interactive image segmentation), the label is defined as the object/background index obtained by using user specified prior information (color, texture, or location). Moreover, the label in image and video completion can be regarded as the source patch index, which guides the process of filling in the missing regions. Besides, there are many more applications involving the labelling problem, such as image matting, image photomontage, etc.

It has been known for decades that such labelling problems can be elegantly expressed as Markov Random Fields (MRFs), since for each pixel we only consider pairwise connections with its neighboring pixels. Therefore the Markov random fields based image and video processing can be formulated as an MRF-based energy minimization problem, which has been demonstrated to well model these problems. The MRF models provide a robust and unified framework, which is justified in terms of maximum a posteriori (MAP) estimation of a Markov random field in the Bayesian framework. In the next section, we will briefly introduce the basic Markov random fields model.

1.1 Markov Random Fields Model

Markov Random Fields model is first introduced in computer vision in [25]. One of the reasons why this framework is so popular is that it can be justified in terms of maximum a posteriori (MAP) estimation of a Markov random field. An MRF, which from graphic model perspective is an undirected graph with each node being a random variable and each undirected link denoting the neighboring connection [6], has several components in the context of an image and video processing application. For every pixel p, its label l_p is a random variable taking a value in some label set. Different applications have different label sets and different physical meanings of the label. $\mathcal{N}$ denote some neighborhood system. For a single image, only spatial neighborhood system (e.g., 4−connected or 8−connected neighborhood system) is used. For a video, a temporal neighborhood system (e.g., corresponding pixels in different frames) is considered as well. $L = [l_1, l_2, ..., l_n]$ is the collection of all pixel label assignment, where n is the number of pixels in the image. With each particular label assignment (label configuration), L corresponds to a realization of the field.

In order to be an MRF, L must satisfy

$$P(l_p|l_{S-\{p\}}) = P(f_p|f_{\mathcal{N}(p)}), \quad \forall p \in \{1, 2, ..., n\}, \tag{1.1}$$

where S is the set of all pixels, and $\mathcal{N}(p)$ is the neighbors of p. This condition states that each random variable l_p only depends on its neighbors. Based on the Hammersley-Clifford theorem [4], the prior probability $P(L)$ (joint probability represented by an MRF undirected graph) can be modeled by MRFs whose clique potentials involve pairs of neighboring pixels, defined as

$$P(L) \propto \exp(-\sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} V_{ij}(l_i, l_j)), \tag{1.2}$$

where $V_{ij}(l_i, l_j)$ is the clique potential representing the prior knowledge of the

label relationship, which is also called the smoothness penalty function imposing the pairwise smoothness. Therefore the MRF framework can express a wide variety of spatially and temporally varying priors by choosing different forms of $V_{ij}(l_i, l_j)$, which we will discuss below.

In general, the field L need to be estimated based on the observation or some prior information obtained from the data, which is represented by X. $P(X|L)$ is a likelihood function and can be represented by the sensor noise model [11]:

$$P(X|L) \propto \prod_{i=1}^{n} \exp(-D_i(l_i)), \tag{1.3}$$

where $D_i(l_i)$ is called the data penalty function that penalizes the inconsistency between the labels and the data.

With the prior probability $P(L)$ and the likelihood function $P(X|L)$, by using the Bayes' rule and removing $P(X)$, which is a constant with respect to L, maximizing the posterior probability $P(L|X)$ is to find $\widehat{L}$ such that

$$\widehat{L} = \underset{L}{\operatorname{argmax}}\, P(X|L)P(L). \tag{1.4}$$

From (1.4), (1.3), and (1.2), we can see that the MAP estimation is equivalent to minimizing the following energy function:

$$E(L) = \sum_{i=1}^{n} D_i(l_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} V_{ij}(l_i, l_j), \tag{1.5}$$

On the other hand, the general MRF-based energy is composed of a data energy E_d and a smoothness energy E_s, defined as

$$E = E_d + \lambda E_s. \tag{1.6}$$

The data energy is simply the sum of a set of per-pixel data cost $D_i(l_i)$

$$E_d = \sum_{i=1}^{n} D_i(l_i). \tag{1.7}$$

And the smoothness energy is

$$E_s = \sum_{\{i,j\} \in \mathcal{N}} V_{ij}(l_i, l_j). \tag{1.8}$$

With the definition of the data energy (1.7) and the smoothness energy (1.8) substituting to the MRF energy (1.6), the MRF energy function is consistent with the MAP estimation derived in (1.5). Therefore the formulation of the MRF model is justified in terms of the MAP estimation of the MRFs in the Bayesian framework.

In general the data term $D_i(l_i)$ has various forms modeling the sensor noise in different applications. The choice of the clique potential function $V_{ij}(l_i, l_j)$ is a critical issue and many clique potential functions have been proposed. Based on the piecewise smoothness assumption on L which is an intrinsic property of the image and video, a good clique potential should be able to enforce spatial homogeneity in low contrast regions and preserve discontinuity in high contrast regions such as object boundaries. Popular clique potentials include the generalized Potts Model [11], $w_{ij}(1-\delta(l_i-l_j))$, where $\delta(\cdot)$ is the unit impulse function, the truncated linear clique potential [12], $w_{ij} \cdot \min\{|l_i - l_j|, K\}$, and the truncated quadratic function [12], $w_{ij} \cdot \min\{(l_i - l_j)^2, K\}$. The factor w_{ij} in these functions denotes the affinity value between pixels i and j and spatially varies to control the degree of the smoothness constraint in different neighborhoods. It is obvious that the MAP estimation $\widehat{L}$, which maximizes $P(X|L)P(L)$ or equivalently minimizes $E(L)$, tends to be consistent with the observed data X as well as to preserve the piecewise continuity.

Some special cases of MRF-based energy minimization can be solved by fast exact algorithms. For example, if there are only two labels, the Potts model can be solved exactly with graph cuts [31]. However, the MRF-based energy minimization with multi-labels in the discrete domain is generally NP-hard, which implies that the vast majority of MRF-based energy functions are intractable. The major obstacle of the optimization is the large computational

cost owing to the high dimensional computing space. In the next section, we will introduce some related optimization approaches for minimizing the MRF-based energy functions.

1.2 Optimization Approaches

Despite the elegance and power of the MRF model in image and video processing, its early adoption was slowed by computational considerations. The optimization problem is generally NP-hard in the discrete domain. Many approaches have been proposed to solve the MRF-based energy optimization such as simulated annealing, the iterated conditional modes (ICM), recently developed graph cuts (GC), belief propagation (BP) and tree-reweighted message passing (TRW-S), and some continuous optimization algorithms. The details of these optimization approaches are described as follows.

Simulated annealing is used to carry out the MRF optimization in [25]. It can handle arbitrary energy functions and achieve the global optimum theoretically. However, the main problem with simulated annealing is that it is very time consuming and usually cannot obtain the global optimum in limited running time. The iterated conditional modes (ICM) [4] is also used to solve the problem using a deterministic greedy strategy to find a local minimum. It starts with an estimate of the labelling, and then, for each pixel, it chooses the label giving the largest decrease of the energy function. This process is repeated until convergence, which is guaranteed to occur, and, in practice, is very rapid. Unfortunately, the results are extremely sensitive to the initial estimate.

Over the last few years, energy minimization approaches have had a renaissance, primarily due to the development of new powerful optimization algorithms such as graph cuts [11], [12], [42], belief propagation [85], [82], and tree-reweighted message passing (TRW-S) [40].

Graph cuts techniques map the energy function to a proper constructed graph and find the labelling of the nodes that minimizes the energy function by using the min-cut/max-flow [10]. The two most popular graph cuts algorithms, called *expansion-move* algorithm and *swap-move* algorithm, which perform comparably well and both converge to a strong local optimum in different criteria, are introduced in [12]. Both algorithms work by repeatedly computing the global minimum of a binary labelling problem in their inner loops.

For a pair of labels α and β, a swap move takes some subset of the pixels currently given the label α and assigns them the label β and vice versa. The swap-move algorithm finds a local minimum such that there is no swap move, for any pair of labels α and β, which will produce a lower energy labelling. Analogously, an expansion move for a label α is defined to increase the set of pixels that are given this label. The expansion-move algorithm finds a local minimum such that no expansion move, for any label α, yields a labelling with lower energy.

The expansion-move algorithm and the swap-move algorithm always obtain good results. However, as shown in [42], the two algorithms only can be used when the following *regularity condition* holds: in the expansion-move algorithm, for all labels α, β, and γ,

$$V_{pq}(\alpha,\alpha) + V_{pq}(\beta,\gamma) \leq V_{pq}(\alpha,\gamma) + V_{pq}(\beta,\alpha); \tag{1.9}$$

in the swap-move algorithm, for all labels α and β,

$$V_{pq}(\alpha,\alpha) + V_{pq}(\beta,\beta) \leq V_{pq}(\alpha,\beta) + V_{pq}(\beta,\alpha). \tag{1.10}$$

If the energy does not obey these constraints, graph cuts algorithms can still be applied by "truncating" the violating terms [69]. In this case, however, we are no longer guaranteed to find the optimal labelling with respect to expansion or swap moves.

Belief propagation is a message passing algorithm. The max-product and sum-product algorithms are two typical BP algorithms [23]. Normally the

two algorithms are both defined in terms of probability distributions. In implementation, the sum-product algorithm computes the marginal probability distribution of each node directly and finds the minimum mean-squared error estimation of the labels. But for max-product, an equivalent computation can be performed with negative log probabilities, where the max-product becomes a min-sum. In this thesis, we consider this max-product formulation because it is less sensitive to numerical artifacts, and it uses the energy function definition more directly.

The max-product BP algorithm works by passing messages around the graph defined by image grid. Each message is a vector of dimension given by the number of possible labels. Let m_{pq}^t be the message that node p sends to a neighboring node q at time t. All entries in m_{pq}^0 are initialized to zero, and at each iteration new messages are computed in the following way,

$$m_{pq}^t(l_q) = \min_{l_p} \left(V_{pq}(l_p, l_q) + D_p(l_p) + \sum_{s \in \mathcal{N}(p) \backslash q} m_{sp}^{t-1}(l_p) \right), \tag{1.11}$$

where $\mathcal{N}(p) \backslash q$ denotes the neighbors of p other than q. After T iterations a belief vector is computed for each node,

$$b_q(l_q) = D_q(l_q) + \sum_{\{p,q\} \in \mathcal{N}} m_{pq}^T(l_q). \tag{1.12}$$

Finally, the label l_q^* that minimizes $b_q(l_q)$ individually at each node is selected.

TRW-S is a message passing algorithm similar to belief propagation, and often performs as well as BP and GC. An interesting feature of TRW-S is that it computes a lower bound on the energy. In this dissertation, we do not use TRW-S, and therefore will not describe the details of this method.

The approaches introduced above are discrete optimization techniques. Alternatively, label relaxation provides a strategy to convert the combinatorial optimization problem to a continuous optimization problem, which can be solved easier in the continuous domain [18, 35, 63, 67, 79]. In this thesis, continuous optimization for MRF-based energy with quadratic smoothness term

	Domain	Results	Condition	Optimization Mode
GC	discrete	local	regularity	iterative
BP	discrete	local	/	iterative
CF	continuous	global	L_2 smoothness	closed form

Table 1.1: Comparison of algorithms GC, BP, and CF.

is considered. For such a problem, closed form global optimal solutions can be obtained in the continuous domain. We call this a closed form (CF) optimization. Detailed explanation of how to get the closed form solution is given in the context of specific applications in this dissertation.

Table 1.1 is the comparison of the algorithms GC, BP, and CF in aspects of application domain, results, work condition, and optimization mode.

1.3 Our Work and Contributions

With the flourish of these optimization techniques, MRF-based algorithms have been widely used in many image and video processing such as image denoising [11, 12], interactive segmentation [7, 9, 68], stereo correspondence [37, 41, 78], image completion [83], etc. In this thesis, we propose three algorithms to solve the problems of interactive image segmentation, video completion, and image denoising, which are all formulated as MRF-based energy minimization problems.

The goal of interactive image segmentation is to find the region of a foreground object using as little interactive effort as possible. We propose a general framework to address the problem, which can well utilize and expand user provided information to iteratively refine the optimization target to solve the problem, leading to more precise results naturally. In each iteration, we design an MRF-based energy function and use the label relaxation optimization scheme to find the optimal label configuration with respect to the current confidence seeds and obtained color models. Recursively, the continuous label

configurations are refined to indicate the foreground as accurately as possible, leading to the final precise foreground object extraction result when the algorithm converges.

The target of video completion is to restore the spatial-temporal missing regions of a video in a visually plausible way. We propose a novel global optimization based approach for video completion. Our algorithm consists of two stages: motion field completion and color completion via global optimization. The local motions are completed greedily, and the video completion is formulated as a global energy minimization problem by MRFs with a well-defined MRF-based energy function involving both spatial and temporal coherence relationship. To avoid the computational impracticability caused by the large number of label candidates in the optimization process of belief propagation, we propose a hierarchical BP optimization scheme to optimize the energy and obtain good results.

Image denoising is to restore the noise-free image. We formulate it as a continuous label assignment problem based on a Gaussian MRF model and obtain a closed form global optimal solution. Since the Gaussian MRFs tend to over-smooth images and blur edges, we incorporate pre-estimated edge information into the energy function to better preserve image structures. Patch similarity based pairwise interaction is also involved to better preserve image details and make the algorithm more robust to impulse noise.

Promising results obtained by the proposed algorithms, with both quantitative and qualitative comparisons to the state-of-the-art methods, demonstrate the effectiveness of our algorithms in these image and video processing applications.

Our work has been published in or submitted to [49, 51, 16, 52, 50, 47, 48]. The details of the three proposed algorithms are presented in Chapters $2-4$.

Chapter 2

Iterative Foreground Object Extraction

In this chapter we propose a general framework to address the problem of interactive foreground object extraction from an image, which can well utilize and expand user provided information to iteratively refine the optimization target to solve the problem, leading to more precise results naturally. The basic operations in the framework are objective energy construction and its optimization. We incorporate an initial probabilistic map of the image associated with foreground into the energy construction, which is calculated from prior foreground and background color models. By iteratively expanding confidence seed sets to train the color models, we can improve the accuracy of the initial probabilistic map, ensure a better energy function to be optimized, and achieve good foreground extraction results.

Based on our general framework, we design a Markov random field (MRF) based objective energy function and propose a label relaxation optimization scheme in each iteration to find the optimal label configuration with respect to the current confidence seeds and obtained color models. Specifically, in each iteration we estimate two Gaussian mixture models from the confidence seeds, one for foreground and the other for background, and define two quantities to measure the initial probabilities of each pixel belonging to the foreground and

the background respectively, which generate the initial probabilistic maps of the image associated with the foreground and background. With the energy function constructed based on the initial probabilistic map and the boundary and coherent region information, a closed form global optimal solution can be achieved by relaxing the *hard* binary segmentation to a *soft* labelling problem in the continuous domain. The global optimum can be regarded as an optimized probabilistic map, which directly provides us the clues to update the confidence seed sets. Recursively, the continuous label configurations are refined to indicate the foreground as accurately as possible, leading to the final precise foreground object extraction result when the algorithm converges.

Our algorithm is simple and accurate, as demonstrated by high-quality segmentation results on natural images and qualitative and quantitative comparisons with state-of-the-art methods on a segmentation database.

2.1 Introduction

Our proposed algorithm in this chapter addresses the problem of user desired foreground object extraction from an image, which is of great practical importance in many applications such as image retrieval, object recognition, and photo/movie editing. Without additional prior information, segmentation process is generally an ill-posed problem. Current fully automatic segmentation methods [87, 72, 17] are far from satisfactory and often generate results deviating from what the user wants. Therefore, interactive image segmentation guided by user provided information becomes popular and draws much attention from researchers.

The aim of interactive segmentation is to accurately find the region of a foreground object using as little interactive effort as possible. There are two main ways for the user to provide foreground and background information: boundary-based and region-based.

A boundary-based tool, such as [26, 58, 59, 64] requires the user to draw along the object boundary and then adjusts the curve to snap to the real boundary piecewise. When the tool cannot adjust the curve well, the user needs to provide additional boundary seed points in order to avoid the deviation of the curve from the desired boundary. These tools need plenty of user interaction and attention to obtain satisfactory results. Besides, they cannot be easily generalized to 3D images. Region-based tools are developed recently and highly improve the efficiency of interactive segmentation. Instead of accurately indicating the object boundary, a region-based tool requires the user to specify two small sets of pixels belonging to the foreground and the background. Then an underlying algorithm carries out the segmentation based on the user input cue. Region-based methods are more convenient to use than boundary-based methods. Next we review recent region-based methods.

Magic Wand in Photoshop fulfills the segmentation by grouping similar pixels using only the color statistics of the user-specified pixels or regions. Intelligent paint [66] merges image regions by using a hierarchical tobogganing algorithm and interactively selects the foreground based on the properties of the underlying regions obtained from each stroke indicated by the user. These two approaches are based on some variations of traditional region growing, where the segmentation boundary is not optimized and may lead to unsatisfactory results.

Recently, many approaches have been proposed to solve the segmentation problem by modeling it as an energy minimization problem [9, 68, 46, 7, 24, 88], where an objective energy function is derived from Markov Random Fields (MRFs) [25] and hard constraints specified by the user [9]. The key of this kind of algorithms is the design of the energy function and the optimization scheme. The energy function should precisely model the problem with the full use of the user provided information, and the optimization strategy should try to obtain the global optimum. Graph cuts, as an efficient optimization

technique, is rapidly developed recently [12, 10, 42]. Thus graph cut based image segmentation methods are highly promoted and achieve promising results [68, 46, 7, 88]. Inspired by [9], GrabCut in [68] and Lazy Snapping in [46] are proposed. With a similar objective energy function to that in [9], Lazy Snapping utilizes pre-segmentation to increase its running speed. The pre-segmentation may cause the final result not to be optimal. Extended from [46], Progressive Cut proposed in [88] models the user's intention into a graph cut framework for the segmentation. The user's intention is analyzed from additional strokes and the results obtained from previous strokes. GrabCut in [68] incorporates Gaussian mixture color models into an MRF-based energy function (GMMRF) and iteratively uses graph cuts for optimization. In addition, shape priors can also be incorporated into graph cuts based formulation to generate more effective segmentation algorithm [24]. Graph cuts optimization is carried out in discrete domain, which can be regarded as a *hard* segmentation technique in the context of segmentation problem. Although these graph cuts based approaches can generally achieve impressive segmentation results, they may fail in low-contrast boundaries and other ambiguous regions. Thus further border editing is often required to capture the desired object [68, 46]. Moreover, for those methods that utilize seed-based estimated color models in the objective energy function [68, 46], generating precise prior color models from limited user provided information is critical to achieve high quality extraction results.

In contrast to graph cut based discrete optimization, the algorithms in [30, 28, 29] carry out the segmentation in continuous domain, which are essentially based on relaxation from discrete optimization to continuous one, and achieve more impressive results. They are random walk based interactive segmentation approaches, obtaining results by assigning each unlabelled pixel to the label with the highest reaching probability of a random walker starting from this unlabelled pixel to all pre-labelled ones (user-specified seeds). The approach

in [28] is an extension of the methods in [30, 29] with prior model involved to solve the problem in [30, 29] that they can only produce a segmentation where each segment is connected to a labeled pixel.

Our proposed algorithm is also region-based where several strokes (seeds) are provided by the user to indicate the foreground and the background and then the algorithm does the segmentation automatically. It is worth mentioning that except GrabCut [68], the region-based approaches discussed above with prior models have the limitation that they are very sensitive to the number and the locations of user provided seeds because the prior color models for foreground and background rely on the seeds and the locations of seeds encode the spatial foreground/background confidence information. In our experiments, we observe that the more confidence seeds for training the color models, the more accurate models and spatial information can be achieved, resulting in precise segmentation outputs. Therefore, we propose an iterative optimization framework to conquer the drawback. Our algorithm iteratively estimates more and more confidence seeds to update the color models, which provide more useful spatial information. In each iteration, two Gaussian mixture models (GMMs) for modeling the foreground and background are trained using the updated confidence seeds. An initial probabilistic map associated with the foreground is estimated from the updated models. We formulate the segmentation as a problem of labelling the non-seed pixels as 1 (the foreground) and 0 (the background). Thus the labels can be regarded as the probabilities of the pixels belonging to the foreground. By optimizing a well defined objective energy function in the continuous domain, a global optimal label configuration, which is the best probabilistic map of the image with respect to the current color models and coherent region information, can be achieved. Once obtaining the optimal probabilistic map, we update the confidence seeds by the lower and upper thresholding of it to find more accurate color models. The final label configuration in the continuous domain can be generated when the

algorithm converges. Finally, by thresholding the probabilistic map, we have the binary segmentation result. The effectiveness of the proposed algorithm is demonstrated by our high-quality segmentation results and favorable qualitative and quantitative comparisons with the state-of-the-art methods on a segmentation database.

