POSTECH-MLG-2011-011

Machine Learning Group

Department of Computer Science, POSTECH

Patch-Based Markov Random Fields for Fast Face Occlusion Recovery [a]

Jeong-Min Yun and Seungjin Choi

Machine Learning Lab
Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu
Pohang 790-784, Korea
Email: {kshkawa,seungjin}@postech.ac.kr

Abstract

In this paper we present Markov random field (MRF) models for face occlusion detection and recovery. For occlusion detection, we use a pixel-based pair-wise MRF model (which is similar to the Ising model) where the binary mask on each pixel is inferred to decide the presence of occlusion. Then we construct a patch-based nonparametric pair-wise MRF model for occlusion recovery, which is learned using occlusion-free face images in the training set. Probabilistic inference using α-expansion leads to fast occlusion recovery, compared to the existing method. Numerical experiments confirm that our method speeds up the existing method by several orders of magnitude, while the quality of recovery is as good as the existing one.

[a] to be presented at MLSP-2011.

Contents

Machine Learning Lab
Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu
Pohang 790-784, Korea

1 Introduction

Image inpainting (also known as image completion) is an important and challenging problem in computer graphics and vision, which involves restoring damaged or missing regions in an undetectable form [1]. The basic idea behind image inpainting is to propagate structure and texture information from the known or neighbor regions, to fill-in damaged or missing regions, exploiting the spatial continuity in images [2, 11].

In order to model the spatial continuity in images, a Markov random field (MRF) model has been widely used in various computer vision problems [7] since it firstly introduced in the computer vision community [4]. For image inpainting problem, various MRF models have also been proposed to capture local structure and texture information in images [11, 6]. Although they suggested reasonable MRF models and showed high-quality inpainting results, it seems that the perfect MRF model for this problem hasn't been developed yet. Furthermore, even if the perfect one is designed, finding optimal configuration of general MRFs is NP-hard, so there is still room for improvement.

Recently, MRF modeling technique was successfully applied to a face occlusion recovery problem [8], which is a task highly related to image inpainting and aims at restoring occluded regions in face images. They propose an automatic face occlusion detection and recovery algorithm with MRF models for each process. In the detection process, the spatial continuity between adjacent pixels are modeled by a simple pair-wise MRF, and the binary result, either occluded or non-occluded, are easily obtained by the graph cut method. In the recovery process, each pixel is modeled by a more complicated model, they call it the quality assessment model, involving more than two neighborhood pixels, and a greedy search algorithm is proposed for inference. Although they show the quite better quality of recovery results than previous methods, the model of the recovery process is inefficient, so it takes a lot of time to solve. This may be a serious problem, because the face occlusion recovery problem has attracted increasing attention with the needs of the more practical face recognition system, where an immediate recovery result is usually needed.

In this paper we adopt the generative model in [8], but propose a more efficient recovery algorithm. We model the complex relationship between several neighborhood pixels by patch-based pair-wise MRF, in which the number of vertices in the graphical model is significantly reduced, but the model steel keeps both global correlation and local patterns of the quality assessment model in [8]. With an efficient graph cut based inference algorithm, α-expansion, we compare our method with theirs on the recovery quality and execution time.

2 Probabilistic Model

As in [8], we consider the generative model shown in Fig. 1. We assume that the input image is an 8-bit grayscale image with the m total number of pixels. $\boldsymbol{s} \in \{0, ..., 255\}^m$ denotes a true face image, which is our final goal. $\boldsymbol{b} \in \{-1, 1\}^m$ denotes a binary mask that determines whether each pixel is occluded or not; 1 for non-occluded pixels and -1 for occluded pixels. $\boldsymbol{o} \in \{0, ..., 255\}^m$ denotes an occluded face image which is generated by $\boldsymbol{s}$ and $\boldsymbol{b}$. With the index i, s_i, b_i, and o_i represent values of the i-th pixel in $\boldsymbol{s}, \boldsymbol{b}$, and $\boldsymbol{o}$.

