Robust Patch Estimation for Exemplar-based Image Inpainting

Wang Jiawen and Zhang Hongbin
College of Computer Science, Beijing University of Technology, Beijing, China

Abstract - The exemplar-based image inpainting algorithm consists of two procedures: patch selection and patch estimation. In this paper, we focus on the latter. Conventional exemplar-based approaches cannot fully exploit the information existing in the image, which makes the patch estimation unstable and produces false colors and artifacts. In this paper, we propose a robust patch estimation algorithm for exemplar-based image inpainting. A set of most relevant patches is extracted as reference patches from the known region of the image. A dictionary is learned both globally and adaptively to best represent the target patch, which is then approximated under the assumption that similar patches admit similar decompositions. Experiments show that both quantitative and qualitative improvements are achieved by our method.

Keywords: Image Inpainting, Image Completion, Sparse Representation

1 Introduction

Image inpainting, also known as image completion, is a technique that aims to automatically filling in missing or damaged regions of an image in a visually plausible way. This technique has been widely applied in various fields, including object removal, image and video restoration, data compression and network data transmission, and become a fundamental area of research in image processing. Many works have been proposed to investigate this problem. In general, the approaches of image inpainting can be classified into two major categories: diffusion-based and exemplar-based.

The pioneering work of the first category was proposed by Bertalmio et al. [1], whose algorithm performed inpainting by continuously propagating linear structures into the missing region via diffusion using nonlinear partial differential equation (PDE). Following their work, Chan and Shen proposed a variational regularization framework with total variation (TV) [2]. One of the limitations of diffusion-based approaches is that they only use local information surrounding the missing region, which restricts them to filling in a smooth and narrow region like a speckle or scratch. When dealing with a large and textured region, a noticeable blur effect could be produced.

The second category of approaches performs inpainting in a copy-and-paste manner, i.e., finds the best-match sample from the known region and copies it directly into the missing region. Criminisi et al. [3] proposed a best-first filling algorithm which preferred patches along the image structure to be filled first. Following this line of research, Wu et al. [4] put forward a cross isophotes inpainting algorithm, in which a cross-isophotes patch priority term was designed to determine the filling order. Instead of a one-pass greedy way, Bugeau et al. [5] proposed an iterative inpainting method by minimizing an energy functional of the correspondence map. The main benefit of exemplar-based techniques is that they utilize non-local information, especially when the corresponding information for filling locates far away from the missing region. Thus, exemplar-based approaches can achieve a more plausible result for a large missing region in most cases. Nevertheless, the known region sometimes cannot provide enough information for filling, in which case an unexpected result may be obtained.

Recently, sparse representation is introduced to the inpainting problem, in which the image patch is synthesized as a sparse linear combination of atoms over a dictionary. Elad et al. [6] proposed an inpainting algorithm by separating the image into cartoon and texture layers, and respectively represented these layers using two incoherent dictionaries. Fadili et al. [7] put forward an EM algorithm for image inpainting based on a penalized maximum likelihood formulated using linear sparse representations. Mairal et al. [8] proposed a sparse representation scheme for color image restoration, in which an adaptive dictionary is learned to sparsely represent the image patch. Assuming that the dictionary is abundant, the missing region still can be recovered when the known region cannot provide enough information. Nevertheless, similar patches sometimes admit very different estimates due to the potential instability of sparse decompositions, particularly when the target patch to inpaint is partially known.

The main contribution of this paper is a robust patch estimation method for image inpainting. All available information, including the known pixels of the target patch, candidate patches in the known region and a learned dictionary, is fully exploited to improve the inpainting quality under a sparse representation framework. Our approach can also be integrated into other inpainting schemes to achieve good visual quality.
2 Motivation

Before describing the proposed method, it is important to provide a brief overview of exemplar-based inpainting techniques. The overall process of exemplar-based inpainting is illustrated in Fig. 1. Let us denote by \( \Omega \) the missing region (called the target region). The known region \( \Phi \) (called the source region) is defined as the entire image \( \mathbf{I} \) minus the target region, i.e., \( \Phi = \mathbf{I} - \Omega \). For a fixed size patch \( \Psi_p \), centered at pixel \( p \), the algorithm first calculates the priority of patches along the fill front \( \partial \Omega \), and selects one with the highest priority as the target patch \( \hat{\Psi}_p \), which is then filled with the best-match sample \( \hat{\Psi}_q \) found from the source region.