2.2 Our Approach

Before formulating the segmentation problem, we first define some notations used in this chapter.

$\mathcal{SF}$ and $\mathcal{SB}$ are used to denote the foreground and background confidence seed sets, respectively, which are comprised of the pixels used to train the foreground and background color models. $P = [p_{f1}, p_{f2}, ..., p_{fn}]^T$ is the initial probabilistic map of the image in each iteration associated with the foreground, where p_{fi} is the initial probability that pixel i belongs to the foreground, and n is the number of image pixels. $L = [l_1, l_2, ..., l_n]^T$ denotes the label configuration of the image, where l_i is the label of pixel i taking 1 (foreground) or 0 (background).

2.2.1 A General Framework

Given the seed sets $\mathcal{SF}$ and $\mathcal{SB}$, we can train a color model M_f for the foreground and a color model M_b for the background. However, using only the user-specified seeds, which are usually limited, cannot accurately model the color statistical properties of the foreground and background. Our motivation is to iteratively expand $\mathcal{SF}$ and $\mathcal{SB}$ through obtained label configurations, and then to iteratively achieve more and more precise segmentation results.

The strategy of our framework is as follows. First we construct an objective energy function $f(L, P)$ replying on an initial configuration P from the color models. Then by applying some optimization technique to the optimization

problem $\text{argmin}_L f(L, P)$, we obtain the optimal label configuration L^*. Based on L^*, more pixels can be stamped as foreground or background seeds to update $\mathcal{SF}$ and $\mathcal{SB}$, which are used to refine the color models to generate a better configuration P. Recursively, we can finally reach the optimal label configuration L^* corresponding to a good segmentation result.

It is worth mentioning that this framework is general and can be applied to different color models, energy functions, and optimization schemes. In the next section, we elaborate our iterative optimization based algorithm, including the color model, objective energy function, optimization inference, and iteration strategy.

2.2.2 Iterative Optimization Based Object Extraction

In this section, we specifically describe our iterative optimization based foreground object extraction algorithm in detail. The main parts of our algorithm include the calculation of the initial probabilistic map from the prior color models, the construction of the MRF-based objective energy function, the inference of the closed form global optimal solution in continuous domain, and the iterative optimization strategy, which are all elaborated below.

Initial Probabilistic Map

In our approach, two GMMs each with K components ($K = 10$ in our algorithm) are used to model the color distributions of the foreground and the background, which are defined as

$$GM_f(c) = \sum_{k=1}^{K} w_{fk} g_{fk}(c; \mu_{fk}, \Sigma_{fk}), \tag{2.1}$$

$$GM_b(c) = \sum_{k=1}^{K} w_{bk} g_{bk}(c; \mu_{bk}, \Sigma_{bk}), \tag{2.2}$$

where c denotes the vector consisting of the R, G and B components of a pixel. g_{fk} (g_{bk}) is the k-th Gaussian component of the foreground (background)

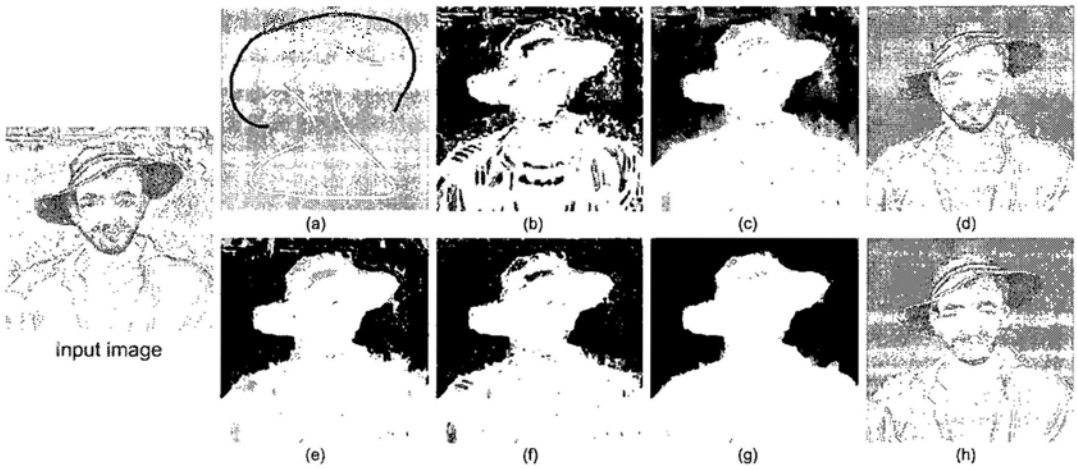

Figure 2.1: Iterative optimization results, where the first row and the second row are for iterations 1 and 5, respectively. In each row from left to right: confidence seed sets $\mathcal{SF}$ (pixels in white regions) and $\mathcal{SB}$ (pixels in black regions) in their corresponding iteration, probabilistic map P, optimal label configuration L^*, and foreground object extraction result by thresholding L^*.

with its mean $\mu_{fk}(\mu_{bk})$ and covariance matrix Σ_{fk} (Σ_{bk}), $0 \leq w_{fk} \leq 1$ with $\sum_{k=1}^{K} w_{fk} = 1$ and $0 \leq w_{bk} \leq 1$ with $\sum_{k=1}^{K} w_{bk} = 1$ are weighting factors.

The parameters w_{fk}, w_{bk}, μ_{fk}, μ_{bk}, Σ_{fk} and Σ_{bk}, $1 \leq k \leq K$, are estimated by $\mathcal{SF}$ and $\mathcal{SB}$, as showed in Fig. 2.1(a) and Fig. 2.1(e). The details about how to obtain the estimations can be found from [6]. Here $GM_f(c)$ $(GM_b(c))$ can be considered as the likelihood of c belonging to the foreground (background). Then each component p_{fi} of the initial probabilistic map P associated with the foreground and the initial probability p_{bi} of pixel i belonging to the background are defined as

$$p_{fi} = \frac{GM_f(c_i)}{GM_f(c_i) + GM_b(c_i)}, \quad p_{bi} = \frac{GM_b(c_i)}{GM_f(c_i) + GM_b(c_i)}. \tag{2.3}$$

Note that as the sizes of the confidence seed sets increase, GM_f and GM_b become more distinguishable from each other in the color space (say, RGB), leading to a more accurate initial probabilistic map P at the beginning of each iteration. From Fig. 2.1 we can see that p_{fi} in Fig. 2.1(f), which is obtained from the updated color models trained by more confidence seeds in Fig. 2.1(e),

provides a better initial probabilistic map than the one in Fig. 2.1(b).

An Objective Function

From Fig. 2.1, it is obvious to see that there are some error parts close to the foreground object boundary in the initial probabilistic map P although P reflects the main object region well. The global optimization of a MRF-based energy function involving spatial coherent constraint can solve this problem.

We define a data cost function $D(l_i, p_{fi}, p_{bi})$ to measure the inconsistency between the assigned label l_i and the initial probabilities, where

$$D(l_i = 1, p_{fi}, p_{bi}) = (1 - p_{fi})^2, \tag{2.4}$$

$$D(l_i = 0, p_{fi}, p_{bi}) = (1 - p_{bi})^2. \tag{2.5}$$

Since $p_{fi} = 1 - p_{bi}$, the above two equations can be combined as the following one:

$$D(l_i, p_{fi}, p_{bi}) = (l_i - p_{fi})^2. \tag{2.6}$$

On the other hand, an natural image usually has the property of pairwise smoothness. Thus we define a smoothness penalty term $w_{ij}(l_i - l_j)^2$ to impose this constraint. Now our objective function is defined as

$$f(L, P) = \sum_{i=1}^{n}(l_i - p_{fi})^2 + \sum_{i=1}^{n}\sum_{j \in \mathcal{N}(i)} w_{ij}(l_i - l_j)^2, \tag{2.7}$$

where $\mathcal{N}(i)$ is a neighborhood of pixel i (8-neighbors of a pixel are chosen as the neighborhood system in our approach), and w_{ij} is a factor defined later. Our aim is to find the label configuration that minimizes $f(L, P)$. This label configuration should best balance the data cost and the piecewise smoothness constraint.

A good smoothness term should be able to not only enforce spatial homogeneity in low contrast regions but also preserve discontinuity in high contrast

areas such as object boundaries. We incorporate the contrast information into w_{ij}, which denotes the affinity value between pixels i and j, and spatially varies to control the degree of the smoothness constraint in different neighborhoods. It is defined as

$$w_{ij} = a \cdot k(i,j) \cdot \exp(-\frac{(p_{fi} - p_{fj})^2}{2\sigma_i^2}), \tag{2.8}$$

where a is a positive factor to control the influence of the smoothness term, σ_i^2 is the deviation of $(p_{fi} - p_{fj})$ over the neighborhood of pixel i, and $k(i,j) = \exp(-d_{ij}^2/2)$ is a Gaussian kernel function to measure the contributions of the neighboring pixels with different distances d_{ij}. Note that if the colors of pixels i and j are similar, $(p_{fi} - p_{fj})^2$ is small; if the colors are quite different, $(p_{fi} - p_{fj})^2$ is larger. Apparently, the local contrast information encoded in w_{ij} imposes the coherence in homogeneous regions and ensures the discontinuity in object boundaries. Therefore, our objective function $f(L,P)$ is contrast-sensitive that is crucial for accurate segmentation, especially in low-contrast or blurred regions.

Here it should be mentioned that the energy function rely on iteratively updated initial probabilistic map P. Therefore the more accurate P is, the better optimization target we obtain, resulting in precise optimal solution.

A Closed Form Solution

With the objective function $f(L,P)$, we need to carry out an optimization process to obtain the optimal label configuration. In hard segmentation, l_i takes 1 or 0. Many methods can be used to perform the minimization of $f(L,P)$, such as graph cuts [12], [42] and belief propagation [23]. However, in our iterative optimization framework, we need to find new confidence seeds based on the optimal label configuration in each iteration. The binary result achieved from graph cuts or belief propagation is not capable of providing this kind of information. However, instead of binary labels, if an optimization

procedure can generate continuous labels in $[0, 1]$ with 1 denoting definite foreground and 0 definite background, we can use these labels to update $\mathcal{SF}$ and $\mathcal{SB}$. Based on the above analysis, we relax the binary labelling problem to the continuous one, ranging from 0 to 1. Consequently, we can obtain a closed form global optimal solution to this continuous optimization problem.

At first, we construct an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices denoting all the image pixels and $\mathcal{E}$ is the set of weighted edges. The elements of the adjacency matrix $W = [W_{ij}]_{n\times n}$ of $\mathcal{G}$ are obtained by

$$W_{ij} = \begin{cases} w_{ij}, & \text{if } i \neq j, j \in \mathcal{N}(i) \\ 0, & \text{if } i \neq j, j \notin \mathcal{N}(i) \\ c, & \text{if } i = j, \end{cases} \tag{2.9}$$

where $c > 0$ is some constant. Let D be an $n \times n$ diagonal matrix with the (i, i)-th entry $D_{ii} = \sum_{j=1}^{n} W_{ij}$. By using a positive constant c, we have $D_{ii} \neq 0$ and W becomes nonsingular. Moreover, c builds up numerical stability for our solution. With the design of w_{ij}, D_{ii} may be very small for some pixels that have edges with small weights. Since the final closed form solution derived below involves D^{-1} and $D^{-\frac{1}{2}}$, a proper c can make the computation of them numerically stable. We find that it is a good choice for c to be comparable with the value of the parameter a in (2.8). Thus, in our algorithm, we set $c = a$.

With the objective function $f(L, P)$ and the corresponding graph $\mathcal{G}$, a closed form solution can be achieved to the following problem:

$$\min_{L} f(L, P) = \min_{L} \big(\sum_{i=1}^{n} (l_i - p_{fi})^2 + \sum_{i,j=1}^{n} W_{ij}(l_i - l_j)^2 \big). \tag{2.10}$$

From (2.10), we can see that different c has no effect on $f(L, P)$ since $c(l_i - l_i) = 0$. Let $R = [r_1, r_2, ..., r_n]^T$, where $r_i = \sqrt{D_{ii}} l_i$, $1 \leq i \leq n$, form a set of medium

variables. Then we have $L = D^{-\frac{1}{2}}R$, and $f(L,P) = f_1(R,P)$ with

$$\begin{aligned} f_1(R,P) &= \sum_{i=1}^{n}(\frac{r_i}{\sqrt{D_{ii}}} - p_{fi})^2 + \sum_{i,j=1}^{n} W_{ij}(\frac{r_i}{\sqrt{D_{ii}}} - \frac{r_j}{\sqrt{D_{jj}}})^2 \\ &= ||D^{-\frac{1}{2}}R - P||_{\mathcal{F}}^2 + 2(R^T\overline{L}R), \end{aligned} \tag{2.11}$$

where $||\cdot||_{\mathcal{F}}$ is the Frobenius norm of a matrix,

$$\sum_{i=1}^{n}(\frac{r_i}{\sqrt{D_{ii}}} - p_{fi})^2 = ||D^{-\frac{1}{2}}R - P||_{\mathcal{F}}^2, \tag{2.12}$$

$$\sum_{i,j=1}^{n} W_{ij}(\frac{r_i}{\sqrt{D_{ii}}} - \frac{r_j}{\sqrt{D_{jj}}})^2 = 2(R^T\overline{L}R), \tag{2.13}$$

and $\overline{L} = D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}$ (called the normalized Laplacian matrix of $\mathcal{G}$).

To minimize $f_1(R,P)$, taking its derivative with respect to R and setting it to zero yields

$$\frac{\partial f_1(R,P)}{\partial R} = 2D^{-\frac{1}{2}}(D^{-\frac{1}{2}}R - P) + 4\overline{L}R = 0, \tag{2.14}$$

which results in $(D^{-1} + 2\overline{L})R = D^{-\frac{1}{2}}P$. Since W and $\overline{L}$ are positive semi-definite, $(D^{-1} + 2\overline{L})$ is nonsingular. Finally, the closed form global solution is:

$$L^* = D^{-\frac{1}{2}}R = D^{-\frac{1}{2}}(D^{-1} + 2\overline{L})^{-1}D^{-\frac{1}{2}}P. \tag{2.15}$$

We consider L^* as a refined probabilistic map with each component being the probability of the pixel belonging to the foreground. An example is showed in Fig. 2.1(c) and Fig. 2.1(g). We can update $\mathcal{SF}$ and $\mathcal{SB}$ to refine the color models by the upper and lower thresholding of L^* (upper threshold $t_u = 0.7$ and lower threshold $t_l = 0.3$ in our work). The algorithm stops at iteration t when $||L_t^* - L_{t-1}^*||_2 < \epsilon$ or $t = T$, where L_t^* is the optimal continuous label configuration in the t-th iteration, ϵ and T are manually set constants to control the algorithm convergence. Finally we use a threshold (0.5 here) on the final continuous global optimum to obtain a binary segmentation result.

Input and Initialization

a) Input: an image and user provided strokes to indicate foreground seed set F and background seed set B.

b) Initialization: confidence seed sets $SF = F$ and $SB = B$.

Iterative Optimization

1. Train the Gaussian mixture color models GM_f and GM_b for the foreground and the background using the current confidence seed sets SF and SB, respectively.
2. Calculate the initial probabilistic map P using the current color models with (2.3).
3. With the energy function $f(L,P)$ in (2.7), find the continuous global optimal label configuration L^* with (2.15).
4. Use the current L^* with the upper and lower thresholds, $t_u = 0.7$ and $t_l = 0.3$, to update the confidence seed sets SF and SB.
5. Repeat steps 1 -- 4 until convergence.
6. Obtain a binary segmentation result by thresholding (with 0.5) the final global optimal probabilistic map L^*.

Figure 2.2: Our framework for foreground object extraction.

Our foreground object extraction framework is summarized in Fig. 2.2. It should be noticed that since the global optimum can be achieved in the optimization procedure, the main factor affecting the result quality is the energy construction, in other words, the accuracy of the initial probabilistic map. Thus we iteratively refine the color models trained from updated confidence seeds to ensure the accuracy improvement of the initial probabilistic map. Besides, the user provided information is iteratively propagated and expanded in our algorithm, which makes it not sensitive to the number and locations of user provided seeds and require less cost of users' interaction and attention.

2.3 Application Extensions

Besides the application of two class interactive image segmentation, our algorithm can be easily extended to some other tasks such as multi-object segmentation and object extraction from a group of images sharing the same object with similar color and texture.

In the context of the multi-object (say, m foreground objects) segmentation problem, we can apply our algorithm with respect to each object and obtain m final probabilistic maps $\{l^*_{oi}\}^{i=1,...,n}_{o=1,...m}$ with each corresponding to one object. At the end, pixel i is assigned to label $o^*_i \in \{1, 2, ..., m\}$, where $o^*_i = \text{argmax}_{o=1,...,m} l^*_{oi}$. To extract similar objects of interest from a group of images using only the user provided seeds in one reference image, we can first find the final object color/texture model in the reference image by our approach, and then utilize the model to obtain the initial probabilistic maps and optimal label configurations of the other images until convergence, which naturally leads to object extraction from all the images in the group.

Moreover, by applying the thresholding process in our algorithm, the final probabilistic map of the image we obtain can be directly transformed to a trimap for the initialization of image matting. Our experiments show that the trimaps generated by our algorithm are much better than those generated by the uniform boundary erosion and dilation of rough segmentation results regardless of local image characteristics [68].

2.4 Experimental Results

In our experiments, we compare our algorithm with belief propagation [23], graph cuts [9], the adaptive GMMRF (AGMMRF) based algorithm [7], and random walk based method [28], which are abbreviated to BP, GC, AGMMRF, and RW, respectively. Note that they are not iterative optimization approaches. The objective energy function optimized by BP and GC is from

[68]. The parameters used in these algorithms are all best tuned. In our algorithm, the number of iterations is set to 5. The algorithms proposed in [46], [68], and [88] are not compared here because their outputs are obtained through multiple user interactions to fine tune the results.

Before showing the qualitative and quantitative comparisons between our algorithm and other related approaches, to demonstrate the effectiveness of our iterative optimization scheme, we first present the visual comparisons of the binary segmentation results generated in the first and the last iterations of our algorithm in Fig. 2.3 on two natural images. Similar observation can be found in Fig. 2.1, which shows the results of our algorithm on a challenging case with complicated foreground and background color characteristics, including the confidence seed sets, initial probabilistic maps, optimal label configurations, and corresponding binary results obtained in iterations 1 and 5 of our algorithm. We can see the label configuration and binary segmentation result in Fig. 2.1(g) and Fig. 2.1(h) are much better than those in Fig. 2.1(c) and Fig. 2.1(d), and our final segmentation results in Fig. 2.3 are much more accurate than those obtained in first iteration. Therefore the advantage of our iterative refinement scheme of the initial probabilistic maps for achieving precise object extraction results is explicitly demonstrated, especially in the challenging case as showed in Fig. 2.1. In addition, with the same user specified seeds and prior color models, the object extraction result obtained by RW [28] is essentially the same as the one obtained in the first iteration of our algorithm since they both are the binary versions of the continuous global optimal label configuration for the same objective energy function $f(L, P)$. Thus the above experimental comparisons display the better performance of our algorithm than RW as well.