With the assumption that $\boldsymbol{s}$ and $\boldsymbol{b}$ are independent, the joint probability is:

$$p(\boldsymbol{s}, \boldsymbol{b}, \boldsymbol{o}) = p(\boldsymbol{s})p(\boldsymbol{b})p(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{b}). \tag{1}$$

Formulation of $p(\boldsymbol{s})$ and $p(\boldsymbol{b})$, the prior distributions, would be explained in next section. $p(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{b})$, the conditional distribution of $\boldsymbol{o}$, can be formulated as follows.

For each pixel i, $p(o_i|s_i, b_i)$ describes a generation process of o_i according to the values of b_i and s_i. If $b_i = 1$, which means the position i in $\boldsymbol{o}$ is non-occluded, o_i and s_i should be

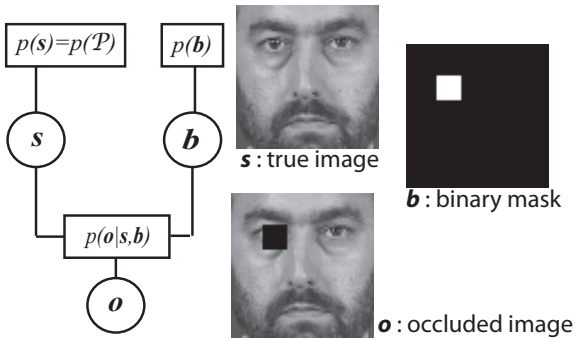

Figure 1: The factor graph of the generation of the occluded face image.

same. Thus, $p(o_i|s_i, b_i = 1)$ represents the similarity between o_i and s_i using a Gaussian:

$$p(o_i|s_i, b_i = 1) = \frac{1}{Z_i} \exp\left\{-\frac{1}{2\sigma^2}(o_i - s_i)^2\right\}, \tag{2}$$

where Z_i is a normalization constant. If $b_i = -1$, the conditional probability may be related with the covering object on the occluded region, which is omitted from our generative model. Instead of a complicated model for that, the probability with fixed-value over o_i would be enough, such as

$$p(o_i|s_i, b_i = -1) = \frac{1}{Z_i} \exp\{-c\}, \tag{3}$$

where c is a fixed constant over all pixels. Since its distribution can be implicitly changed by the effect of Z_i in (2), the probability becomes higher when the pixel difference between $\boldsymbol{s}$ and the covering object is bigger.

The objective of our problem is to infer the hidden variables $\boldsymbol{b}$ and $\boldsymbol{s}$ from the observable variable $\boldsymbol{o}$. However, computing the both at once is infeasible. Therefore, we solve the problem one by one; at first, compute $\boldsymbol{b}$, and then, compute $\boldsymbol{s}$. The former and latter correspond to the detection and recovery process, respectively.

Given $\boldsymbol{o}$ and the initial guess of the true face image $\boldsymbol{s}^0$, the detection process should find the optimal $\boldsymbol{b}$ which maximizes the posterior probability $p(\boldsymbol{b}|\boldsymbol{s}^0, \boldsymbol{o})$. Bayes' rule for this is:

$$p(\boldsymbol{b}|\boldsymbol{s}^0, \boldsymbol{o}) = \frac{p(\boldsymbol{o}|\boldsymbol{s}^0, \boldsymbol{b})p(\boldsymbol{b})}{p(\boldsymbol{o}|\boldsymbol{s}^0)}, \tag{4}$$

since $p(\boldsymbol{b}|\boldsymbol{s}^0) = p(\boldsymbol{b})$ from their independence assumption. Then, the maximum a posteriori (MAP) estimate of $\boldsymbol{b}$ is:

$$\boldsymbol{b}^* = \arg\max_{\boldsymbol{b}} p(\boldsymbol{o}|\boldsymbol{s}^0, \boldsymbol{b})p(\boldsymbol{b}). \tag{5}$$