![Fig. 1. Illustration of exemplar-based inpainting. (a) Determine the target patch \( \hat{\Psi}_p \). (b) Find the best-match sample \( \hat{\Psi}_q \). (c) Copy the content from \( \hat{\Psi}_q \) to \( \hat{\Psi}_p \) correspondingly.](image)

2.1 Patch estimation

The success of the exemplar-based algorithms [3-5] is mainly attributed to exploiting self-similarity and coherence of natural images. Self-similarity refers to an amount of repetitions of local information in natural images, as illustrated in Fig. 2. This feature is firstly utilized by Efros et al. [9], in which the texture is synthesized by pixels with similar neighborhood. The spatial distribution of texture induces coherence between neighboring pixels. Thus, it is reasonable to estimate the target patch with a similar one.

Denote by \( \Psi_p \) the target patch to be inpainted, an exhausted search is performed in the source region \( \Phi \) to find the best-match sample \( \Psi_q \) in [3, 4]:

\[
\Psi_q = \arg \min_{\Psi_q \in \Phi} d \left( \mathbf{M}_p \Psi_p, \mathbf{M}_q \Psi_q \right)
\]  

(1)

Where \( d \) is a distance metric, which usually is the Sum of Squared Differences (SSD). \( \mathbf{M} \) and \( \bar{\mathbf{M}} \) are binary masks to extract the known and missing pixels of \( \Psi_p \) respectively. Then, image data is copied from \( \Psi_q \) to \( \Psi_p \):

\[
\bar{\mathbf{M}} \Psi_p = \bar{\mathbf{M}} \Psi_q
\]  

(2)

![Fig. 2. Illustration of self-similarity. Similar patches are marked with black squares.](image)

2.2 From best match to multiple candidates

The estimation by the best-match sample sometimes is unstable for image inpainting. Particularly, when the missing region is large, the false information may be duplicated to produce obvious artifacts (garbage growing [3]). This can be improved by estimating over multiple similar samples (called candidate patches), as illustrated in Fig. 3. [10] and [11] exploits this idea from different views. In [10], the target patch was estimated by the non-local means of candidate patches in the known region. [11] assumed that the target patch admitted a sparse linear combination of candidate patches. Denote by \( \Psi_p \) the target patch, the estimation \( \hat{\Psi}_p \) by both of the two methods can be interpreted as a linear combination of \( n \) candidate patches \( \{ \Psi_i \} \), \( 1 \leq i \leq n \), which is formulated as follows:

\[
\hat{\Psi}_p = \sum_{i=1}^{n} a_i \Psi_i
\]  

(3)

Where \( a \) is the coefficient vector. Compared with Eq. (1), multiple candidates can alleviate the influence of each sample, and strengthen the stability of estimation.

![Fig. 3. Illustration of estimation over multiple candidate patches. (a) Patch estimation by Eq. (1). (b) Patch estimation by Eq. (3).](image)
2.3 From correspondence map to sparse representation

There is a close relationship between the exemplar-based inpainting problem and a correspondence map [12] \( F: \Omega \rightarrow \Phi \), which associates each pixel of the image to an original pixel such that

\[
F(p) = \begin{cases} 
  p, & p \in \Phi \\
  q \in \Phi, & p \in \Omega
\end{cases}
\] (4)

In an extreme condition, if the known region cannot provide enough information for filling, the target patch may be filled by some irrelevant sample, which can lead to an unexpected result. Note that estimation over multiple candidate patches still suffers from this case.