The second experiment is the visual comparison of the results obtained by BP, GC, and our algorithm. Since the AGMMRF algorithm is not available publicly, its results are not included in this experiment. Fig. 2.4 shows the

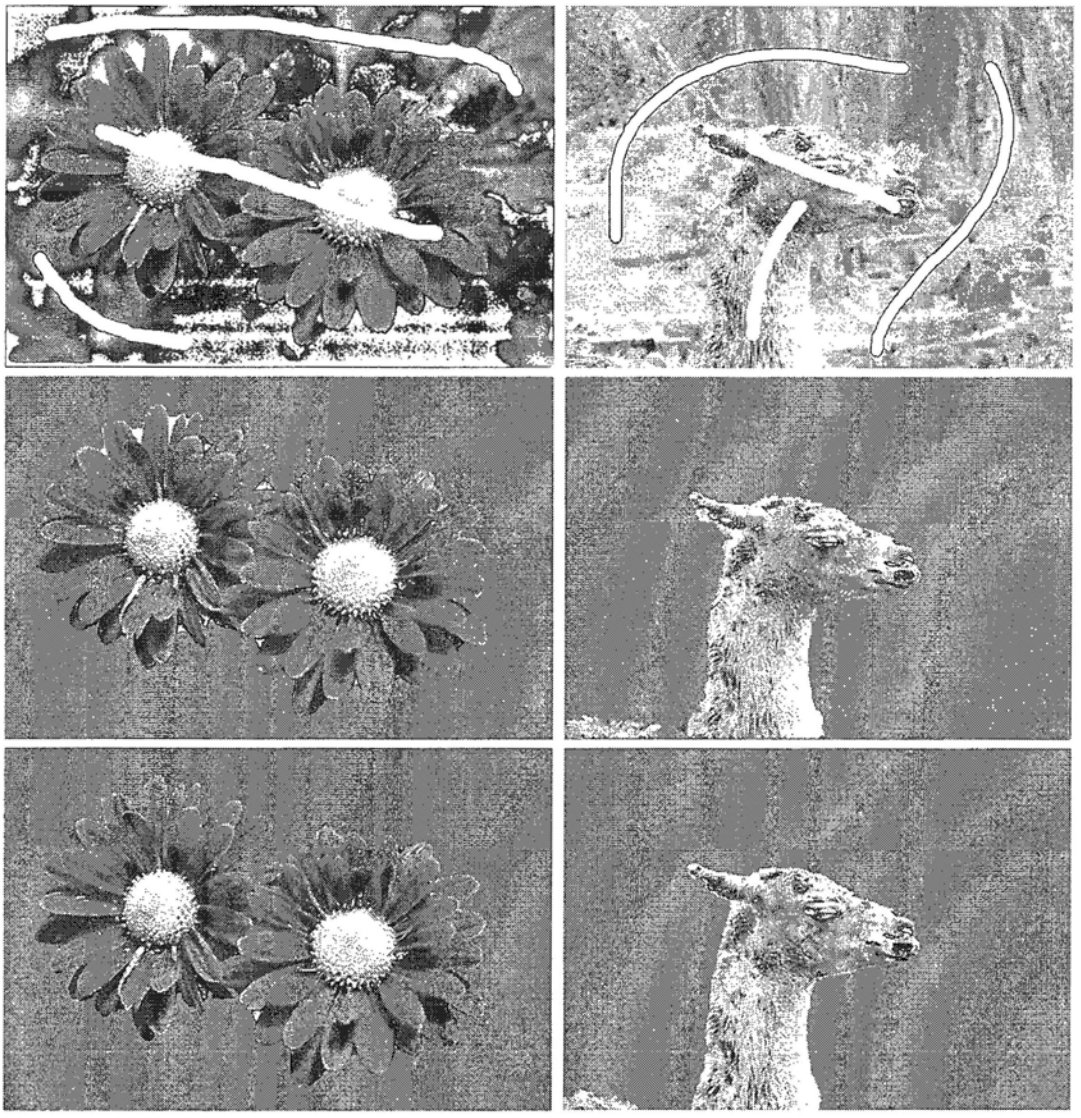

Figure 2.3: Foreground object extraction results on two natural images. From top to bottom: input images with user specified strokes, the results generated in the first and the last iterations of our algorithm.

results on two images with initial simple user indications of the foreground and the background. These results clearly demonstrate the better performance of our algorithm with the precise extraction of the foreground objects, especially on the edges of the central flower in the first image, and on the hair and the collar near the neck of the person in the second image.

To make our experimental comparison more convincing, we conduct a quantitative comparison among BP, GC, AGMMRF, RW, and our algorithm. The authors of AGMMRF [7] provided a database containing 50 natural images with both user initial foreground and background inputs and the ground truth results[1] . Although there are many segmentation databases with ground truth results, the database in [7], to the best of our knowledge, is the only one with seed regions provided, as showed in Fig. 2.5. With the results generated by these algorithms based on the same user provided information, we can quantitatively compute their average error rates using the ground truth. Table. 2.1 clearly shows that our algorithm performs the best. This quantitative comparison over a large set of natural images convincingly demonstrates the excellent performance of our algorithm. Fig. 2.5 illustrates some typical visual and quantitative outputs of our approach, from which we can see that our results are of great visual quality.

Here it should be mentioned that the provided information about the foreground and background for each image in this database is abundant since all pixels are labelled except for a narrow band around the boundary of the object, as showed in Fig. 2.5. Therefore, the virtue of our proposed framework that the accuracy is improved iteratively as the expansion of the seed sets is not fully exploited. In practical applications, the user is not expected to give so abundant information. When fewer seeds are provided, the improvement between our algorithm and the others is more obvious, as showed in Fig. 2.3 and Fig. 2.4.

[1] Available at http://research.microsoft.com/vision/cambridge.

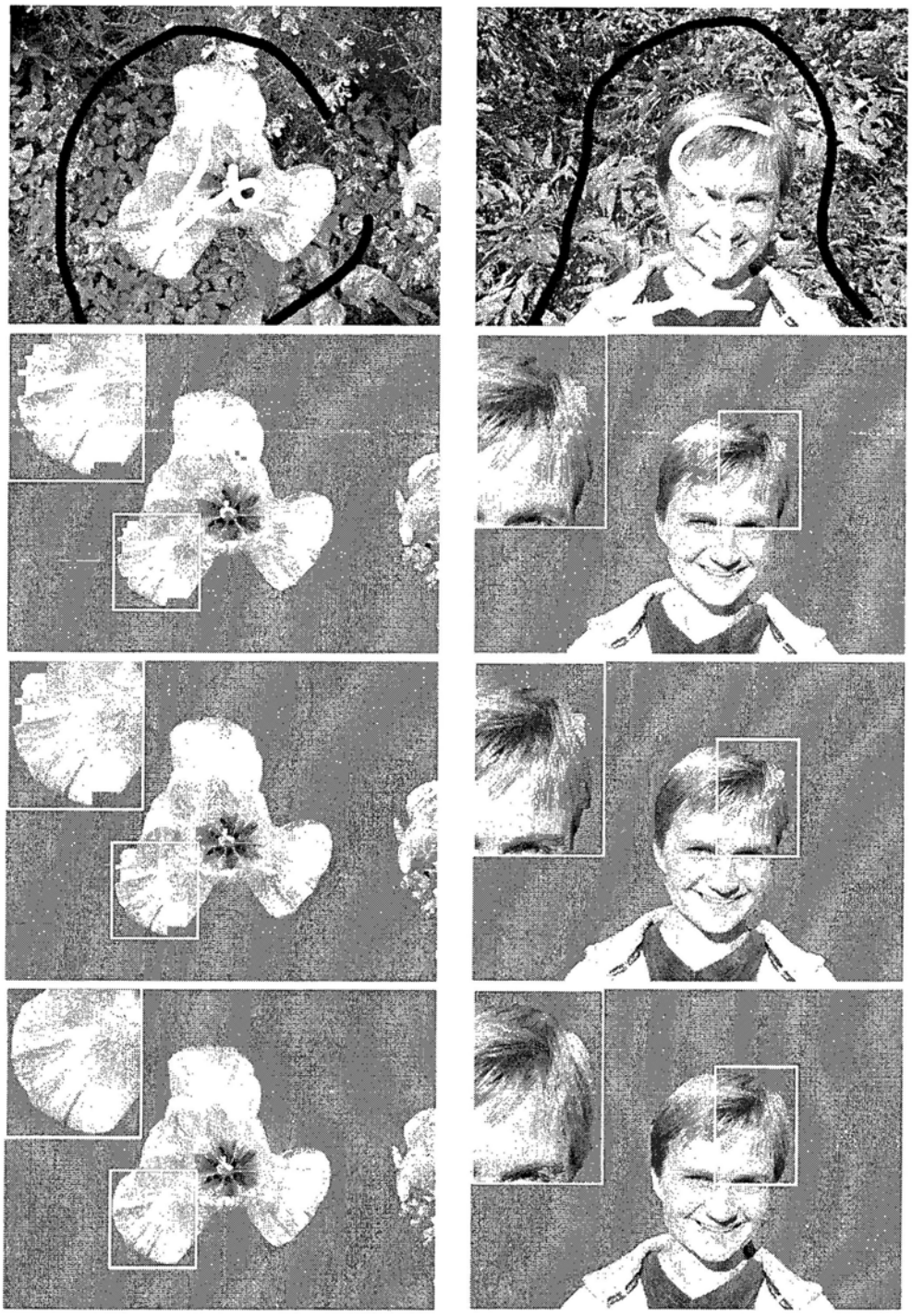

Figure 2.4: Results on "Flower" and "Person" images. From top to bottom: input images with user guided strokes, the results of BP, GC, and our algorithm. We also zoom in some regions for better observation.

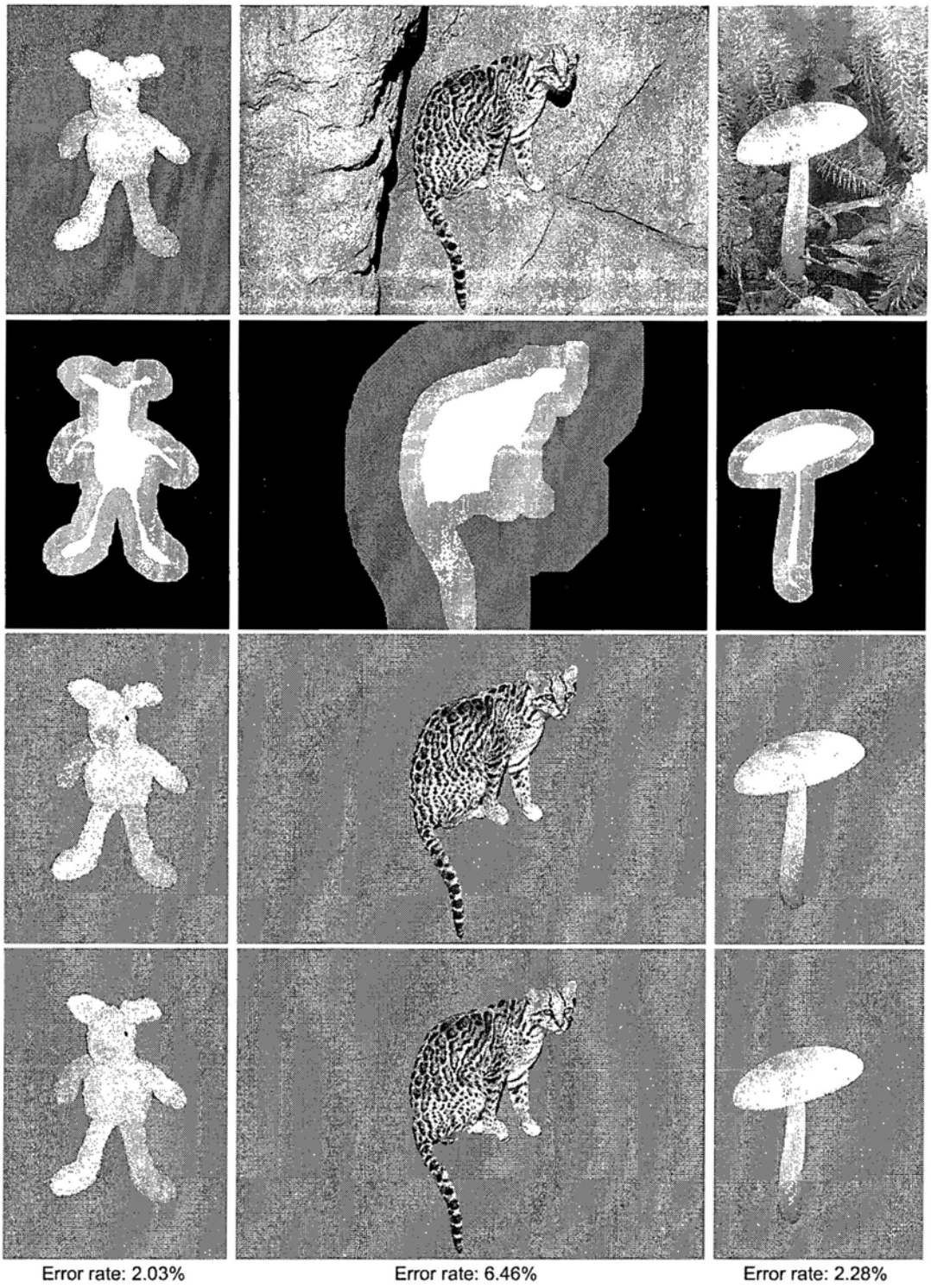

Figure 2.5: Results on "Teddy", "Cat", and "Mushroom" images in the database [7]. From top to bottom: input images, provided seed images, ground truth results, and our object extraction results. The error rates computed from our results and ground truth are showed here as well.

Algorithm	BP	GC	AGMMRF	RW	Ours
Error rate (%)	8.4	7.2	7.9	5.8	5.5

Table 2.1: Comparison of the error rates by the algorithms on all the 50 natural images in the database [7].

To validate the good performance of our algorithm on foreground object extraction with little user interaction, we apply our algorithm to many natural images with diversified objects and background. Fig. 2.6 shows some excellent results by our algorithm.

2.5 Conclusions

In this chapter, an iterative optimization based framework is proposed to address the problem of foreground object extraction from an image. We model the problem as an iterative energy minimization problem to find the optimal label configuration. In our algorithm, the foreground and background color models are iteratively refined by expanding the confidence seed sets, which improves the accuracy of the initial probabilistic map at the beginning of each iteration. Based on the initial probabilistic map and the boundary and coherent information in the image, we construct an MRF-based energy function in our optimization problem. Then by relaxing the hard segmentation to the soft one, a closed form global optimal solution can be achieved, which can be regarded as a refined probabilistic map providing us the clues for updating the confidence seed sets. The more accurate initial information we use in the objective energy construction, the more precise label configuration we can obtain by energy minimization. Therefore, through the iterative optimization scheme, high-quality foreground object extraction results can be achieved by our algorithm. We have compared our algorithm with several related approaches on many natural images both visually and quantitatively. The results demonstrate the excellent performance of our algorithm.

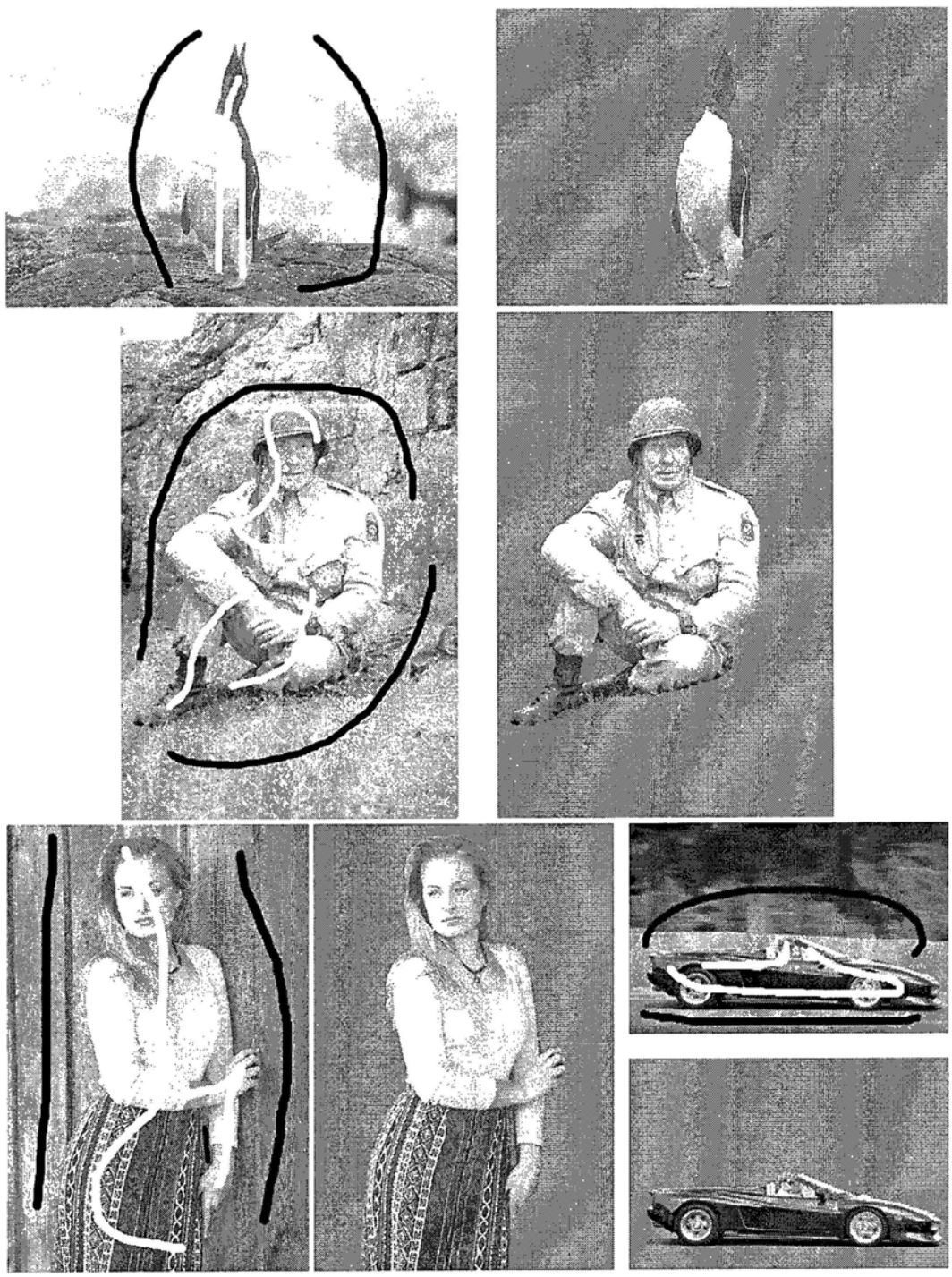

Figure 2.6: Some experimental results obtained by our algorithm.

Chapter 3

Video Completion via Spatial-Temporal Global Optimization

The target of video completion is to restore the spatial-temporal missing regions of a video in a visually plausible way. These missing pixels are caused by some damage to the video or the removal of unwanted objects. In this chapter, a novel global optimization based approach is proposed for video completion. Our algorithm consists of two stages: motion field completion and color completion via global optimization. First, local motions within the missing parts are completed patch-by-patch greedily using pre-computed available motions in the video. Then the missing regions are filled by sampling patches from available parts of the video. We formulate the video completion as a global energy minimization problem by Markov random fields (MRFs). Based on the completed motion field of the video, a well-defined energy function involving both spatial and temporal coherence relationship is constructed. Belief propagation is used to solve the problem. To avoid the computational impracticability caused by the large number of label candidates in the optimization process of belief propagation, we utilize a coarse-to-fine optimization scheme whose

essential idea is to carry out belief propagation multiple times with sharply reduced numbers of label candidates, instead of running belief propagation only one time with a large number of label candidates. Based on our motion guided spatial-temporal global optimization framework, good video completion results are obtained in the experiments, which demonstrate the excellent performance of our algorithm.

3.1 Introduction

Image and video completion, also known as image and video inpainting, are of great importance in many computer vision and computer graphics applications such as photo and movie editing and post-production. Their goal is to automatically reconstruct missing regions in an image/video in a non-detectable form, which is challenging and wide-open for researchers.