After detection, the recovery process finds the optimal $\boldsymbol{s}$ which maximizes $p(\boldsymbol{s}|\boldsymbol{b}^*, \boldsymbol{o})$. Similar to above, the MAP estimate of $\boldsymbol{s}$ is:

$$\boldsymbol{s}^* = \arg\max_{\boldsymbol{s}} p(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{b}^*)p(\boldsymbol{s}). \tag{6}$$

3 Markov Random Field Models

For $p(\boldsymbol{b})$ and $p(\boldsymbol{s})$, the main characteristics we want to capture from them are the spatial continuity among the pixels or patches. These can be modeled by a Markov random

field (MRF). When variables $X = \{x_1, x_2, ..., x_n\}$ are modeled as a MRF with the graph $G = (V, E)$, where vertices and edges correspond to variables X and dependencies between variables respectively, [4] showed that the joint probability of X can be represented as

$$p(X) = \frac{1}{Z_X} \exp\left\{-\frac{U(X)}{T}\right\}, \tag{7}$$

where

$$U(X) = \sum_{c \in C} V_c(X) \tag{8}$$

and C is a collection of maximum cliques of G. We usually say that $U(X)$ is an energy function, Z_X is the partition function, T is the temperature, and $V_c(X)$ is a clique potential.

When G is a 2D grid graph, which is our case, the energy function can be further simplified as

$$U(X) = \sum_{i \in V} V_i(x_i) + w \sum_{(i,j) \in E} V_{i,j}(x_i, x_j), \tag{9}$$

where w is the weight between V_i and $V_{i,j}$. V_i is called a node potential or data energy, and it is related to the likelihood between the observable variable and the hidden variable. $V_{i,j}$ is called a edge potential or smoothness energy that models spatial coherence of hidden variables.

3.1 Pixel-based MRF for Detection

In order to use (5) for detection, we need to specify $\boldsymbol{s}^0$, the initial guess of $\boldsymbol{s}$, in advance. Although more complicated method may produce better results, the average of all face images in the training set is enough in our case.

For the detection, $p(\boldsymbol{b}|\boldsymbol{s}^0, \boldsymbol{o})$ is modeled as a pixel-based MRF. Given a 2D grid graph $G^{det} = (\boldsymbol{b}, E)$, a vertex $b_i \in \boldsymbol{b}$ corresponds to a binary mask of i-th pixel position in the image, and the occlusion usually tends to appear in contiguous region like sunglasses and a scarf. Then, this prior knowledge can be modeled by the classical Ising model in $p(\boldsymbol{b})$ as

$$p(\boldsymbol{b}) = \frac{1}{Z_b} \exp\left\{\frac{1}{\sigma_b^2}\left(\sum_{b_i \in \boldsymbol{b}} \lambda_i b_i + \sum_{(i,j) \in E} \lambda_{ij} b_i b_j\right)\right\}, \tag{10}$$

where λ_i reflects prior knowledge about the possibility of occlusion of the pixel, and we simply set to 0 for all i, implying that we don't give any prior possibility of occlusion. λ_{ij} controls the relation between adjacent pixels, and we set to 1 for all (i, j), so we give same weights to all adjacent pixel pairs.

Plugging (2), (3), and (10) into (5), $p(\boldsymbol{b}|\boldsymbol{s}^0, \boldsymbol{o})$ can be formulated as another pixel-based MRF:

$$p(\boldsymbol{b}|\boldsymbol{s}^0, \boldsymbol{o}) = \frac{1}{Z_{b'}} \exp\{-U(\boldsymbol{b})\}, \tag{11}$$

where

$$U(\boldsymbol{b}) = \sum_{b_i \in \boldsymbol{b}} V_i^{det}(b_i) + w_b \sum_{(i,j) \in E} V_{i,j}^{det}(b_i, b_j). \tag{12}$$

From (2) and (3), the node potential is

$$V_i^{det}(b_i) = \begin{cases} V_i^{det}(b_i = 1) = (o_i - s_i^0)^2 \\ V_i^{det}(b_i = -1) = c \end{cases}. \tag{13}$$