More prior information needs to be considered. Given a dictionary matrix \( D = [d^1, \ldots, d^K] \) with \( K \) columns referred to as atoms, we assume that this matrix is known and fixed. The sparse representation model suggests that every image patch \( p_\Psi \) could be sparsely represented over this dictionary, which satisfies:

\[
\min_\alpha \| \alpha \|_0 \quad \text{s.t.} \quad p_\Psi \approx \alpha D \alpha
\] (5)

Where \( \alpha \) is the representation coefficient, \( \| \cdot \|_0 \) norm which promotes sparsity. \( \| \alpha \|_0 \) stands for the count of the nonzero entries in \( \alpha \). In practice, the dictionary \( D \) can be chosen in advance. Assuming that the dictionary is abundant, the target patch still can be recovered when the known region cannot provide enough information for filling. Then, the estimated patch \( ˆp_\Psi \) is computed by

\[
  ˆp_\Psi = \sum_{i=1}^{\xi} \alpha_i d_i
\] (6)

We can see from Eq. (6) that there is a close relationship between estimation over multiple candidate patches and sparse representation. If we consider candidate patches as atoms, Eq. (3) can be viewed as a special case of sparse representation. Nevertheless, similar patches sometimes admit very different estimates due to the potential instability of sparse decompositions, particularly when the target patch to inpaint is partially known. Thus, we need to exploit all prior information, including the known pixels of the target patch, candidate patches in the known region and a learned dictionary, and estimate the target patch under the sparse representation framework with local and global patch consistency constraints described in Section 3.1.

3 Proposed algorithm

In this section, we show how to improve the stability of estimation in a sparse representation framework under the assumption that similar patches admit similar decompositions.

3.1 Robust patch estimation

Given the target patch \( p_\Psi \) to be inpainted, \( M \) is a binary mask to extract the known pixels in \( p_\Psi \). Assuming that \( p_\Psi \) admits a sparse representation over a dictionary \( D = [d^1, \ldots, d^K] \) with \( K \) atoms, one can find a linear combination of atoms from \( D \) that is close to the target patch \( p_\Psi \) over the known pixels. Under a square loss, this optimization problem can be written as

\[
\min_{a_\Psi} \| a_\Psi \|_2 \quad \text{s.t.} \quad M(p_\Psi - Da_\Psi) \leq \varepsilon_1
\] (7)

Where \( \varepsilon_1 \) is a parameter to control the error tolerance of this approximation. \( a_\Psi \) is the representation coefficient of \( p_\Psi \). Eq. (7) is referred to as the local patch consistency constraint. We further assume that the estimated patch \( ˆp_\Psi \) should share similar decompositions with similar patches. Let us define the set of candidate patches sampled from the source region as

\[
S_p \triangleq \{ i = 1, \ldots, N \quad \text{s.t.} \quad d(p_\Psi - p_i) \leq \xi \}
\] (8)

Where \( N \) is the number of candidate patches, \( d \) is a metric function to evaluate the similarity between two patches, \( \xi \) is a threshold. The sparse representations of candidate patches are defined as

\[
\min_{a_i} \| a_i \|_2 \quad \text{s.t.} \quad \sum_{i \in S_p} \| p_i - Da_i \|_2 \leq \varepsilon_2
\] (9)

Under the constraints (7) and (9), a global patch consistency constraint is imposed, which is formulated as

\[
\min_{A} \| A \|_{u=0} \quad \text{s.t.} \quad \sum_{i \in S_p} \| p_i - Da_i \|_2 \leq \varepsilon_1 \\
\sum_{i \in S_p} \| p_i - Da_i \|_2 \leq \varepsilon_2
\] (10)

Where \( \lambda \) is a parameter balancing the strength of the constraints in (7) and (9). \( A \) is the coefficient matrix defined as \( A = [a_\Psi, a_i] \), \( i \in S_p \). \( \| \cdot \|_{u=0} \) is a group-sparsity regularizer, which stands for the count of nonzero rows. The global patch
consistency constraint forces the target patch and its candidate patches should be decomposed by the same set of atoms, i.e. similar patches admit similar sparse decompositions, as illustrated in Fig. 4(b).