3.1.1 Related Work

A number of methods have been proposed to deal with the problem of image completion [3, 45, 21, 19, 43, 83]. The algorithm presented in [3] is PDE-based and can effectively restore small missing portions with strong structures in the image while failing to reconstruct large holes in texture areas. Following [3], the authors in [45] solve the problem based on prior image knowledge by using global image statistics. Inspired by the texture synthesis technique in [22], exemplar-based technique [19], which repairs the missing regions by merging with the best source patches obtained by searching the available regions of the image, are more suitable for the completion of large textured holes. However, this greedily filling process may lead to visual inconsistency since in the greedy scheme a pixel cannot change its value once it has been filled in each processing iteration. Besides, it may lose some important structure information, and its high computational complexity is also a limitation for practical using. In

[19], the order by which the process is carried out is well studied, and the confidence map is proposed to guide the filling priority of each pixel to enforce the structure propagation. To overcome the shortcomings of the greedy completion techniques, global optimization based algorithms have been proposed recently [43, 83]. In [43], image completion problem is posed in the form of a discrete global optimization problem with a well-defined objective function, and a new scheme called Priority-BP is proposed to solve the optimization problem efficiently. The algorithm proposed in [83] follows the spirit of [43] and add a structure constraint into the objective function to ensure the proper structure propagation.

Intuitively, video completion can be considered as an extension of 2D image completion to 3D video completion. However, compared with image completion, video completion is more challenging in two aspects. First, it is more important to enforce temporal coherency than spatial coherency in the completion process since human visual system is more sensitive to motion distortion. Simply treating video as a set of independent images and then applying an image completion method to them are not advisable. Second, video completion contains much more data and thus needs more efficient algorithms.

One of the first efforts for video completion is made in [2], which is a PDE-based approach and handles the video frame by frame. It works well in small structured holes, but fails to complete large holes in a video sequence and does not utilize the temporal information from the video.

Many segmentation based or layer extraction based video completion algorithms are developed recently [38, 89, 71, 90, 74, 62]. The algorithm in [38] extracts the background and foreground and repairs them separately frame by frame with different strategies. Although achieving impressive results, this method is limited owing to its complexity, requirement of user interaction to manually indicate layers with different depths, and its restriction to periodic motion. Another motion layer segmentation based approach in [89] combines

motion compensation and region completion to restore the motion layers in the reference frame and then transfers the information to other frames to generate a completed video sequence. Similarly, the authors in [71] utilize the motion manifold obtained after the foreground and background segmentation to reconstruct the missing parts of the video. This approach can handle camera motion and distortion, but it is still restricted to the periodic motion. An interesting algorithm in [62] first builds the mosaics of background, foreground, and optical flow, and then fulfills the motion inpainting and background inpainting in turn guided by the pre-computed priority. While the layer extraction based algorithms are effective for video completion, it is very difficult to obtain accurate layers in general, especially for the scenes with complex motions. In addition, all these methods are restricted to the videos with only periodic motion.

Extending the exemplar-based approach to video completion, the algorithm in [86] treats video completion as a global optimization problem with a well-defined objective function. It fills the missing portions by exhaustive searching for the most similar space-time source patches available in the video and weighted blending of the selected candidates. The spatial-temporal consistency is enforced by the global optimization. However, the algorithm also relies on the assumption of periodic motion and is computationally inefficient due to the pixel-by-pixel filling process and the exhaustive search for candidates. In addition, the similarity-based merging scheme for calculating each pixel's value leads to noticeable blurring artifacts. The authors in [39] present another exemplar-based approach which uses tracking to reduce the search space and applies graph cuts algorithm for merging the source and target patches to maintain details.

A newly published algorithm in [75] restores local motion in the holes of the video by sampling spatial-temporal motion patches, instead of directly using the color copy-and-paste scheme. With the completed motion volume, color

is propagated into the holes to complete the video. As discussed in [75], the algorithm is more sensitive to noise than directly using color sampling and does not work well for the completion of videos with large motions. Moreover, the results of this algorithm have blurring effects due to the weighted average scheme in color propagation. It is worth noting that the assumption in this algorithm that motion information is sufficient to fill holes in videos cannot be true in some cases such as the video with the same missing regions in all the frames, in which little motion appears.

3.1.2 Our Framework

In our work, we propose a motion guided spatial-temporal global video completion algorithm to combine motion field completion and global exemplar-based color completion. The two steps in our algorithm are briefly described as follows.

Motion field completion. We use hierarchical Lucas-Kanade optical flow computation method [53, 8] to calculate local motion vector of each pixel in the video except the pixels in the holes. Then the motion in the data missing regions is completed patch-by-patch using the computed motion vectors in the available regions, where a copy-and-paste scheme is adopted based on a pre-defined similarity measurement.

Spatial-temporal global optimization. The global exemplar-based color completion in our algorithm is formulated as a discrete global optimization problem with a well defined objective function, which is constructed under the Markov random field (MRF) models incorporating both spatial and temporal constraint terms to enforce the spatial and temporal consistency. The design of the temporal constraint term is guided by the completed motion fields obtained in our first stage above. To carry out the optimization process, we propose a coarse-to-fine belief propagation (BP) technique, which can deal

with the intolerable computational cost caused by the large number of label candidates in the optimization.

Our algorithm combines the motion and color information to accurately fill the missing parts of the video, which preserves the temporal consistency based on the completed motion field, and globally optimizes the color completion process. It avoids the blurring effect caused by the sampling and blending process, while maintaining the video details and structures well. Besides, the proposed framework unifies the problems of image completion and video completion and solves them in a consistent form. Moreover, our algorithm is not restricted to videos containing periodic motion only and can handle a wide variety of videos, producing visually natural results without obvious artifacts. The experimental results have demonstrated the excellent performance of our algorithm.

3.2 Motion Guided Spatial-Temporal Global Optimization

We formulate the video completion problem as a labeling problem modeled by discrete Markov Random Fields (MRFs). The target regions are filled globally by using exemplar patches taken from the source region of the video.

Let $f = \{f^t\}_{t=1}^T$ be the input video of T frames with the region $\Pi = \{\Pi^t\}_{t=1}^T$, where Π^t is the region of f^t. Suppose that $\Phi = \{\Phi^t\}_{t=1}^T$ is the source region and $\Omega = \{\Omega^t\}_{t=1}^T$ is the target region (data missing region). Then we have $\Phi + \Omega = \{\Phi^t + \Omega^t\}_{t=1}^T = \{\Pi^t\}_{t=1}^T = \Pi$.

3.2.1 Model Construction

Firstly, we sparsely sample each frame with a horizontal spacing hs and vertical spacing vs. Then we can obtain sampled pixels $P = \{\{p_i^t\}_{i=1}^{N^t}\}_{t=1}^T$ in the target

region, where N^t is the number of sampled pixels in the target region of the tth frame. The process of video completion is to fill the target region by pasting some $w \times h$ patches taken from the source region to the locations centered at the positions in P.

We construct an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \{\{v_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ contains all the pixels in P, and $\mathcal{E}$ is the set of edges connecting each node to nodes in its neighborhood system. A 4-neighborhood system is used to enforce the *spatial consistency* constraint in the same frame, while some nodes in sequential frames, called *temporal neighbors*, are included in our neighborhood system to enforce the *temporal consistency* constraint. The detail of temporal neighbors is described as follows.

Let $\mathcal{L} = \{l_k\}_{k=1}^{K}$ be the set of label candidates containing all the $w \times h$ patches taken from the source region. Then our labeling problem is to find the best label configuration $X = \{\{x_i^t\}_{i=1}^{N^t}\}_{t=1}^{T}$ such that an energy function is minimized, where $x_i^t \in \mathcal{L}$ and $x_i^t = l_k$ represents that the label (patch) for node v_i^t is l_k. In our approach, the best label configuration is estimated by minimizing the following energy function:

$$E(X) = E_s(X) + \alpha E_t(X), \tag{3.1}$$

where $E_s(X)$, called *spatial term*, enforces the spatial consistency constraint, $E_t(X)$, called *temporal term*, enforces the temporal consistency constraint, and α is a positive constant to balance these two terms. Fig. 3.1 illustrates the spatial and temporal terms. The details of them are discussed in the following two sections. It is worth noting that if α is set to 0, our energy function has only spatial term to be optimized, which is a suitable model for the problem of single image completion.

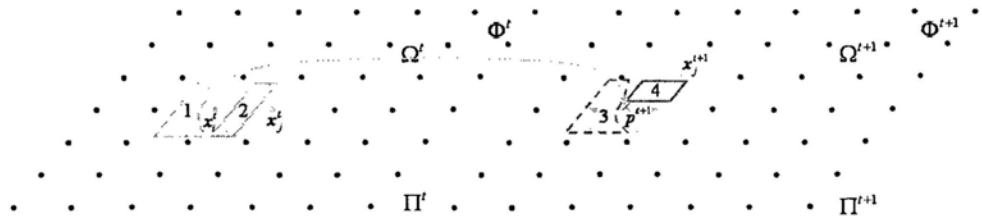

Figure 3.1: Illustration of the spatial and temporal terms. The dots indicate the sampled pixels which correspond to the vertices in the graph. Regions 1, 2, 3, and 4 are overlapping parts for the calculation of $E_1(x_i^t)$, $E_2(x_i^t, x_j^t)$, $E_3(x_i^t)$, and $E_4(x_i^t, x_j^{t+1})$, respectively. The patch centered at p^{t+1} (the cross) is copied from x_i^t.

3.2.2 The Spatial Term

The spatial term is used to enforce the spatial consistency, of which the implied assumption is that the overlapping parts of patches should have consistent texture and structure information in the patch pasting process. Based on the MRF model, it is defined as:

$$E_s(X) = \sum_{v_i^t} E_1(x_i^t) + \sum_{(v_i^t, v_j^t) \in \mathcal{E}_s} E_2(x_i^t, x_j^t), \tag{3.2}$$

where $\mathcal{E}_s$ is the spatial 4-neighborhood system, $E_1(x_i^t)$ is the cost for label x_i^t, and $E_2(x_i^t, x_j^t)$ is the consistency cost for label pair (x_i^t, x_j^t).

Similar to [19] and [83], the *confidence map* is also used in our algorithm to represent the importance of nodes in the filling process. In the map, the pixels in the target region closer to the source region in each frame have larger confidence values. Fig. 3.2 is an example of the map of the confidence values.

With the confidence map, the cost for label x_i^t is defined as:

$$E_1(x_i^t) = C_i^t \cdot d(x_i^t, \Phi^t) \tag{3.3}$$

where C_i^t is the confidence value for node v_i^t and $d(x_i^t, \Phi^t)$ constrains the synthesized patch x_i^t to match well with the source region which overlaps with the node v_i^t. $d(x_i^t, \Phi^t)$ is calculated as the sum of the squared differences (SSD)

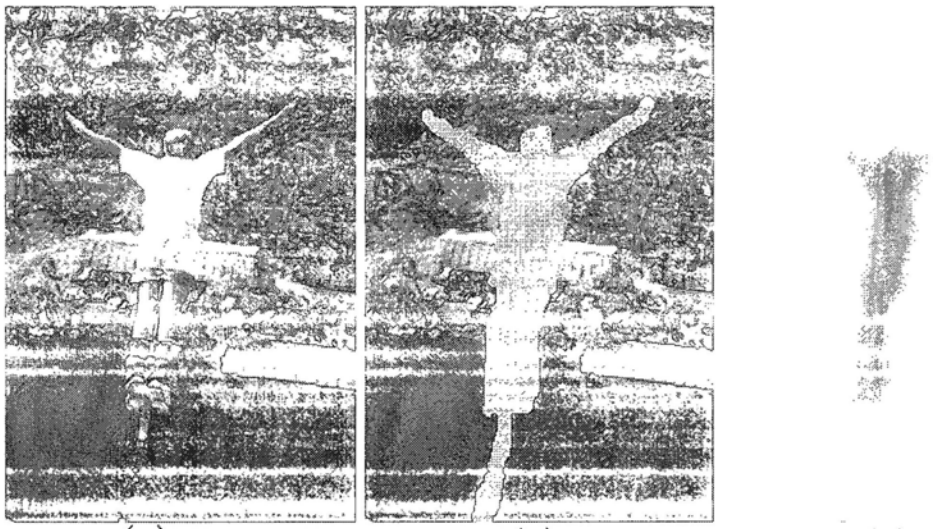

Figure 3.2: An example of the confidence map. (a) One frame of an input video. (b) The mask (in green) of the object to be removed. (c) The confidence map in the mask, in which the brighter a pixel is, the larger the confidence value is.

of pixel colors in the overlapping part between x_i^t and Φ^t (e.g., region 1 surrounded by the red dashed curve in Fig. 3.1). When x_i^t and Φ^t do not overlap, $E_1(x_i^t) = 0$.

Since structure (e.g., lines, curves) continuity is important for human perception and texture reflects the details of an image, we incorporate both structure and texture in the completion process. The consistency cost $E_2(x_i^t, x_j^t)$ in (3.2) is thus defined as

$$E_2(x_i^t, x_j^t) = \left[\frac{C_i^t + C_j^t}{2}\right] \left[\lambda_1 E_2'(x_i^t, x_j^t) + \lambda_2 E_2''(x_i^t, x_j^t)\right], \tag{3.4}$$

where C_i^t and C_j^t are the confidence values of nodes v_i^t and v_j, respectively, $E_2'(x_i^t, x_j^t)$ is used to enforce consistency for texture propagation, $E_2''(x_i^t, x_j^t)$ is for structure propagation, and λ_1 and λ_2 are two factors to balance E_1, E_2', and E_2''.

In our algorithm, $E_2'(x_i^t, x_j^t)$ is computed by

$$E_2'(x_i^t, x_j^t) = d(x_i^t, x_j^t), \tag{3.5}$$

where $d(x_i^t, x_j^t)$ is the SSD in the overlapping part between the patches centered at nodes v_i^t and v_j^t (e.g., region 2 surrounded by the red solid curve in Fig. 3.1). $E_2''(x_i^t, x_j^t)$ is computed by

$$E_2''(x_i^t, x_j^t) = d_{gh}^2(x_i^t, x_j^t) + d_{gv}^2(x_i^t, x_j^t), \tag{3.6}$$

where $d_{gh}(x_i^t, x_j^t)$ and $d_{gv}(x_i^t, x_j^t)$ are the gradient differences between x_i^t and x_j^t in the image horizontal and vertical directions, respectively. The gradient of a patch is denoted as the maximum gradient of the pixels in the patch, which describes the structure of the patch. The constraint of gradient consistency propagates the structure information.

With the design of the spatial term $E_s(X)$, which is an MRF-based energy function incorporating the texture and structure information, our framework well models the problem of image completion by removing the temporal term ($\alpha = 0$). The experimental results demonstrate the excellent performance of our framework for image completion.

3.2.3 The Temporal Term

The temporal term enforces the consistency constraint between two sequential frames, meaning that two corresponding patches in two sequential frames should have consistent colors. In our algorithm, the correspondence is found via local motion estimation. There exits many optical flow algorithms [53, 34, 60] for motion estimation. In our algorithm, the hierarchical Lucas-Kanade algorithm [73] is used.

If dense motion is estimated, correspondence for all patches without missing pixels in a video can be constructed. The current problem is that in the context of video completion there are many data missing regions in the input video, and thus optical flows cannot estimate the motions for the pixels in these regions. To obtain a completed dense motion map, we first calculate the motions for all the pixels in the source region. Then motion field transfer technique [75]

is used to compensate the motion in the target region, which utilizes a copy-and-paste scheme to carry out the motion completion based on some similarity measurement criteria. The details are described as follows.

Motion Completion

In our approach, a copy-and-paste process, i.e., copying the best motion patch from the source region and pasting it to the target region, fills in the motion in the target region patch by patch. Here the compensation process, which is different from the global color propagation, is completed greedily.

Motion completion in our approach starts from the boundary of the target region and goes towards the inner region. When one motion patch is filled, the new target region is obtained by assigning the filled pixels to the source region.[1] At each time, the selected target patch is centered at the boundary of the target region. Therefore, each selected target patch includes both pixels in the source region and pixels in the target region. The selection order of the target patch is determined by the number of pixels belonging to the source region in the target patch. The target patch with the largest number is selected for being filled in first.

Before defining the criteria for choosing the best source patch for a target region, the motion difference measurement is introduced. Suppose that the motion vector of pixel q in frame t is $(u_q^t, v_q^t)^T$. If we regard the 2D motion as a 3D vector in the spatial-temporal domain by padding the temporal element t, then the 3D vector is defined as $\mathbf{m}_q^t = (u_q^t, v_q^t, t)^T$. The difference between two motion vectors $\mathbf{m}$ and $\mathbf{m}'$ is defined as the angular difference [1]:

$$d_m(\mathbf{m}, \mathbf{m}') = 1 - \frac{\mathbf{m} \cdot \mathbf{m}'}{|\mathbf{m}||\mathbf{m}'|} = 1 - \cos\theta, \tag{3.7}$$

where θ is the angle between the two motion vectors $\mathbf{m}$ and $\mathbf{m}'$. Since this

[1] Target region update is only for motion compensation. For color completion, the target region is kept unchanged.

expression is defined in homogenous coordinates, the measurement relies on the differences in both direction and magnitude.

For a source motion patch A_s and a target motion patch A_t (A_s and A_t are 3D in the spatial-temporal domain), the difference measurement between them is defined as:

$$d_{mp}(A_s, A_t) = \frac{1}{|Q_s|} \sum_{q_t \in Q_s} d_m(\mathbf{m}_{q_s}^{t_s}, \mathbf{m}_{q_t}^{t_t}), \tag{3.8}$$

where Q_s is the set of points in A_t belonging to the source region, $|Q_s|$ is the number of pixels in Q_s, q_s and q_t are two corresponding pixels in A_s and A_t respectively, and t_s and t_t are the frames in which A_s and A_t are respectively. Then for A_t the best source patch $\widehat{A}_s$ is chosen by minimizing (3.8):

$$\widehat{A}_s = \underset{A_s}{\operatorname{argmin}}\, d_{mp}(A_s, A_t). \tag{3.9}$$

Temporal Energy Function

Once the motion compensation is completed, a dense motion map can be obtained for all the pixels in the input video. With the estimated local motions of all the pixels, the relationship between two sequential frames can be constructed. Before defining the temporal term, the temporal neighborhood is introduced first. For a sampled pixel p_i^t whose corresponding graph vertex is v_i^t, if its motion is known, then we can find its corresponding point p^{t+1} in the next frame. We call the set of the four vertices in frame $t+1$ corresponding to the four sampled vertices nearest to p^{t+1} the *temporal neighborhood* of v_i^t (see Fig. 3.3). If v_j^{t+1} is a temporal neighbor of v_i^t, then we denote them as $(v_i^t, v_j^{t+1}) \in \mathcal{E}_t$.

The definition of the temporal term is similar to the spatial term, which is expressed as the sum of two parts:

$$E_t(X) = \sum_{v_i^t} E_3(x_i^t) + \sum_{(v_i^t, v_j^{t+1}) \in \mathcal{E}_t} E_4(x_i^t, x_j^{t+1}), \tag{3.10}$$

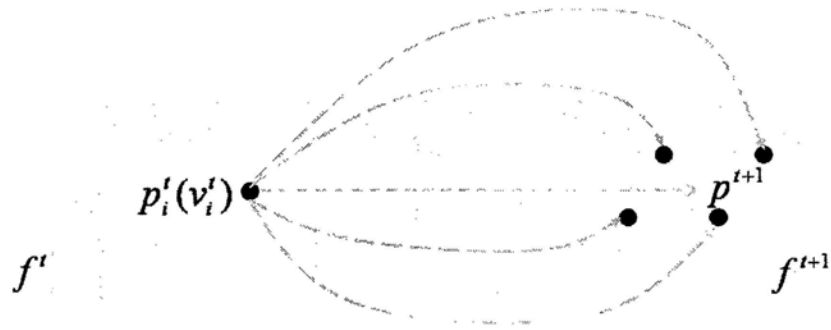

Figure 3.3: Illustration of the temporal neighborhood system. p_i^t is a sampled pixel in frame t with its corresponding graph vertex v_i^t. The cross in frame $t+1$ is the corresponding position of p_i^t based on the motion estimated. Then vertices corresponding to the four nearest sampled pixels in frame $t+1$ are the temporal neighbors of vertex v_i^t (connected with v_i^t by red dashed lines).

where $\mathcal{E}_t$ is the temporal neighborhood system, $E_3(x_i^t)$ represents the temporal inconsistency between x_i^t and its corresponding source region in frame $t+1$, and $E_4(x_i^t, x_j^{t+1})$ represents the temporal inconsistency between x_i^t and x_j^{t+1}.