It follows from (10), the edge potential is

$$V_{i,j}^{det}(b_i, b_j) = -b_i b_j. \tag{14}$$

Since $\boldsymbol{b}$ is a binary vector, and for all edges (i, j) in the graph, (14) satisfies the submodular condition: $V_{i,j}(b_i = -1, b_j = -1) + V_{i,j}(b_i = 1, b_j = 1) \leq V_{i,j}(b_i = -1, b_j = 1) + V_{i,j}(b_i = 1, b_j = -1)$, the problem can be represented as a binary graph cut problem, and the global optimum can be found in polynomial time using the max-flow algorithm [5].

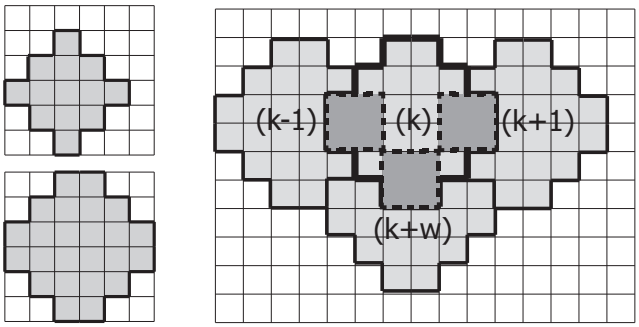

Figure 2: (a) Diamond-shaped 3×3 and 4×4 patches, where each square corresponds to each pixel in the image. (b) Example mapping of 4×4 patches on the image, where $w = \lceil (\text{the width of the image})/4 \rceil$

3.2 Patch-based MRF for Recovery

For the recovery process, a vertex $\mathcal{P}_k$ in the graph is a diamond-shaped $r \times r$ patch (Fig. 2 (a)) in $\boldsymbol{s}$, not a pixel. This type of the patch is introduced motivated by [9] to reduce the inference time of the recovery process. Let $\mathcal{P}$ be the collection of $\mathcal{P}_k$ in S. It is exactly same as $\boldsymbol{s}$ when we construct the image only using $r \times r$ center pixels of each patch, so (6) can be reformulated as

$$\arg\max_{\mathcal{P}} p(\boldsymbol{o}|\mathcal{P}, \boldsymbol{b}^*) p(\mathcal{P}). \tag{15}$$

However, unlike the detection case, $p(\boldsymbol{o}|\mathcal{P}, \boldsymbol{b}^*)$ is not a suitable probability measure to the recovery process: After the detection, $\boldsymbol{b}^*$ provides binary values for all pixels. If a pixel i in the patch k is occluded ($b_i = -1$), The recovery of that pixel should be represented by appropriate modeling, but $p(o_i|\mathcal{P}_k, b_i^* = -1)$ is always constant (3).

Therefore, we first generate the intermediate recovery result $\boldsymbol{s}^g = \mathcal{P}^g$ as an observation for the later patch-based MRF model of the recovery process. Same as the quality assessment model in [8], we capture the global correlation of faces from the generation of $\boldsymbol{s}^g$ using the probabilistic PCA model [13]. With the m total number of pixels in $\boldsymbol{s}$ and q ($q < m$) dimensional hidden variables, consider the linear generative model of $\boldsymbol{s}$:

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{16}$$

where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}_{q\times 1}, \boldsymbol{I}_q)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}_{m\times 1}, \sigma^2 \boldsymbol{I}_m)$. From this model, $\boldsymbol{s}$ satisfies the following Gaussian: $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}_m)$. Then, the maximum-likelihood estimators for $\boldsymbol{W}\boldsymbol{W}^\top$ and σ^2 can be explicitly computed:

$$(\boldsymbol{W}\boldsymbol{W}^\top)_{ML} = \boldsymbol{U}_q(\boldsymbol{\Lambda}_q - \sigma_{ML}^2 \boldsymbol{I}_q)\boldsymbol{U}_q^\top, \tag{17}$$