Denote by $A_i$ the $i$-th column of $A$, above constrained optimization model can also be formulated as an energy minimization problem:

$$
\hat{A} = \arg \min_{A} \left\{ \beta \|A\|_{\text{row}} + \|M(\Psi_p - DA_p)\|_2^2 + \gamma \sum_{i \in S_p} \|\Psi_i - DA\|_2^2 \right\}
$$

(11)

It is equivalent to the constrained optimization problem in (10) when $\beta$ and $\gamma$ are properly chosen. Then, the estimated patch $\hat{\Psi}_p$ is obtained by

$$
\hat{\Psi}_p = DA_p
$$

(12)

### 3.2 Implementation details

- **Solution to Eq. (10)**

Eq. (10) is a NP-hard problem due to its combinatorial nature. We solve it by a modification of Simultaneous Orthogonal Matching Pursuit (S-OMP) [13] for its simplicity and efficiency in a greedy manner. The procedure is shown in Fig. 5.

- **Choice of the dictionary**

A dictionary with 12000 atoms is learned in advance based on 24 photos of Kodak PhotoCD. 10000 samples are extracted from each image for training. This pre-learned global dictionary is used for each image to be inpainted. An adaptive dictionary is also learned from samples in the source region. Then, the target patch is represented by a joint one, which takes advantage of the benefits of these two dictionaries. K-SVD [14] is used for dictionary learning due to its efficiency for small scale training.

- **Determination of candidate patches**

When building the set of candidate patches $S_p$, we restrict the search in a fixed size neighborhood. This semi-local approach can reduce the time-cost and exclude irrelevant samples. The SSD metric is used in Eq. (8) to evaluate the similarity between two patches.

- **Extension to color images**

The extension to color can be easily performed by a simple concatenation of the RGB values to a single vector. This usually gives better results and produces fewer artifacts compared with performing in each channel separately.
In this paper, we have introduced a robust patch estimation method for exemplar-based image inpainting. All available information, including the known pixels of the target patch, candidate patches from the source region and a learned dictionary, is taken advantage of to approximate the target patch in a sparse representation framework under the assumption that similar patches admit similar decompositions. The experimental results showed that the proposed method achieved better visual quality compared with current exemplar-based inpainting approaches. Future work includes investigating an alternative priority computation method and an extension to video restoration.

6 References


Table 1. PSNR/SSIM results for text and scratch removal.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 6(a)</td>
<td>28.85/0.89</td>
<td>29.31/0.91</td>
<td>30.26/0.90</td>
<td><strong>30.57/0.91</strong></td>
</tr>
<tr>
<td>Fig. 6(b)</td>
<td>29.18/0.88</td>
<td>29.78/0.90</td>
<td>30.04/0.89</td>
<td><strong>30.64/0.89</strong></td>
</tr>
<tr>
<td>Fig. 6(c)</td>
<td>33.88/0.92</td>
<td>35.11/0.93</td>
<td>34.93/0.93</td>
<td><strong>35.27/0.93</strong></td>
</tr>
<tr>
<td>Fig. 6(d)</td>
<td>29.78/0.96</td>
<td><strong>30.49/0.97</strong></td>
<td>28.68/0.95</td>
<td>29.26/0.96</td>
</tr>
<tr>
<td>Mean</td>
<td>30.42/0.91</td>
<td>31.17/0.93</td>
<td>30.98/0.92</td>
<td><strong>31.43/0.92</strong></td>
</tr>
</tbody>
</table>


Fig. 6. Comparison for text and scratch removal. The first to sixth rows show the original images, the degraded images, the inpainted results of the best-match method [3], the non-local means method [10], the sparse representation method [8], and our method.
Fig. 7. Comparison for object removal. (a) The original image. (b) The degraded image. (c) – (f) The inpainted results of the best-match method [3], the non-local means method [10], the sparse representation method [8], and our method.