The definitions of $E_3(x_i^t)$ and $E_4(x_i^t, x_j^{t+1})$ are similar to those of $E_1(x_i^t)$ and $E_2(x_i^t, x_j^t)$, respectively. Compared with $E_1(x_i^t)$ and $E_2(x_i^t, x_j^t)$, there is no confidence and structure information in $E_3(x_i^t)$ and $E_4(x_i^t, x_j^{t+1})$. They are defined as:

$$E_3(x_i^t) = d(x_i^t, \Phi^{t+1}), \tag{3.11}$$

$$E_4(x_i^t, x_j^{t+1}) = d(x_i^t, x_j^{t+1}). \tag{3.12}$$

As in the spatial term, here d is the SSD value in the overlapping region of the two parts. Suppose that the corresponding point of p_i^t in frame $t+1$ is p^{t+1}. To calculate $d(x_i^t, \Phi^{t+1})$ and $d(x_i^t, x_j^{t+1})$, the first step is to put the center of the patch x_i^t at p^{t+1}. Then $d(x_i^t, \Phi^{t+1})$ is the SSD value in the overlapping region between the patch and Φ^{t+1} (e.g., region 3 surrounded by the purple dashed curve in Fig. 3.1), and $d(x_i^t, x_j^{t+1})$ is the SSD value in the overlapping region between the patch and x_j^{t+1} (e.g., region 4 surrounded by the purple solid curve in Fig. 3.1).

3.3 Optimization Scheme

Recall our discussion in Chapter 1. The problem of minimizing the energy function (3.1) is an NP-hard problem. Belief propagation can find a local optimum for such an MRF energy function. The max-product and sum-product are two typical BP algorithms [23]. In our algorithm, the max-product algorithm is used since it is less sensitive to numerical inaccuracy and is derived directly to the problem of energy minimization.

The max-product BP works iteratively by passing messages along the graph. For a graph with N nodes and K label candidates, the running time for T iterations is $O(TNK^2)$. In our video completion approach, the main problem with such a standard BP algorithm is that the number of label candidates K is too large to be used in practice. For instance, there may be more than $100,000$ patches for a video clip with 30 frames of size 256×256. It takes the standard BP several days to run to achieve the video completion result. Therefore, to overcome this problem, we use a coarse-to-fine scheme to greatly reduce the computational time. The main idea of this scheme is to perform BP R times with K_r label candidates in each time, $r = 1, ..., R$, instead of running BP only one time with K candidates, where K_r is much smaller than K. The steps of our coarse-to-fine strategy are described as follows.

First, an R-level patch pyramid is constructed (see Fig. 3.4). In the bottom level of the pyramid, elements are all label candidates $\mathcal{L}$ taken from the source region. Then we use the k-means algorithm to classify all patches in $\mathcal{L}$ into $\frac{K}{K_R}$ clusters. The means of the $\frac{K}{K_R}$ clusters are regarded as the elements in level $R-1$ of the pyramid. Then $\frac{K}{K_R}$ elements are clustered into $\frac{K}{K_R K_{R-1}}$ clusters and the mean values are the elements of level $R-2$. The rest is deduced similarly and we can obtain the patch pyramid with the element number K_1 in the top level.

After the patch pyramid is completed, we perform BP R times from the

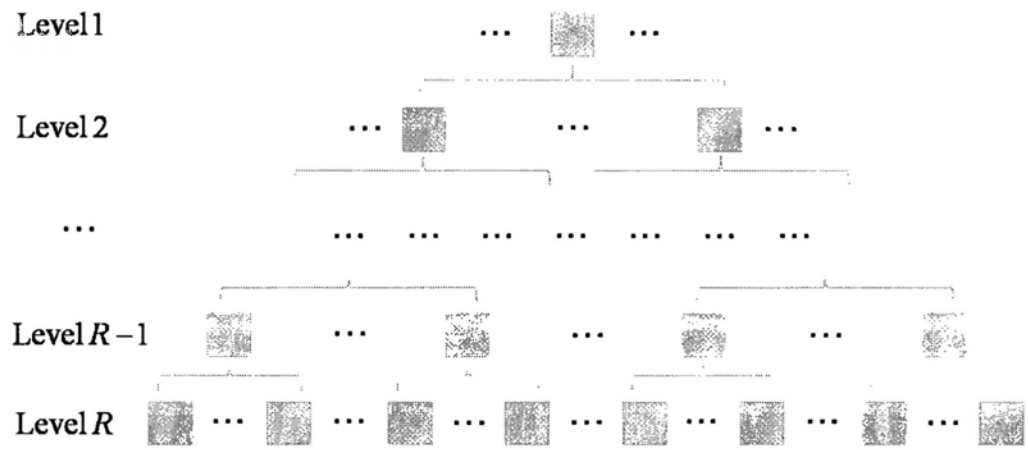

Figure 3.4: The pyramid of patch candidates. Except in the bottom level, patches in level R of the pyramid are the mean values of their corresponding patch sets in level $R+1$.

top level to the bottom level. Except for BP execution in the first level, where K_1 elements are regarded as the label candidates for all the nodes, in level R's BP optimization, $R \neq 1$, different nodes have different label candidates. For each node, its label candidates are all the elements in the cluster corresponding to the center label assigning to the node in the $R-1$ level of BP. Totally, BP is performed R times, and the result obtained by BP in level R is the final result. Since the higher the pyramid level is, the more blurring the elements are, BP in the lower level obtains more detailed results. Recall that the computational complexity of BP is proportional to the square of the number of label candidates. Therefore, our coarse-to-fine BP is at least $K^2/\sum_r K_r^2$ times faster than the standard BP.

For simplicity, we take a two-layer pyramid as an example to explain the scheme. Let K_1 and K_2 be the numbers of candidates in the first and the second BP executions respectively. We first use the k-means algorithm to classify all the patches in $\mathcal{L}$ into K_1 clusters, denoted as $S_1, S_2, ..., S_{K_1}$, i.e., $\mathcal{L} = \{S_1, S_2, ..., S_{K_1}\}$. The first running of BP takes the K_1 cluster centers as the label candidates $\mathcal{L}^1 = \{c_1, c_2, ..., c_{K_1}\}$ to find the best label configuration $X_1 = \{x_1^1, x_2^1, ..., x_N^1\}$ that minimizes the objective energy function, where $x_i^1 \in$

$\mathcal{L}^1$, $1 \leq i \leq N$. Then we perform BP again. Suppose that after the first BP, the best label for node v_i is $x_i^1 = c_{k_1}$. In the second round BP, the new label candidates for node v_i are all the elements[2] belonging to the cluster with center c_{k_1}. Using such different label candidate sets for different nodes, the second BP runs to find the best label configuration.

Obviously, such a coarse-to-fine BP scheme leads to a result different from that obtained with the standard BP. However, our experiments show that this scheme can achieve satisfactory results and is not sensitive to the initialization of the k-means algorithm. The most important benefit of this scheme is that it can make our algorithm practical. Such a coarse-to-fine BP can also be used to speed up some other MRF based applications in computer vision and computer graphics.

3.4 Experimental Results

In our experiments, we validate our algorithm on various videos representing different interesting and challenging cases to demonstrate its effectiveness. Here we show a few selected results from four representative videos, 120-frame "performance" (180 × 240) [75], 88-frame "beach" (80 × 170) [86], 40-frame "running" (240 × 320) [62], and 19-frame "car" (240 × 320) [62]. For all our experiments, the parameters in our algorithm are chosen as $\lambda_1 = 1$, $\lambda_2 = 1.5$, and $\alpha = 5$. The number of levels R in the multi-level BP is chosen as 2 or 3, depending on the size of a video.

Fig. 3.5 shows the visual comparison results for the video "performance" between the algorithm in [75] and our algorithm. The first row gives 4 original frames. We want to remove the walking spectator. The second row shows the manually removed regions roughly covering the spectator. The last two rows

[2]To limit the maximum label candidate number, if the number is larger than K_2, K_2 candidates are randomly selected.

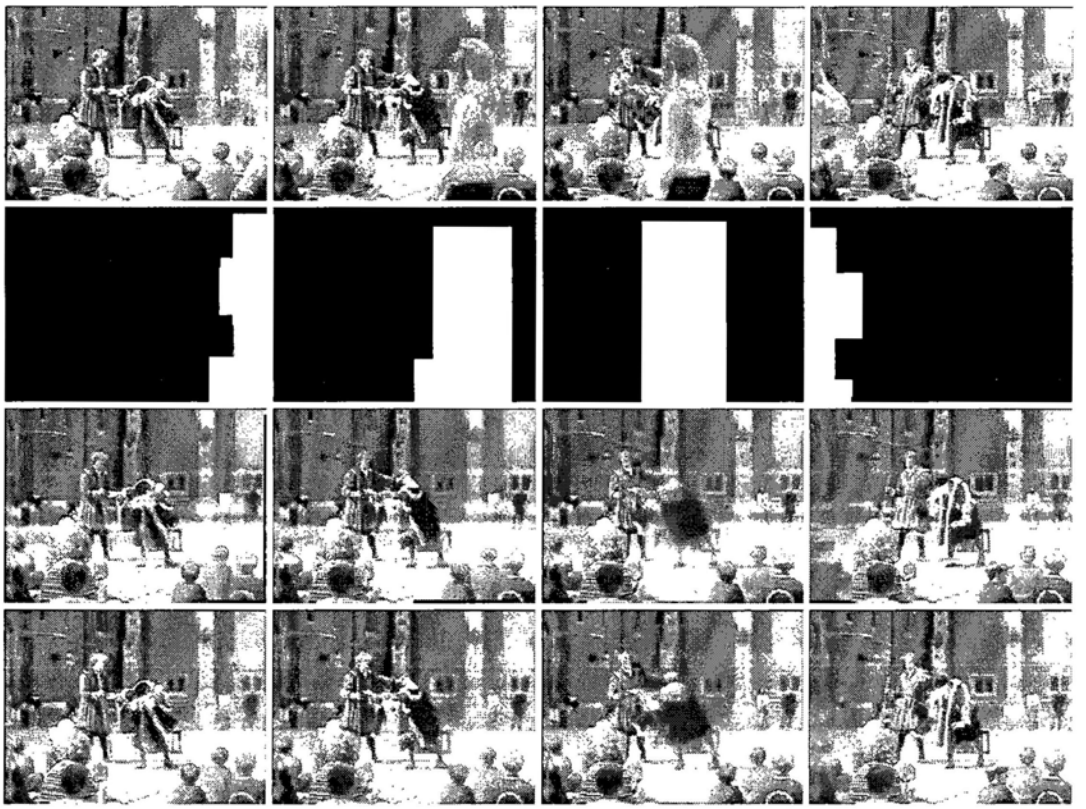

Figure 3.5: Some results on the "performance" video. The four rows show the original frames, the manually removed regions, the video completion results by [75], and the results by our algorithm, respectively.

display the completion results by [75] and our algorithm. Fig. 3.6 and Fig. 3.7 show the other results by our algorithm.

As shown in Fig. 3.5, the spectator takes a large space in each frame, and non-periodic motion happens in this video. The approach in [62], therefore, cannot handle this video completion well due to its periodic motion constraint and the large data missing. From Fig. 3.5, we can see that the algorithm [75] leads to serious blurring results for this video, as stated in [75], because of its simple weighted average scheme in color propagation. However, our algorithm generates promising results on this challenging case. In the "running" video, the camera taking the video is also moving. Our algorithm can fill in the holes well. Another challenging case in video completion is to complete the regions

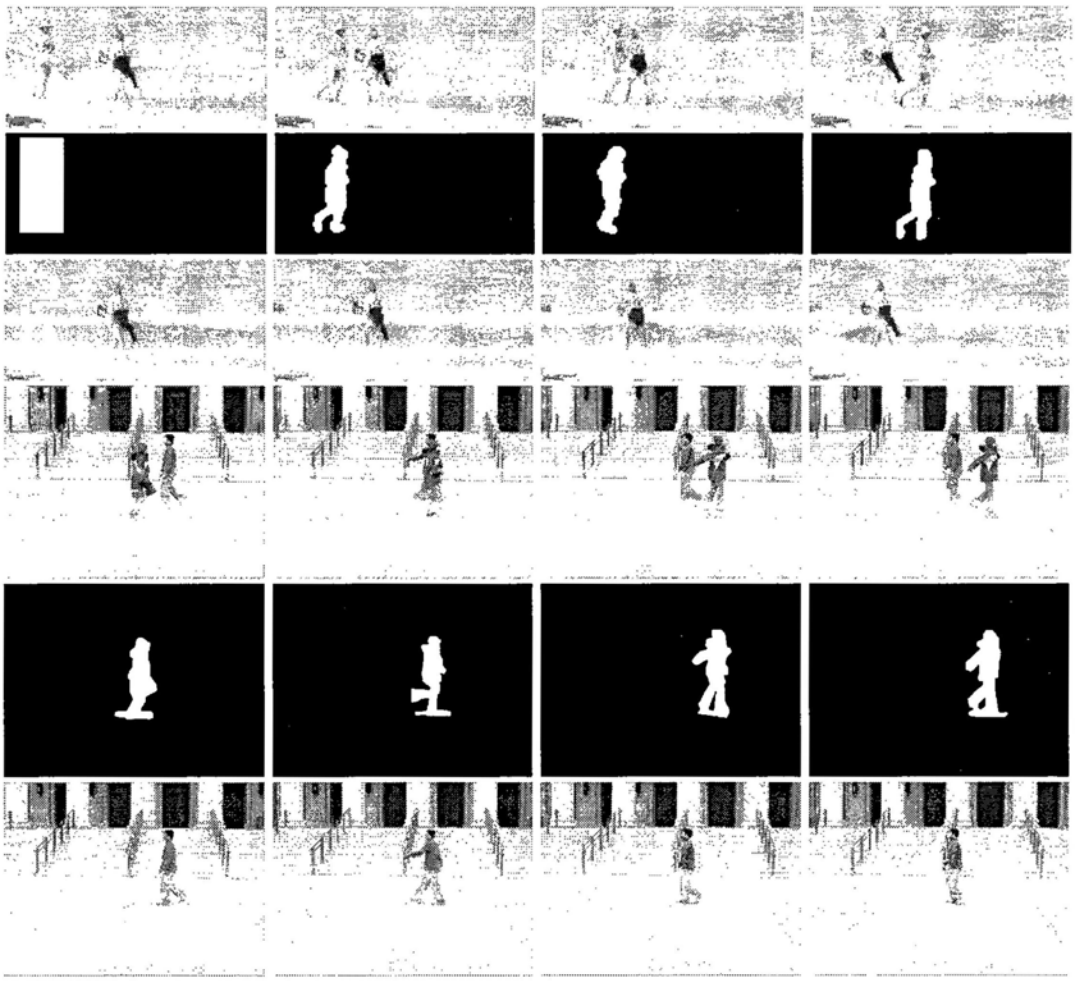

Figure 3.6: Some results on the "beach" and "running" videos. For each video, the original frames, the manually removed regions, and the video completion results by our algorithm are showed.

where the sizes of the objects change. Fig. 3.7 is such an example where the car moves closer to the camera. Our algorithm is still successful to complete the removed sign post.

From the experimental results, we can see that our algorithm can handle a variety of video completion tasks with different situations, such as dynamic foreground and background, camera motion, object scale changing, and large data missing. Besides, there is no periodic motion restriction imposed on our algorithm.

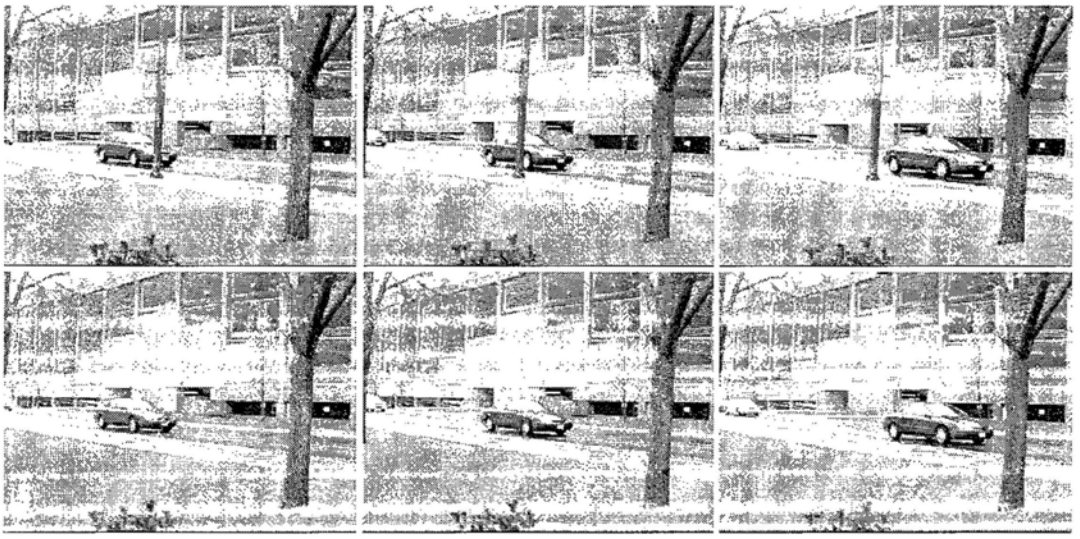

Figure 3.7: Some results on the "car" video.

As we discussed, the proposed framework can handle the problem of image completion by removing the temporal consistency constraint. We validate our framework on many natural images and make comparisons with related algorithms [19, 43]. Fig. 3.8 gives the comparative results. We can see that the results generated by [19] have obvious artifacts due to its greedy scheme, [43] leads to the results losing some strong structures, and our algorithm produces visually natural results with good texture and structure visual consistency. Fig. 3.9 shows more results by our algorithm, which illustrate the excellent performance of the proposed framework on image completion.

3.5 Conclusions

In this chapter, a novel video completion algorithm has been proposed by combining motion completion and global exemplar-based color completion. For a video with holes, the motion field in the holes is filled locally first. Based on the completed motion field, color is restored in a global exemplar-based scheme by minimizing an MRF energy function. The color completion is a patch copy-and-paste process, i.e., copying patches in the source region and pasting them

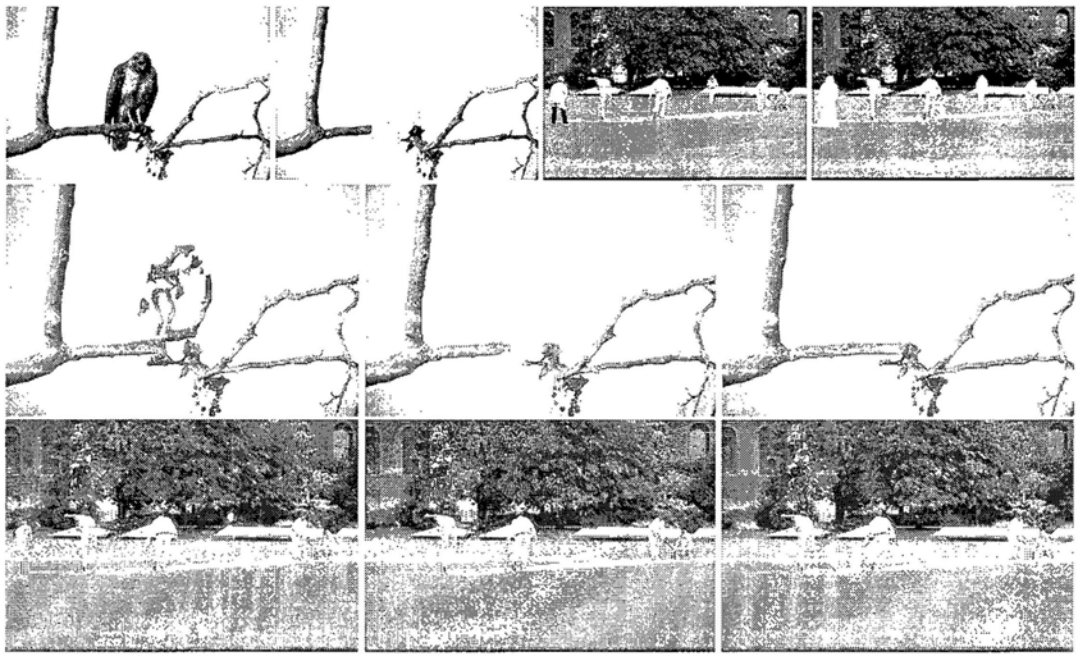

Figure 3.8: Comparative results of image completion. The first row contains two pairs of original and masked images. On the second and the third rows, from left to right: the results obtained by [19], [43], and our algorithm.

to the data missing region. The proposed objective function enforces both spatial and temporal consistency constraints in the color completion process. Besides, our framework can also well handle the problem of image completion as a special case of video completion when there are no temporal information and constraints (set $\alpha = 0$ in our energy function). Belief propagation is used to solve the energy minimization problem. To avoid the computational impracticability caused by the large number of label candidates in BP optimization process, we utilize a coarse-to-fine optimization scheme whose essential idea is to carry out BP multiple times with sharply reduced number of label candidates, instead of running BP only one time with a large number of label candidates. Such a coarse-to-fine scheme makes BP practicable in our algorithm. The experimental results on a variety of videos and images have demonstrated the effectiveness of our proposed uniform framework for image and video completions.