where $\mathbf{\Lambda}_q$ is a diagonal matrix with the q largest eigenvalues, $\boldsymbol{U}_q$ is an $m \times q$ matrix with corresponding eigenvectors, and σ^2_{ML} is the average of the remaining eigenvalues. Now, the probability distribution of $\boldsymbol{s}$ can be represented by

$$p(\boldsymbol{s}) = \frac{1}{Z_s} \exp\left\{-(\boldsymbol{s}-\boldsymbol{\mu})^\top \boldsymbol{G}(\boldsymbol{s}-\boldsymbol{\mu})\right\}, \tag{18}$$

where

$$\begin{aligned} Z_s &= \sqrt{(2\pi)^m |(\boldsymbol{W}\boldsymbol{W}^\top)_{ML} + \sigma^2_{ML}\boldsymbol{I}_m|}, \\ \boldsymbol{G} &= \frac{1}{2}\left((\boldsymbol{W}\boldsymbol{W}^\top)_{ML} + \sigma^2_{ML}\boldsymbol{I}_m\right)^{-1}. \end{aligned}$$

With other pixels $s_j, j \neq i$ fixed, (18) can be reformulated with s_i as

$$\begin{aligned} p(s_i) &= \frac{1}{Z_{s_i}} \exp\Big\{-G_{ii}(s_i-\boldsymbol{\mu}_i)^2 \\ &\quad -2\sum_{j\neq i} G_{ij}(s_i-\boldsymbol{\mu}_i)(s_j-\boldsymbol{\mu}_j)\Big\}. \end{aligned} \tag{19}$$

By maximizing this, we can compute the recovery result of an occluded pixel s_i using other non-occluded pixels as

$$s_i^g = \arg\max_{s_i} p(s_i) = \boldsymbol{\mu}_i - \frac{1}{\boldsymbol{G}_{ii}} \sum_{i\neq j} \boldsymbol{G}_{ij}(s_j - \boldsymbol{\mu}_j). \tag{20}$$

In addition, to avoid the use of pixel values of unrecovered region, the indicator function $\mathbf{1}_s$ is introduced; 0 for occluded pixels which haven't been recovered yet, 1 for the other pixels. Finally, each pixel s_i^g can be computed as

$$\begin{cases} \boldsymbol{\mu}_i - \frac{1}{\boldsymbol{G}_{ii}} \sum_{i\neq j} \mathbf{1}_s(j)\boldsymbol{G}_{ij}(s_j-\boldsymbol{\mu}_j), & \text{for } b_i = 1, \\ o_i, & \text{for } b_i = -1. \end{cases}. \tag{21}$$

Now, $p(\mathcal{P}|\mathcal{P}^g)$ is modeled by the patch-based nonparametric MRF [9]. Before doing this, from the result of the detection, we define the overlap function of patches. $O(\mathcal{P}_k)$ denotes a set of pixels in $\mathcal{P}_k$ which is detected as occluded region. Using this, occluded region is represented as $O(\mathcal{P}_k) \neq \emptyset$. $O(\mathcal{P}_k, \mathcal{P}_l)$ denotes a set of pixels in their overlapping region (dark gray region in Fig. 2 (b)).

Given a 2D grid graph $G^{rec} = (\mathcal{P}, E')$,

$$p(\mathcal{P}|\mathcal{P}^g) = \frac{1}{Z_\mathcal{P}} \exp\{-U(\mathcal{P}|\mathcal{P}^g)\}, \tag{22}$$

where

$$U(\mathcal{P}|\mathcal{P}^g) = \sum_{\mathcal{P}_k \in \mathcal{P}} V_k^{rec}(\mathcal{P}_k|\mathcal{P}_k^g) + w_\mathcal{P} \sum_{(k,l)\in E'} V_{k,l}^{rec}(\mathcal{P}_k, \mathcal{P}_l). \tag{23}$$