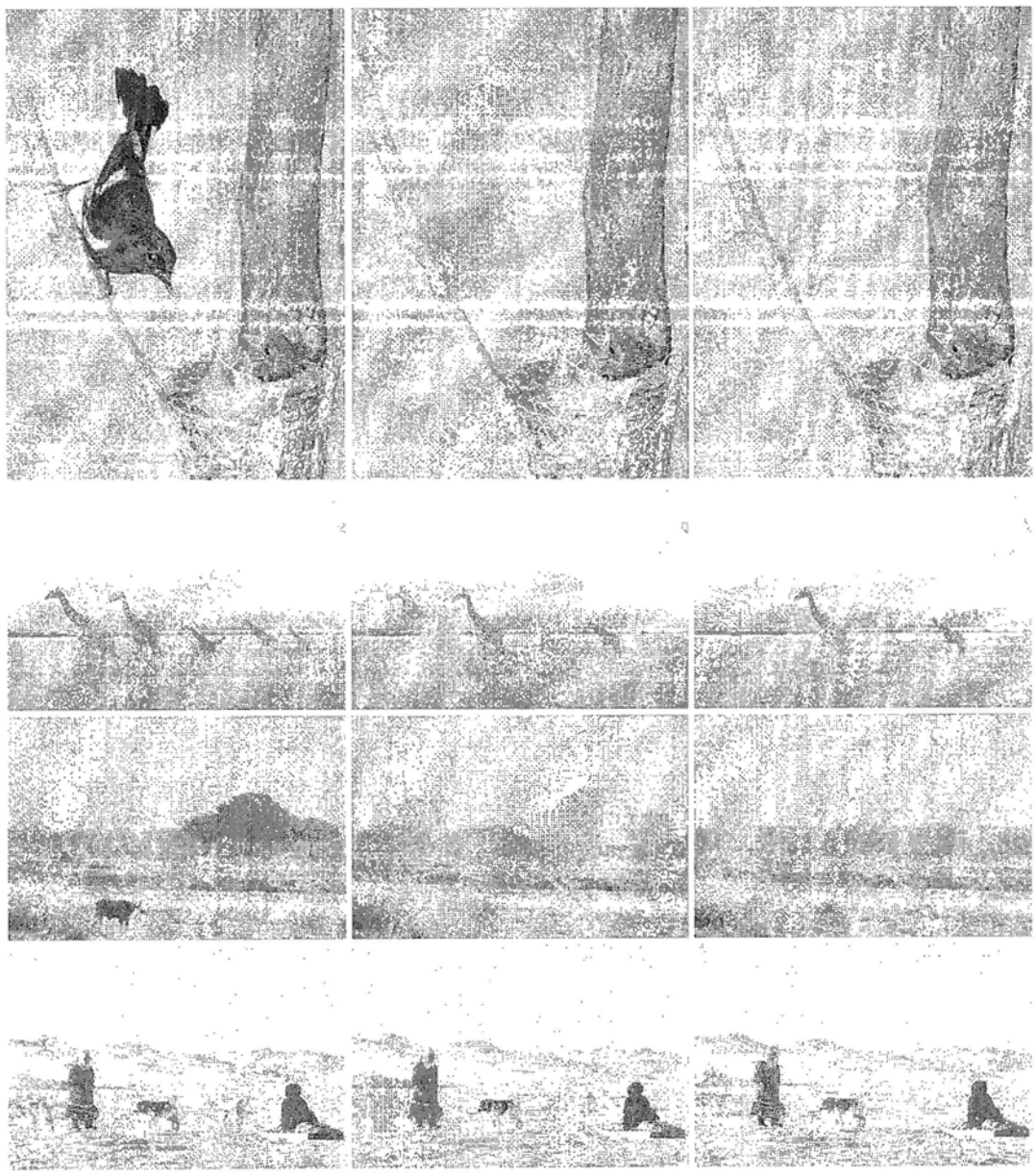

Figure 3.9: Image completion results by our algorithm.

Chapter 4

Continuous MRF Based Image Denoising

In this chapter, we tackle the problem of image denoising by formulating it as maximum a posterior (MAP) estimation problem using Markov random fields (MRFs). The estimation, which has been demonstrated to well model the problem, is equivalent to a maximum likelihood estimation constrained on spatial homogeneity and is generally NP-hard in discrete domain as our discussion in Chapter 1. To make it tractable, we convert it to a continuous label assignment problem based on a Gaussian MRF model and obtain a closed form global optimal solution, which is similar to our derivation in Chapter 2 for interactive foreground object extraction. Since the Gaussian MRFs tend to over-smooth images and blur edges, we incorporate pre-estimated edge information into the energy function to better preserve image structures. Patch similarity based pairwise interaction is also involved to better preserve image details and make the algorithm more robust to impulse noise. Both quantitative and qualitative comparative experimental results are given to demonstrate the better performance of our algorithm over several state-of-the-art related algorithms.

4.1 Introduction

Due to the imperfection of image acquisition and transmission systems, images are often corrupted by noise. The contamination on images not only affects their visual quality but also precludes many further higher-level computer vision tasks such as image/video coding, recognition, scene understanding, and object tracking. Therefore, either as a stand-alone processing or as a pre-processing, it remains one of the most active topics in image processing. In most cases, a noisy image can be modeled as $X = F + N$, where X, F, and N represent the observed noisy image, the noise-free image, and the noise that is often considered as Gaussian with zero mean. The goal of image denoising is to remove the noise while maintaining and recovering the details and structures of the image as much as possible.

Many denoising approaches have been developed over the past decades. They can be grouped into two basic categories: filtering in the spatial domain and filtering in the frequency domain. The former in essence estimates the value of each pixel with its neighboring pixels in some way. The basic idea of the latter is to project an image onto a set of orthogonal bases, usually referred to as wavelet bases, and then to discard small coefficients (mainly representing the noise) in the transformed representation using some kind of thresholding or by shrinking [20]. In [14], the authors consider the evaluation of a good denoising algorithm in three aspects: no structure loss, no artifact generation, and optimal neighborhood selection. Their analysis indicates that algorithms in both the spatial and frequency domains are of great importance for image denoising. Some good algorithms in the frequency domain hold the property of no structure loss in denoised images but generate artifacts. On the other hand, some classic spatial filters are artifact free but destroy useful structures of original images. Since our algorithm is developed in the spatial domain, we focus on the discussion and comparison of the spatial filtering methods in this

chapter.

Traditional image denoising methods in spatial domain use linear local smoothing filters to do the work. The most common and simplest one is the *Gaussian filter* [27] that has the advantage of fast computation. However, since these linear filters are all based on the assumption of stationarity of the whole image, which is not true in common real-world images, they are incapable of preserving edges and details well. Nonlinear models, on the other hand, can preserve edges better and reduce the blurring effect. Many nonlinear filters are based on partial differential equations (PDEs). *Total Variation (TV) filter* [70] solving a 2nd-order nonlinear PDE suffers the staircase effect and the loss of image texture information, although it can keep edges well. To avoid this effect, a 4th-order PDE filter combined with the TV filter is presented in [54] and some algorithms based on iterated total variation [81], [61] are also proposed. Another popular model of PDEs based approaches is the *anisotropic filter* [65], [15] which uses an anisotropic diffusion (AD) equation to smooth a noisy image. While maintaining boundaries well, the AD removes small details and fine structures of the image.

In order to use the grey level information of neighborhoods in the local smoothing process, the sigma-filter is developed by Lee in [44], whose idea is to average neighboring pixels with similar grey levels to the reference pixel's. Two popular algorithms, called *SUSAN filter* [76] and *bilateral filter* [84], take the average value of the pixels close to the reference pixel in both grey level and spatial location, while other local filters only consider the geometric closeness of pixels. The bilateral filter can keep relatively sharp image edges and maintain the structures well. However, relying on the grey levels between two single pixels is not robust for the denoising of a seriously noisy image. Recently, considerable interests have been given to the use of image partition for denoising [32], [56], [57]. These approaches share the same idea: smoothing the reference pixel or region by using the ones belonging to the same cluster

obtained through a segmentation procedure to maintain sharp edges.

Buades et al. present a non-local means (NL-means) algorithm for image denoising [14], [13]. They argue that the local smoothing methods aim at noise reduction and the reconstruction of the main geometrical configurations of the image, but not at the preservation of fine details and textures. To address this problem, the NL-means approach estimates the "true" value for a pixel as the weighted average grey level of all pixels whose Gaussian neighbors look like the neighbors of the reference pixel with a close neighborhood configuration. This approach is based on the assumption that neighborhood similarities of each pixel exist in nature images. It is suitable for denoising images with periodicity texture patterns, but it fails in images with strong noise due to the corruption of the image structures. Besides, the nature of the simply weighted average calculation may cause grey-level inconsistency in some regions.

Recall our introduction to the Markov random field (MRF) models [25] in Chapter 1. The MRF models can handle the problem of image denoising through an MRF-based energy minimization problem. The formulation is justified in terms of maximum a posteriori (MAP) estimation of a Markov random field in the Bayesian framework. However, as a multi-label assignment problem, image denoising modeled as the MRF energy minimization in the discrete domain is generally NP-hard. The major obstacle of the optimization is the large computational cost owing to the high dimensional computing space. The approaches proposed to solve the MRF energy optimization have been described in Chapter 1, including simulated annealing [25] being very time consuming, the iterated conditional modes (ICM) [4] with a deterministic greedy strategy to obtain a local minimum, and recently developed algorithms based on graph cuts [11, 12, 42, 10], belief propagation [85, 82, 80], and tree-reweighted message passing (TRW-S) [40]. Graph cuts, which minimizes the energy function by using the min-cut/max-flow on a properly constructed graph, has demonstrated its good performances to handle MRF optimization

problems, although it converges to a local optimum in the context of image denoising. Belief propagation and TRW-S as message passing algorithms often performs as well as graph cuts. All these algorithms work in the discrete domain and usually can only find a local optimum. Moreover, they carry out the optimization iteratively without a closed form solution. We also find that these methods cannot well preserve image edges and the large labels in the denoising task, especially for color images, cause the optimization procedure to be very slow.

In this chapter, we also formulate image denoising as an MRF energy minimization problem with elaborately defined pairwise relationship between neighboring pixels. Our optimization approach is based on label relaxation. We solve the label estimation by transforming it to a continuous optimization problem, where the labels of the pixels are relaxed from discrete values to continuous values. Compared with the related approaches, the contributions of our work are summarized as follows: 1) In the continuous domain, a closed form global optimal solution can be obtained, which provides a good prerequisite for our final result. 2) Image edges and details can be better preserved in our algorithm since pre-estimated edge information and patch based similarity are incorporated into the design for the MRF energy function. 3) While obtaining better or comparable results, our algorithm is more efficient than belief propagation, graph cuts, and NL-means. 4) Our formulation for gray level image denoising can be directly extended to the denoising of color images with the CIE-Lab color space used without increasing the computational complexity. Our experimental results have demonstrated these advantages of our approach and shown that it outperforms several state-of-the-art related methods both quantitatively and qualitatively.

4.2 The Basic MAP-MRF Model

From Chapter 1 we know that the formulation of the MRF model is justified in terms of the MAP estimation of the MRFs in the Bayesian framework. Next we give the specific explanation of the MAP justification for MRF energy optimization in the context of image denoising.

Let an input image and its labelling be represented by $X = [x_1, x_2, ..., x_n]^T$ and $F = [f_1, f_2, ..., f_n]^T$, respectively, where x_i is the intensity of pixel i and f_i is its corresponding label (restored intensity), $1 \leq i \leq n$, and n is the number of the pixels. F is the restored image denoting a realization of the MRF. From probabilistic perspective, image denoising can be regarded as an optimization problem that is to maximize the posterior probability $P(F|X)$ such that the global optimal solution can be found.

We know that the MAP estimation $\widehat{F}$ such that $\widehat{F} = \text{argmax}_F\, P(X|F)P(F)$ is equivalent to minimizing the following energy function:

$$E(F) = \sum_{i=1}^{n} D_i(x_i, f_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} S_{i,j}(f_i, f_j), \tag{4.1}$$

where the data penalty function $D_i(x_i, f_i)$ that penalizes the inconsistency between the labels and the data comes from the likelihood function $P(X|F)$, and $S_{i,j}(f_i, f_j)$ as the clique potential of MRFs representing the prior knowledge of the labels is from the prior probability $P(F)$. $\mathcal{N}(i)$ is the neighborhood of i. Here it is straightforward to explain our claim in Chapter 1 that $P(X|F)$ can be represented by the sensor noise model [11], since the inconsistency between the data and labels in image denoising is just the noise model we assume.

Based on the energy function (4.1), it is obvious that the MAP estimation $\widehat{F}$, which maximizes $P(X|F)P(F)$ or equivalently minimizes $E(F)$, tends to be a balanced label configuration (restored image) that is consistent with the data X (observed image) as well as following the smoothness constraint.

4.3 Continuous MRF Based Image Denoising

In our work, we focus on grey-level images, but the formulation can be directly extended to handling color images.

The data term $D_i(x_i, f_i)$ in our approach is chosen as

$$D_i(x_i, f_i) = (f_i - x_i)^2, \tag{4.2}$$

which models the additive Gaussian noise and is commonly used in image denoising. The clique potential is defined as

$$S_{i,j}(f_i, f_j) = w_{ij}(f_i - f_j)^2, \tag{4.3}$$

where w_{ij} denotes the affinity value between pixels i and j and is used to control the smoothness degree for each pairwise interaction. The quadratic label difference without truncation is not edge preserving. However, with the special design of the affinity value w_{ij}, where pre-estimated edge information is incorporated into its calculation, we can maintain edge sharpness in the denoised image. Furthermore, patch based similarity measurement is also used in the design of w_{ij}, which further preserves image details as much as possible.

Before giving w_{ij} explicitly, we first define the patch based similarity between pixels i and j as

$$\Delta(i, j) = ||\mathbf{x}_{\mathcal{B}(i)} - \mathbf{x}_{\mathcal{B}(j)}||^2 / |\mathcal{B}(i)|, \tag{4.4}$$

where $\mathbf{x}_{\mathcal{B}(i)}$ and $\mathbf{x}_{\mathcal{B}(j)}$ represent the grey-level vectors of the pixels in two same-size blocks $\mathcal{B}(i)$ and $\mathcal{B}(j)$ centered at pixels i and j, respectively. $|\mathcal{B}(i)|$ is the cardinality of $\mathcal{B}(i)$.

We design w_{ij} in this way: 1) if the difference between the blocks centered at neighboring pixels i and j is large in the input image, the smoothness penalty $S_{i,j}(f_i, f_j)$ should be small; 2) the farther the distance between pixels i and j, the less effect of them on $S_{i,j}(f_i, f_j)$; 3) if pixels i and j fall into two

Figure 4.1: Explanation of the region indexes. The curve denotes an edge separating the window into two regions. Pixels i and j_1 have the same region index ($C_i = C_{j_1}$), but pixels i and j_2 have different region indexes ($C_i \neq C_{j_2}$).

regions separated by an edge, they have no effect on $S_{i,j}(f_i, f_j)$. Based on these criteria, we define w_{ij} as

$$w_{ij} = a \cdot \exp(-\frac{\Delta(i,j)}{b}) \cdot k(i,j) \cdot T(C_i = C_j), \tag{4.5}$$

where a and b are two positive factors to control the contribution of w_{ij} to the smoothness penalty, $k(i,j) = \exp(-d_{ij}^2/2)$ is a Gaussian kernel function to reach target 2) above, and $T(C_i = C_j)$ is used towards target 3). $T(\cdot)$ is 1 if its argument is true and 0 otherwise. C_i and C_j are region indexes that can be explained with Fig. 4.1, in which an edge separates the window into two regions and pixels i and j_1 have the same region index but pixels i and j_2 have different region indexes, i.e., $C_i = C_{j_1}$ and $C_i \neq C_{j_2}$. We use Canny edge detector to find edges in the input image since this edge detector is not much sensitive to noise, and then assign indexes to different regions. The design of w_{ij} in (4.5) makes our algorithm not only be able to denoise but also preserve edges and details well.

With the data penalty (4.2) and the smoothness penalty (4.3), the energy function (4.1) can be written as

$$E(F) = \sum_{i=1}^{n}(f_i - x_i)^2 + \sum_{i=1}^{n}\sum_{j \in \mathcal{N}(i)} w_{ij}(f_i - f_j)^2. \tag{4.6}$$

This objective function represents a Gaussian MRF. The approaches in [12, 42, 85, 82, 23] can solve the energy minimization problem by graph cuts or belief propagation. However, solutions obtained by them are locally optimal

in the discrete domain. On the other hand, the algorithm proposed in [36] can exactly optimize the energy function (4.6) in the discrete domain by converting the problem into a min-cut/max-flow [10] problem with a complicated directed graph. However, this algorithm is not suitable for image denoising due to the heavy computational burden caused by the large sets of the nodes and edges of the graph.

Next we give our closed form global optimal solution to this optimization problem. It is based on the relaxation of the labels from discrete values to continuous values and the utilization of the normalized Laplacian matrix corresponding to an undirected weighted graph.

With the smoothness term in (4.6), the MRF is isotropic and we can construct an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices denoting the image pixels and $\mathcal{E}$ is the set of weighted edges. Then the adjacency matrix of $\mathcal{G}$, which has the similar definition to the one proposed in Chapter 2, is $W = [W_{ij}]_{n\times n}$ whose elements are defined as

$$W_{ij} = \begin{cases} w_{ij}, & \text{if } i \neq j, j \in \mathcal{N}(i) \\ 0, & \text{if } i \neq j, j \notin \mathcal{N}(i) \\ c, & \text{if } i = j, \end{cases} \tag{4.7}$$

where $c > 0$ is some constant. It is clear that c has no effect on the energy function $E(F)$ since the term $c(f_i - f_i)^2 = 0$. Let D be an $n \times n$ diagonal matrix with the (i,i)-th entry $D_{ii} = \sum_{j=1}^{n} W_{ij}$. When $D_{ii} = 0$, we call i an isolated vertex in the graph that causes W to be singular. By using the positive constant c, $D_{ii} \neq 0$ and singular W is avoided. Moreover, c builds up numerical stability for our solution and is better to be comparable with the value of the parameter a as we discussed in Chapter 2.

Based on the construction of the energy function $E(F)$ and the corresponding graph $\mathcal{G}$, similar to derivation process in Chapter 2, a closed form global optimal solution in the continuous domain to minimize $E(F)$ can be obtained

as

$$F = D^{-\frac{1}{2}}(D^{-1} + 2\overline{L})^{-1}D^{-\frac{1}{2}}X. \tag{4.8}$$

where $\overline{L}$ is the normalized Laplacian matrix of $\mathcal{G}$ defined as $\overline{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$.

From our formulation and the derived solution, it is not difficult to see that our approach is able to handle color image denoising directly. Given a color image, we treat the color of each pixel as a three dimensional vector. In the CIE-Lab color space, the three channels are relatively independent. Based on the perceptual linearity of CIE-Lab, we use the Euclidean distance in this space as the color difference measurement involved in (4.5), which is consistent with human perception. With the replacement of the gray-level values by the color vectors, we achieve the same closed form solution for color image denoising as in (4.8).

It is worth noting that the above derivation of the global optimal F is based on the assumption that F is a continuous vector, which is preferred if the denoised image is used for further higher-level processing. However, if we want to display the denoised image, the pixel intensities (labels) have to be discrete. After obtaining the optimal continuous solution, we quantize it to obtain its discrete version, which naturally results in the deviation of denoising output from the global optimal solution. Fortunately, we find that discrete result slightly falls away and the error is even within a known bound. The optimal property will be elaborated in the next section.

4.4 Optimal Property

Optimal property discussion is meaningful for MRF energy optimization techniques, especially for those converging to a local minimum according to some

criteria, e.g., iterated conditional modes (ICM) [4], graph cuts based expansion algorithm and swap algorithm [12] and belief propagation [85],[23]. To the best of our knowledge, except for the expansion algorithm by which the optimized energy is within a known factor $2c$ of the global optimum, where $c = \max_{i,j\in\mathcal{N}}(\frac{\max_{f_i\neq f_j} s(f_i,f_j)}{\min_{f_i\neq f_j} s(f_i,f_j)})$, other algorithms converging to the local minimum do not have theoretical analysis for their optimal properties, although it is undeniable that many of them have desirable visually outputs.

To present output image, discretization is carried out on our global optimal solution in continuous domain by setting its components to their closest discrete values. It is easy to see that the the error caused by dicretization of intensity values affects the output energy in (4.6). As an optimization technique for energy minimization, we should consider the worst case to give an insight into the optimal property of our approach. Since the error between global optimal energy and the one after discretization in our algorithm can be proved within a known bound, our algorithm has the guaranteed optimality property.

Before giving the energy error bound, we first introduce some notations. Let f_i^* and f_i denote the optimal continuous label and corresponding discretized one for pixel i, $1 \leq i \leq n$, ranging from 0 to 1. E_{opt}, E_{dis} represent the global optimal energy in the continuous domain and the energy after label discretization, respectively. We have $E_{opt} = E_{data} + E_{smooth}$, where $E_{data} = \sum_{i=1}^{n}(f_i^* - x_i)^2$ and $E_{smooth} = \sum_{i=1}^{n}\sum_{j\in\mathcal{N}(i)} w_{ij}\cdot(f_i^* - f_j^*)^2$. With these notations, the following theorem gives the energy error bound between E_{opt} and E_{dis}.

Theorem 1. *The error between E_{opt} and E_{dis}, represented by $\Delta E = E_{dis} - E_{opt}$, holds the upper bound $2^{\frac{1}{2}(\lceil log_2^{4(1+k)n}\rceil)}\sqrt{cE_{opt}} + A$, that is,*

$$\Delta E \leq 2^{\frac{1}{2}(\lceil log_2^{4(1+k)n}\rceil)}\sqrt{cE_{opt}} + A, \tag{4.9}$$

where $A \sim \mathcal{O}(n)$, is some constant, k is the number of neighborhood for each

pixel and $c = \max(\max_{i,j}(w_{ij}), 0.5)$, *the operation* $\ulcorner * \urcorner$ *denotes taking upper integer.*

Proof. With the definitions of f_i^*, f_i, and the energy function (4.6), we have

$$\begin{aligned}\Delta E =& \sum_{i=1}^{n} \big[(f_i - x_i)^2 - (f_i^* - x_i)^2 + \\ & \sum_{j\in\mathcal{N}(i)} w_{ij} \cdot ((f_i - f_j)^2 - (f_i^* - f_j^*)^2)\big] \\ \leq & 2\sum_{i=1}^{n} \big(|f_i^* - x_i||d_i| + \sum_{j\in\mathcal{N}(i)} w_{ij} \cdot |f_i^* - f_j^*||d_i + d_j|\big) \\ & + \sum_{i=1}^{n} \big(d_i^2 + \sum_{j\in\mathcal{N}(i)} w_{ij} \cdot (d_i + d_j)^2\big), \end{aligned} \tag{4.10}$$

where $d_i = f_i - f_i^*$, $1 \leq i \leq n$. It is obvious that d_i is in $(-0.5, 0.5]$ due to the rounding operation for discretization, and therefore $d_i + d_j \in (-1, 1]$.

Considering the worst case that $|d_i|$ and $|d_i + d_j|$ take their maximal values (say, 0.5 and 1) respectively, which in general can not be achieved simultaneously in all i, $1 \leq i \leq n$ and (i, j), $(i, j) \in \mathcal{N}$, by substituting $|d_i| = 0.5$ and $|d_i + d_j| = 1$ into (4.10) we have

$$\Delta E \leq \sum_{i=1}^{n} \big(|f_i^* - x_i| + \sum_{j\in\mathcal{N}(i)} 2w_{ij} \cdot |f_i^* - f_j^*|\big) + A, \tag{4.11}$$

where $A = \frac{n}{4} + \sum_{i=1}^{n} \sum_{j\in\mathcal{N}(i)} w_{ij}$, is a constant and $A \sim \mathcal{O}(n)$.

Based on the inequation $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ where $a, b \geq 0$, (4.11) can be further advanced as

$$\begin{aligned}\Delta E &\leq 2c \sum_{i=1}^{n} \big(\frac{1}{\sqrt{2c}}|f_i^* - x_i| + \sum_{j\in\mathcal{N}(i)} \sqrt{\frac{w_{ij}}{c}}|f_i^* - f_j^*|\big) + A \\ &\leq 2c(\sqrt{2})^{(\ulcorner log_2^{(1+k)n} \urcorner} \sqrt{\frac{1}{2c}E_{data} + \frac{1}{c}E_{smooth}} + A \\ &\leq 2^{\frac{1}{2}(\ulcorner log_2^{4(1+k)n} \urcorner)} \sqrt{cE_{opt}} + A, \end{aligned} \tag{4.12}$$

where $c = \max(\max_{i,j}(w_{ij}), 0.5)$ □

With Theorem 1, we have the upper bound for the energy deviation ΔE caused by the label discretization, which demonstrates the guaranteed optimal property of the proposed algorithm. Moreover, our experiments show that the energy of the discrete solution is very close to the global optimal energy.

4.5 Experimental Results

To demonstrate the performance of our algorithm, we compare our algorithm with five most related algorithms: swap graph cuts (GC) [12], max-product belief propagation (BP) [80], Gaussian filter (GF) [27], bilateral filter (BF) [84], and NL-means (NL) [14]. GC [12] and BP [80] are representative MRF-based approaches, GF [27] is a linear local smoothing filter, BF [84] is a nonlinear local smoothing filter, and NL [14] is a nonlocal method. We test these algorithms on a set of classic grey level images, "Barbara", "Boat", "House", "Pepper", and "Lena" of size 256×256, and all 300 natural images in the Berkeley segmentation benchmark [55]. These images are contaminated by adding Gaussian noise of five levels with standard deviations $\sigma = 10$, 20, 30, 50 and 100. Both visual quality comparisons and quantitative comparisons are given. Peak signal-to-noise ratio (PSNR) is used for the quantitative evaluations. The experiments are divided into two parts: comparisons with the two MRF-model based algorithms and comparisons with the other three algorithms.

4.5.1 Comparisons with the MRF Based Algorithms

The max-product belief propagation and the swap graph cuts are the most popular algorithms for MRF optimization and have been used successfully in many applications including image denoising. The comparisons with them can better demonstrate the superiority and efficiency of our algorithm. Furthermore, the obtained energy values by these algorithms are also provided.

Different energy functions generate different results. Our energy function is constructed by the quadratic data and smoothness terms with spatially varying w_{ij} to force labelling discontinuity at edges and labelling smoothness in homogenous regions. The energy functions of BP and GC are given in [80] and [12], respectively. The parameters of the three algorithms are tuned best in terms of the largest PSNR at each noise level. In BP and GC, the data truncation constant, the smoothness truncation constant, and the affinity value λ range from 3000 to 10000, 200 to 500, and 2 to 10, respectively. We cannot find better results of BP and GC when these parameters are not in these intervals. The iteration numbers in BP and GC are set to 15 and 20, respectively, which are large enough to ensure the convergence of them. In our algorithm, the parameter a ranges from 1 to 3 and the parameter b is fixed to 100. We find that the best parameter setting in our algorithm is easier to obtain and is more stable than those in BP and GC.

The PSNR comparisons of the denoising results on the five commonly-used testing images at the five noise levels are given in Table 4.1. From the quantitative comparison, we can see that our algorithm outperforms BP and GC. Among all the PSNR values, our algorithm obtains the better results than BP and GC.

Some denoised images "Barbara" and "Boat" are showed in Fig. 4.2 and Fig. 4.3. The presented visual results correspond to the largest PSNR outputs generated by the three algorithms. These visual comparisons indicate that our algorithm performs much better than BP and GC. It well preserves the edges although some small details are lost. In contrast, BP and GC remove the details and main structures and over-smooth the images. There are also some isolated undesirable noise pixels in BP's and GC's results that badly affect the visual quality.

As energy minimization techniques, BP and GC have been proven to be powerful energy minimization techniques converging to strong local optima.

image	σ	PSNR					
		GC	BP	GF	BF	NL	Ours
Barbara	10	29.79	29.76	27.82	30.15	**31.67**	31.21
	20	24.62	24.73	23.89	25.72	**28.40**	28.32
	30	22.15	22.55	22.43	23.82	25.60	**26.15**
	50	20.03	20.84	21.28	21.26	20.92	**21.32**
	100	17.72	18.08	19.04	19.25	14.58	**19.46**
Boat	10	30.55	30.55	29.69	30.42	30.29	**30.79**
	20	26.43	26.60	25.78	26.65	**27.27**	27.12
	30	24.24	24.55	23.90	24.81	25.71	**25.92**
	50	20.85	22.11	22.23	22.34	22.28	**22.54**
	100	18.15	18.70	19.81	19.91	14.54	**20.06**
Pepper	10	28.77	27.29	30.48	32.79	33.34	**33.68**
	20	27.03	25.04	27.05	28.91	30.15	**30.42**
	30	24.64	22.35	25.27	26.42	27.79	**28.43**
	50	22.31	21.02	23.34	23.45	23.89	**24.92**
	100	15.93	13.58	20.10	20.31	14.78	**20.82**
House	10	30.67	30.72	31.17	33.70	35.36	**35.52**
	20	28.98	28.43	28.04	30.05	32.25	**32.34**
	30	27.29	26.21	26.43	27.85	29.21	**30.82**
	50	25.09	23.43	24.42	25.85	25.03	**26.43**
	100	18.71	15.19	21.32	22.27	14.81	**22.54**
Lena	10	31.65	31.74	30.53	32.00	**32.61**	32.35
	20	27.56	27.79	27.13	28.35	**29.53**	29.22
	30	25.10	25.80	25.44	26.36	27.28	**27.51**
	50	22.89	23.56	23.62	24.24	23.88	**25.24**
	100	18.95	19.75	20.65	21.14	14.70	**21.37**

Table 4.1: PSNR values obtained by GC, BP, GF, BF, NL, and our algorithm on the five noisy images at five noise levels.

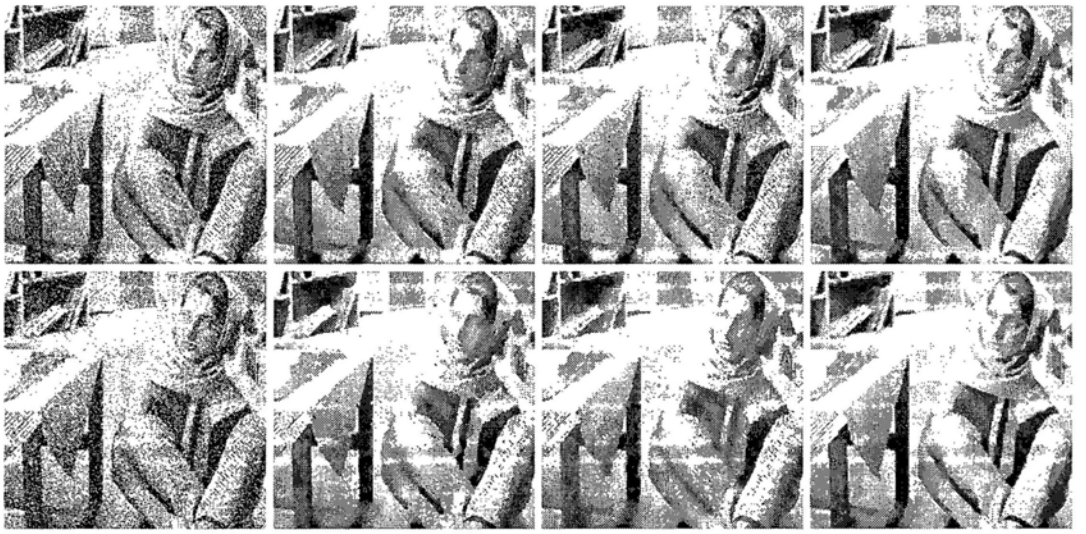

Figure 4.2: Results of the MRF-based denoising algorithms on the "Barbara" image with the noise $\sigma = 20$ (the first row) and $\sigma = 30$ (the second row). From left to right: the noisy image, the results of BP, GC, and our algorithm.

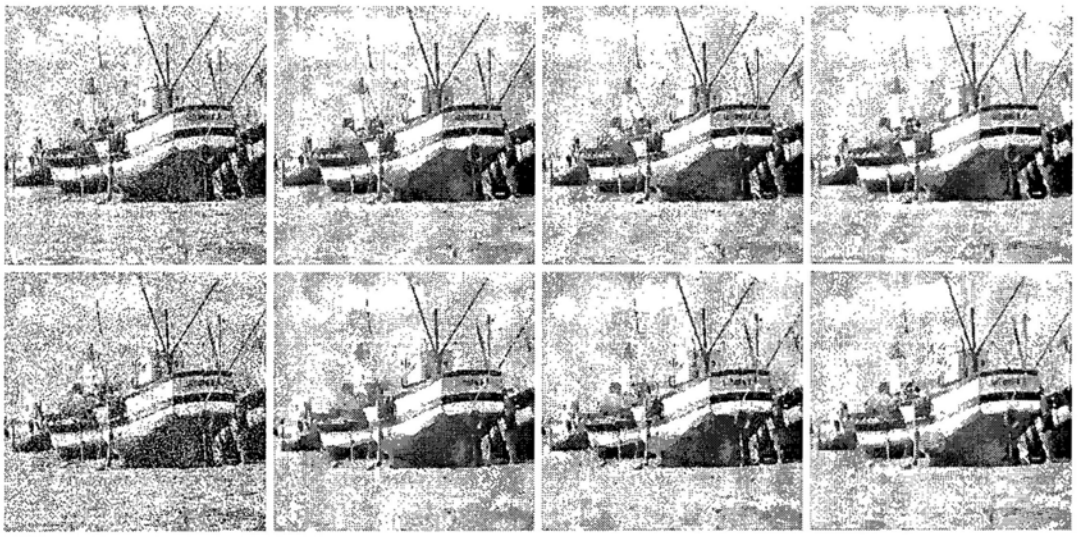

Figure 4.3: Results of the MRF-based denoising algorithms on the "Boat" image with the noise $\sigma = 20$ (the first row) and $\sigma = 30$ (the second row). From left to right: the noisy image, the results of BP, GC, and our algorithm.

image	E_{opt}	E_{ours}	E_{BP}	E_{GC}
Barbara	2.4309	2.4353	2.4863	2.5863
Boat	2.4086	2.4133	2.4353	2.5483
House	0.9128	0.9205	0.9295	1.0181
Pepper	1.8740	1.8798	1.9168	2.0523
Lena	1.3662	1.3730	1.4240	1.5961

Table 4.2: Comparison of the energy values ($\times 10^7$) obtained by the three algorithms on the five noisy images with $\sigma = 20$.

On the other hand, the deviation from the global optimum in the continuous domain to the discrete output in our algorithm is unavoidable. Therefore, the discussion of the optimal property in terms of output energy deviation from global optimum is important. We know that the GC expansion algorithm produces a solution within a known factor $2c$ of the global optimum, where $c = \max_{i,j\in\mathcal{N}}(\frac{\max_{f_i\neq f_j} s(f_i,f_j)}{\min_{f_i\neq f_j} s(f_i,f_j)})$. The factor can be as small as 2 in the best situation where smooth term is Potts model, but Potts model does not suit for denoising task since as the piecewise constant model, Potts model denoising generates serious staircase effect. Under other smoothness constraint model applicable for expansion algorithm like linear truncated one, the optimal factor is always large in image denoising, which leads to the insignificant optimal guarantee. With our energy and optimization strategy that is effective for image denoising, the optimal guarantee is attainable with Theorem. 1. However, the error bound depends on the continuous global optimal energy E_{opt} found by our algorithm in the continuous domain. To demonstrate the optimal property in practice, we calculate the ratio parameter (factor) of our output energy with respect to specific E_{opt} on five noisy images with $\sigma = 20$. The factors are all around 1.5, which validates the optimal property of our algorithm.

Moreover, we compare the output energy values obtained by BP, GC and our algorithm using the same energy function as in (4.6). Table 4.2 shows the results where E_{opt} are also given. From the table, we can see that the output energy values (E_{ours}) of our algorithm are closer to the optimal values than

E_{BP} and E_{GC}, which are obtained by BP and GC, respectively. The results in Table 4.1 and Table 4.2 are consistent, showing that our algorithm works best both in terms of denoising outputs and energy outputs.

4.5.2 Comparisons with the Other Three Algorithms

In this section, we compare our algorithm with the other three state-of-the-art spatial denoising approaches. Similarly, we choose the best parameters for the algorithms in terms of the best PSNR outputs at each noise level. The window size and the standard deviation in GF are from 7 to 19 and 0.5 to 2, respectively. The window size of BF is between 7 and 11 with the standard deviations of the spatial domain and the intensity domain ranging from 1 to 2 and 0.1 to 2, respectively. The search window size in NL is 15 to 21 and the similarity measurement window is 7×7 with the filter degree between 100 and 1000. The parameter setting of our algorithm here is the same as the one in the experiments described in Section 4.5.1.

The PSNR values are given in Table 4.1. From these results, we can see that NL and our algorithm almost always outperform the other two. At less noise levels with $\sigma = 10$ and 20, our algorithm obtains comparable PSNR values to those by NL. For seriously noisy images ($\sigma = 30$, 50, 100), our algorithm achieves the best results. It is worth noticing that NL degrades sharply on strongly noisy images due to its simply weighted averaging scheme for grey-level estimation based on the self-similarity structure in the image. In a strongly noisy image, the noise causes the structures lost and the structure similarity is unreliable.

The visual results corresponding to the best PSNR outputs on the noisy images "Pepper" and "House" with $\sigma = 20$ and 30 are showed in Fig. 4.4 and Fig. 4.5. The figures indicate that the results of our algorithm and NL are much better than the others. The outputs of GF and BF are under-smoothed with

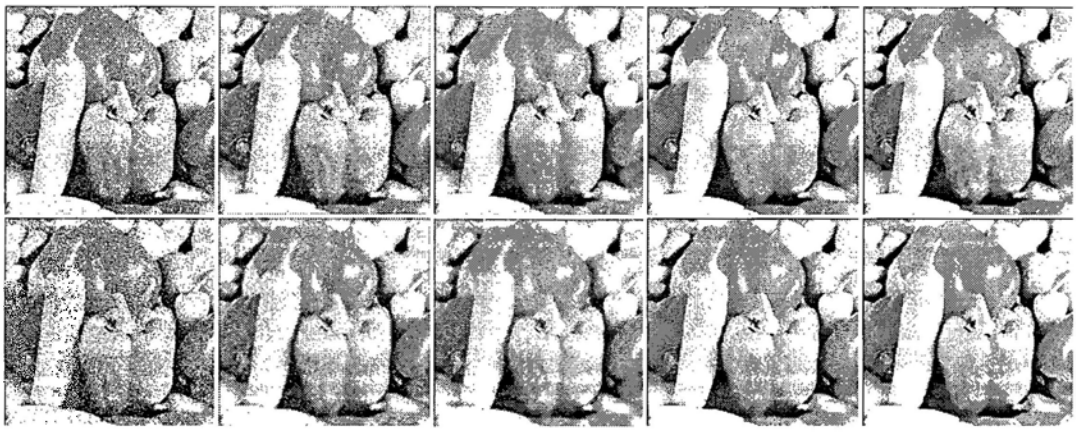

Figure 4.4: Results of the four algorithms on the "Pepper" image with the noise $\sigma = 20$ (the first row) and $\sigma = 30$ (the second row). From left to right: the noisy image, the results of GF, BF, NL, and our algorithm.