Basically, $V_i^{rec}(\mathcal{P}_k|\mathcal{P}_k^g)$ corresponds to the similarity between $\mathcal{P}_k$ and $\mathcal{P}_k^g$. However, two different label sets are used for two different region: $\mathcal{P}^{tr(\cdot)}$ (patches from training images) for occluded region, $\mathcal{P}^g$ for non-occluded region. And for each region, selecting a label from the opposite label set should be prohibited, so this factor is also considered in the node potential using the minimum and maximum pixel value, 0 and MAX_S, 256 for 8-bit grayscale image,

respectively. $\mathcal{P}^{tr(\cdot)}$ is candidate labels only for occluded region:
$V_k^{rec}(\mathcal{P}_k = \mathcal{P}_k^{tr(j)}|\mathcal{P}_k^g)$

$$= \begin{cases} \frac{1}{|O(\mathcal{P}_k)|} \sum_{s_i \in O(\mathcal{P}_k)} |s_i^g - s_i^{tr(j)}|^2, & O(\mathcal{P}_k) \neq \emptyset, \\ MAX_S^2, & \text{otherwise}, \end{cases} \tag{24}$$

where $\mathcal{P}_k^{tr(j)}$ is a k-th patch of the j-th face image in the training set and $s_i^{tr(j)}$ is the i-th pixel of that image. Similarly, $\mathcal{P}^g$ is a candidate only for non-occluded region:

$$V_k^{rec}(\mathcal{P}_k = \mathcal{P}_k^g|\mathcal{P}_k^g) = \begin{cases} MAX_S^2, & O(\mathcal{P}_k) \neq \emptyset \\ 0, & \text{otherwise}. \end{cases} \tag{25}$$

$V_{k,l}^{rec}(\mathcal{P}_k, \mathcal{P}_l)$ models the spatial continuity of patches, and it can be represented using $O(\mathcal{P}_k, \mathcal{P}_l)$:

$$V_{k,l}^{rec}(\mathcal{P}_k, \mathcal{P}_l) = \frac{1}{|O(\mathcal{P}_k, \mathcal{P}_l)|} \sum_{s_i \in O(\mathcal{P}_k, \mathcal{P}_l)} |s_i(k) - s_i(l)|^2, \tag{26}$$

where $s_i(k)$ is s_i in $\mathcal{P}_k$.

Now, the size of the label set is more than 2: the number of images in the training set + 1. In general, solving this multiple labels problem is NP-hard [14]. Thus, we use the approximation algorithm based on the graph cut method called α-expansion [3]. For implementation, we use the publicly available C++ codes made by [3, 12].