Figure 4.5: Results of the four algorithms on the "House" image with the noise $\sigma = 20$ (the first row) and $\sigma = 30$ (the second row). From left to right: the noisy image, the results of GF, BF, NL, and our algorithm.

σ		10	20	30	50	100
PSNR	GF	30.84	26.98	25.40	23.43	20.29
	BF	31.43	27.95	26.36	23.84	20.75
	NL	31.92	28.42	27.02	23.53	15.82
	BP	31.09	27.57	25.95	23.63	19.06
	GC	30.95	27.41	25.65	23.12	18.76
	Ours	**32.22**	**28.93**	**27.49**	**25.17**	**21.26**

Table 4.3: Average PSNR values on the 300 noisy images in the Berkeley segmentation benchmark.

obvious blurring effect. NL and our algorithm can remove the noise effectively while preserving the details and edges very well.

Fig. 4.6 shows the visual comparison among all the six algorithms on the noisy "Lena" image with $\sigma = 20$, 30 and 50. It is easy to see that our algorithm performs best.

The final experiment is carried out on the Berkeley segmentation benchmark [55]. All its 300 nature images are corrupted by the Gaussian noise with $\sigma = 10$, 20, 30, 50, and 100. The parameters for each algorithm are the same as those in the above experiments. Table 4.3 shows the average PSNR values, and indicates again that our algorithm outperforms the others.

It is also worth noticing that our algorithm can obtain the closed form solution and is much faster than BP and GC that are iterative algorithms, and the computational efficiency of our algorithm is also higher than the NL algorithm. For a 256×256 noisy image, the computational times taken by the BP, GC, GF, BF, NL, and our algorithms are about 30, 120, 0.3, 1, 300, and 10 seconds, respectively. GF, BF, NL, and our algorithms are implemented in Matlab, and BP and GC are in VC++. All the algorithms are running on a Pentium IV PC with a 2.8 GHz CPU.

Figure 4.6: Results of all the six algorithms on the "Lena" image with the noise $\sigma = 20$ (the first column), $\sigma = 30$ (the second column) and $\sigma = 50$ (the third column). From top to bottom: the noisy image, the results by BP, GC, GF, BF, NL, and our algorithm.

4.6 Conclusions

In this chapter, a novel image denoising algorithm has been proposed. The image denoising problem is formulated as an energy minimization problem based on the MRF model. The objective function we propose is a Gaussian MRF based energy. With the special design to incorporate pre-estimated edge information and patch similarity based pairwise interaction into the energy function, our algorithm can effectively reduce noise while maintaining image structures and details well. Furthermore, by relaxing the labels from discrete values to continuous values, a closed form global optimal solution can be obtained. In our extensive experiments, we compare our algorithm with two representative MRF-based denoising algorithms and three recent spatial filtering methods. The results clearly show that our algorithm outperforms these state-of-the-art algorithms both qualitatively and quantitatively.

Chapter 5

Summary and Discussion

Label assignment is the essential part in many low-level computer vision tasks such as image and video denoising, segmentation, and completion. With the spatial piecewise smoothness constraint, which is regarded as the intrinsic property of an image, label assignment is elegantly expressed as an MRF-based energy minimization problem, which has been demonstrated to well model these problems. For those video based applications, the temporal coherence constraint is further added into the energy function to ensure the label smoothness of the corresponding pixels in consecutive frames. Many approaches have been proposed to solve the MRF-based energy optimization problem, such as the iterated conditional modes, graph cuts, and belief propagation.

As the MRF model is a powerful general model for many image and video processing applications, for each particular problem in the context of an application, the keys for achieving good results lie in two parts: 1) constructing an objective energy function to well fit the problem; 2) developing an efficient and effective optimization technique to find the optimal results. In this dissertation, we propose three algorithms to handle the problems of interactive foreground object extraction, video completion, and image denoising, respectively. All of the three algorithms utilize MRF model to formulate the problems as MRF-based energy minimization problems, and design different optimization schemes to efficiently achieve the optimal results.

Our work has been published in or submitted to [49, 51, 16, 52, 50, 47, 48]. Next we conclude our work in this dissertation and discuss some future work as follows.

5.1 Contributions of Our Work

In Chapter 2, an iterative optimization based framework is proposed to address the problem of foreground object extraction from an image. We model the problem as an iterative MRF energy minimization problem to find the optimal label configuration. In each iteration, an MRF-based energy function with an iteratively refined initial probabilistic map of the image is designed. By optimizing the energy function in the continuous domain, a global optimal label configuration can be achieved, which can be regarded as a refined probabilistic map providing us the clues for updating the color models to estimate more accurate initial probabilistic map for the energy construction. Through the iterative optimization scheme, user provided information is iteratively propagated and expanded, which makes our work not sensitive to the number and locations of user provided seeds and require less cost of users' interaction and attention. Moreover, as the accuracy of the initial probabilistic map is iteratively improved ensuring a more precise MRF-based energy function as optimization target, high-quality foreground object extraction results can be obtained in the end by our algorithm.

Chapter 3 presents our work on video completion, which combines motion completion and global exemplar-based color completion. In the proposed algorithm, the motion field in the missing region of the video is filled locally first. Based on the completed motion field, color is restored in a global exemplar-based scheme by minimizing an MRF-based energy function. The global optimization problem is solved by a coarse-to-fine belief propagation scheme to avoid the computational impracticability caused by the large number of label

candidates in the optimization process. By using the motion and color information, our work preserves the temporal consistency based on the completed motion field, and globally optimizes the color completion process. It avoids the blurring effect caused by the sampling and blending process, while maintaining the video details and structures well. Besides, the proposed framework unifies the problems of image completion and video completion and solves them in a consistent form. Moreover, our algorithm is not restricted to videos containing periodic motion only and can handle a wide variety of videos, producing visually natural results without obvious artifacts.

In Chapter 4, we focus on the problem of image denoising, which is formulated as an MRF-based energy minimization problem with pre-estimated edge information and patch similarity based pairwise interaction involved. The optimization problem is generally NP-hard in discrete domain. In our work, by relaxing the labels from discrete values to continuous values, a closed form global optimal solution can be achieved. Compared with the related approaches, our work has the contributions: 1) A continuous closed form global optimal solution can be obtained, which provides a good prerequisite for our final result. 2) Image edges and details can be better preserved in our algorithm since pre-estimated edge information and patch based similarity are incorporated into the MRF energy function. 3) While obtaining better or comparable results, our algorithm is more efficient than belief propagation, graph cuts, and NL-means. 4) Our formulation for gray level image denoising can be directly extended to the denoising of color images without increasing the computational complexity.

5.2 Discussion and Future Work

For MRF-based image and video processing algorithms, there are three main parts: particular applications, well-defined MRF-based energy functions to be optimized, and developed optimization techniques. According to these three

parts, we will talk about the future work related to our work in this dissertation.

Other applications

Our model for image denoising with a Gaussian MRF energy and label relaxation based global optimal solution can be extended to the problem of stereo vision, which is to find the disparity (depth) map of a scene by using two images capturing the scene with slight different angles. The label in stereo is the disparity and the data penalty is calculated by using the Birchfield and Tomasi's pixel dissimilarity measurement [5]. The results on "Tsukuba" (16 labels) and "Venus" (20 labels) images are shown in Fig. 5.1. We can see that our work obtains better results than [78] and [41], which are customized for stereo vision with standard MRF energy function optimized by belief propagation and graph cuts respectively. However, there are many particular algorithms with special designed MRF energy functions for stereo, such as [77] and [33], which perform better than our general model.

As our discussion in Chapter 2 that by applying the thresholding process we can conveniently obtain a good trimap for the initialization of image matting, which is a process of estimating the foreground, the background, and the transparent factor (alpha) for each pixel. Since the MRF model is suitable to model the alpha map as well, we can try to formulate the segmentation and matting problems in an unified framework and therefore carry out the two tasks simultaneously.

Energy function design and optimization scheme development

For video based applications, the MRF energy function should not only enforce the spatial coherence but also the temporal consistency. Our work on video completion ensures the temporal consistency by adding the temporal term to the MRF energy function, which is constructed by the estimated local motions.

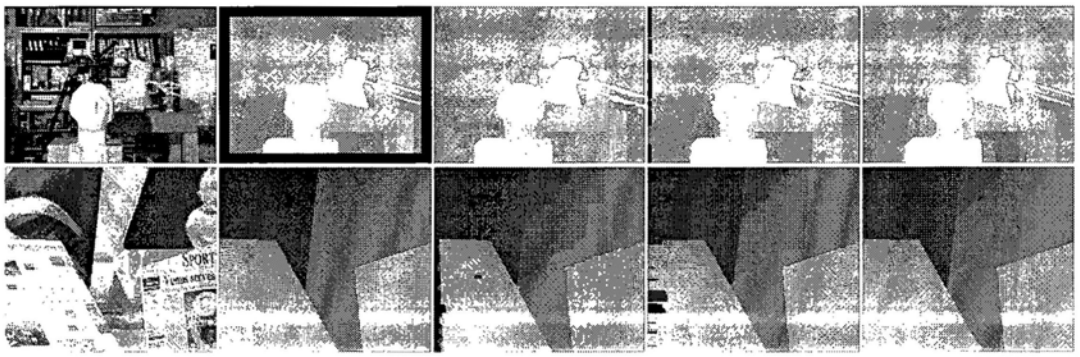

Figure 5.1: Stereo correspondence results on "Tsukuba"' and "Venus" images. From left to right: left image of the input image pair, ground truth, the results of belief propagation [78], graph cuts [41], and our algorithm.

The same idea can be applied to other video based applications, such as video denoising. By adopting our model for image denoising and adding a temporal constraint term in a similar way to the one in Chapter 3, we can construct a spatial-temporal MRF model for video denoising.

In our work on foreground object extraction, the core idea is to iteratively refine the color model and thus the optimization target, which will lead to more precise results and make the algorithm not very sensitive to the user interactions. The proposed framework is general and can be applied to different color models, energy functions, and optimization schemes, which still need further exploration. Moreover, although the analysis of user interaction is not quite related to the algorithm design, it is very important for the performance evaluation of the algorithms on interactive applications. Therefore, we can try to build some user interaction analysis criteria or benchmark data sets for this evaluation.

Bibliography

[1] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal on Computer Vision*, pages 43–77, 1994.

[2] M. Bertalmio, A. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[3] M. Bertalmıo, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. *ACM Transactions on Graphics (SIGGRAPH)*, 2000.

[4] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

[5] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:401–406, 1998.

[6] C. Bishop. *Pattern Recognition and Machine Learning*, New York, Springer, 2006.

[7] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. *Proceedings of European Conference on Computer Vision*, pages 428–441, 2004.

[8] J. Bouguet. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm. *Intel Corporation, Microprocessor Research Labs*, 2000.

[9] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proceedings of International Conference on Computer Vision*, 2001.

[10] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[11] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 648–655, 1998.

[12] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1222–1239, 2001.

[13] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 60–65, 2005.

[14] A. Buades, B. Coll, and J. Morel. Nonlocal Image and Movie Denoising. *International Journal of Computer Vision*, 76(2):123–139, 2008.

[15] F. Catte, P. Lions, J. Morel, and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal on Numerical Analysis*, 29(1):182–193, 1992.

[16] M. Chen, M. Liu, J. Liu, and X. Tang. Isoperimetric Cut on a Directed Graph. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[17] S. Chen, L. Cao, J. Liu, and X. Tang. Iterative MAP and ML Estimations for Image Segmentation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[18] P. Chou and C. Brown. The theory and practice of Bayesian image labeling. *International Journal of Computer Vision*, 4:185–210, 1990.

[19] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[20] D. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[21] I. Drori, D. Cohen-Or, and H. Yeshurun. Fragment-based image completion. *ACM Transactions on Graphics (SIGGRAPH)*, 2003.

[22] A. Efros and T. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of International Conference on Computer Vision*, pages 1033–1038, 1999.

[23] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006.

[24] D. Freedman and T. Zhang. Interactive graph cut based segmentation with shape priors. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[25] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741, 1984.

[26] M. Gleicher. Image snapping. *ACM Transactions on Graphics (SIGGRAPH)*, 1995.

[27] R. Gonzalez and R. Woods. *Digital Image Processing.* Prentice Hall, 2007.

[28] L. Grady. Multilabel random walker image segmentation using prior models. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 2005.

[29] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768, 2006.

[30] L. Grady and G. Funka-Lea. Multi-label Image Segmentation for Medical Applications Based on Graph-Theoretic Electrical Potentials. *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis: ECCV 2004 Workshops CVAMIA and MMBIA*, page 230, 2004.

[31] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *J. Royal Statistical Society*, pages 271–279, 1989.

[32] M. Hansen and W. Higgins. Watershed-based maximum-homogeneity filtering. *IEEE Transactions on Image Processing*, 8(7):982–988, 1999.

[33] L. Hong and G. Chen. Segment-based stereo matching using graph cuts. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[34] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[35] R. Hummel and S. Zucker. On the foundations of relaxation labeling processes. *Morgan Kaufmann Readings Series*, 1987.

[36] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1333–1336, 2003.

[37] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. *European Conference on Computer Vision*, pages 232–248, 1998.

[38] J. Jia, T. Wu, Y. Tai, and C. Tang. Video repairing: Inference of foreground and background under severe occlusion. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 364–371, 2004.

[39] Y. Jia, S. Hu, and R. Martin. Video completion using tracking and fragment merging. *The Visual Computer*, pages 601–610, 2005.

[40] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1568–1583, 2006.

[41] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. *Proceedings of IEEE International Conference on Computer Vision*, pages 508–515, 2001.

[42] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[43] N. Komodakis and G. Tziritas. Image completion using global optimization. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[44] J. Lee. Digital image smoothing and the sigma filter. *Computer Vision, Graphics and Image Processing*, 1983.

[45] A. Levin, A. Zomet, and Y. Weiss. Learning how to inpaint from global image statistics. *Proceedings of IEEE International Conference on Computer Vision*, pages 305–312, 2003.

[46] Y. Li, J. Sun, C. Tang, and H. Shum. Lazy snapping. *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

[47] M. Liu, M. Chen, and J. Liu. Clustering on Dependency Digraphs. *Proceedings of IEEE International Conference on Image Processing*, 2010.

[48] M. Liu, M. Chen, and J. Liu. Dimensionality Reduction via Tangential Learning. *Proceedings of IEEE International Conference on Image Processing*, 2010.

[49] M. Liu, S. Chen, and J. Liu. Precise object cutout from images. *Proceedings of ACM International Conference on Multimedia*, pages 623–626, 2008.

[50] M. Liu, S. Chen, and J. Liu. Continuous MRF Based Image Denoising with a Closed Form Solution. *Proceedings of IEEE International Conference on Image Processing*, 2010.

[51] M. Liu, S. Chen, J. Liu, and X. Tang. Video completion via motion guided spatial-temporal global optimization. *Proceedings of ACM International Conference on Multimedia*, pages 537–540, 2009.

[52] M. Liu, J. Liu, and X. Tang. Iterative Foreground Object Extraction from an Image. *Proceedings of ACM International Cconference on Multimedia*, 2010 (submitted).

[53] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of DARPA Image Understanding Workshop*, 1981.

[54] M. Lysaker and X. Tai. Iterative image restoration combining total variation minimization and a second-order functional. *International Journal of Computer Vision*, 66(1):5–18, 2006.

[55] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, pages 416–423, July 2001.

[56] M. Mignotte. A segmentation-based regularization term for image deconvolution. *IEEE Transactions on Image Processing*, 15(7):1973, 2006.

[57] M. Mignotte. Image denoising by averaging of piecewise constant simulations of image partitions. *IEEE Transactions on Image Processing*, 16(2):523, 2007.

[58] E. Mortensen and W. Barrett. Intelligent scissors for image composition. *ACM Transactions on Graphics (SIGGRAPH)*, pages 191–198, 1995.

[59] E. Mortensen and W. Barrett. Toboggan-based intelligent scissors with a four-parameter edgemodel. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 1999.

[60] A. Ogale and Y. Aloimonos. A roadmap to the integration of early visual modules. *International Journal of Computer Vision*, 72(1):9–25, 2007.

[61] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling. Simulation*, 4(2):460–489, 2005.

[62] K. Patwardhan, G. Sapiro, and M. Bertalmio. Video inpainting under constrained camera motion. *IEEE Transactions on Image Processing*, pages 545–553, 2007.

[63] M. Pelillo. The dynamics of nonlinear relaxation labeling processes. *Journal of Mathematical Imaging and Vision*, 7:309–323, 1997.

[64] P. Perez, A. Blake, and M. Gangnet. Jetstream: Probabilistic contour extraction with particles. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, 2001.

[65] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[66] L. Reese and W. Barrett. Image editing with intelligent paint. *EUROGRAPH*, pages 714–724, 2002.

[67] A. Rosenfeld, R. Hummel, and S. Zucker. Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man and Cybernetics*, 6:420–433, 1976.

[68] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, 2004.

[69] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[70] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992.

[71] Y. Shen, F. Lu, X. Cao, and H. Foroosh. Video completion for perspective camera under constrained motion. pages 63–66, 2006.

[72] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[73] J. Shi and C. Tomasi. Good features to track. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[74] T. Shih, N. Tang, W. Yeh, T. Chen, and W. Lee. Video inpainting and implant via diversified temporal continuations. *Proceedings of ACM International Conference on Multimedia*, pages 133–136, 2006.

[75] T. Shiratori, Y. Matsushita, S. Kang, and X. Tang. Video completion by motion field transfer. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 411–418, 2006.

[76] S. Smith and J. Brady. Susana new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, 1997.

[77] J. Sun, Y. Li, S. Kang, and H. Shum. Symmetric stereo matching for occlusion handling. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[78] J. Sun, N. Zheng, and H. Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:787–800, 2003.

[79] R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *International Journal of Computer Vision*, 5:271–301, 1990.

[80] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1068–1080, 2008.

[81] E. Tadmor, S. Nezzar, and L. Vese. A multiscale image representation using hierarchical (BV, L 2) decompositions. *Multiscale Model Simulation*, 2(4):554–579, 2004.

[82] M. Tappen and W. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. *IEEE International Conference on Computer Vision*, pages 900–906, 2003.

[83] H. Ting, S. Chen, J. Liu, and X. Tang. Image inpainting by global structure and texture propagation. *Proceedings of ACM International Conference on Multimedia*, pages 517–520, 2007.

[84] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of International Conference on Computer Vision*, page 839, 1998.

[85] Y. Weiss and W. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, 47(2):736–744, 2001.

[86] Y. Wexler, E. Shechtman, and M. Irani. Space-time video completion. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[87] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

[88] Q. Yang, X. Tang, C. Wang, M. Chen, and Z. Ye. Progressive Cut: An Image Cutout Algorithm that Models User Intentions. *IEEE Multimedia*, 2007.

[89] Y. Zhang, J. Xiao, and M. Shah. Motion layer based object removal in videos. *IEEE Workshops on Application of Computer Vision, WACV/MOTIONS*, 2005.

[90] S. Zhao and M. Venkatesh. Efficient Object-Based Video Inpainting. *Proceedings of IEEE International Conference on Image Processing*, pages 705–708, 2006.