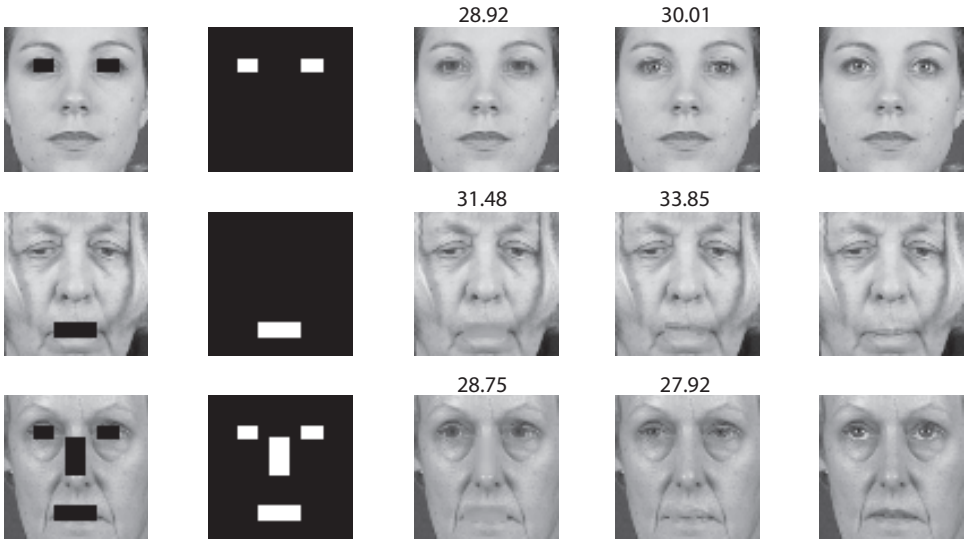

Figure 3: Some results of the occluded face image recovery algorithms for synthetic occlusion case. The first row: eyes occlusion, the second row: mouth occlusion, the third row: combined occlusion of eyes, nose, and mouth. (a) The occluded image, (b) the occlusion mask, (c) the recovery result of Lin and Tang's method [8], (d) the result of our method, and (e) the ground truth of non-occluded image. A score means the PSNR value between the recovery result and the ground truth image.

Table 1: The averaged PSNR results (dB) of the occluded face image recovery algorithms. For each different types of occlusion, eyes, nose, mouth, and combined occlusion of previous three occlusions, a quarter of the number of images are randomly selected from the test set and artificial occlusion is added. The value in $(\cdot)$ is the standard deviation.

Method	eyes	nose	mouth	combined
Lin and Tang [8]	31.12 ($\pm$ 2.02)	35.65 ($\pm$ 3.51)	32.85 ($\pm$ 2.50)	27.57 ($\pm$ 1.43)
Ours	31.08 ($\pm$ 2.00)	37.09 ($\pm$ 3.16)	34.59 ($\pm$ 3.05)	28.40 ($\pm$ 1.50)

4 Numerical Experiments

Our method is compared to the other MRF based method [8], which is the state-of-the-art method of the occluded face image recovery problem to the best of our knowledge. Basically, the main contribution of our algorithm is the reduced computation time in the recovery process, and changes in the detection process is minor. Thus, we only implement the recovery part of [8]. We implement their Confidence-oriented sampling method using MATLAB according to the algorithm and parameters in their paper. All experiments are tested on an Intel Core2Quad Q9400 2.66GHz computer with 8GB main memory.

In [8], they propose the iterative procedure of detection and recovery to refine the detection result iteratively. In our experiments, exact occlusion masks are used for synthetic occlusion case, and our detection algorithm's results are used for both recovery methods in the real occlusion tests. Since we start with same detection results, running their recovery algorithm only once is reasonable choice.

For the fair comparison, we fix the parameters of our method. Because the detection results are shared, we vary w_b in (12) and c in (13) until the detection algorithm correctly finds the occluded region in the real occlusion experiments. In the recovery process, there are three parameters: the number of hidden variables q in PPCA model, patch size r, and weight $w_{\mathcal{P}}$ in (23). We choose q which preserves 99.9% of the energy of the space using the magnitude of the eigenvalues, and use same q for PPCA model in [8]. $w_{\mathcal{P}}$ doesn't seriously affect the results, so we set to 1. For the patch size, we use 3×3 for the synthetic occlusion case and 2×2 for the real occlusion case.

4.1 Recovery with Synthetic Occlusion

4.1.1 XM2VTS database

We use the frontal face image set of XM2VTS database [10]. The dataset consists of two sets, where one set has 1180 color images (295 people $\times$ 4 images) at resolution 720×576, and the other set has 1180 images for same people but take shots on another day. We use one set for the training set, the other for the test set. Using the eyes, nose, and mouth position information available in XM2VTS database web-site[1], we make the cropped image of each images, which focuses on the face and has similar eyes, nose, and mouth position with the other cropped images. Finally, we convert each cropped image to a 64×64 grayscale image.

We consider four different occlusion cases: eyes, nose, mouth, and combined occlusion of previous three. We randomly split 1180 test images into 4 sets, where each set has 295 images. After adding different occlusions for each sets, the recovery algorithms are applied with the correct occlusion mask.

[1] http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

4.1.2 Qualitative evaluation

Some recovery results are shown in Fig. 3. Comparing Lin and Tang's method [8] (c) and ours (d) with the ground truth (e), ours produces more realistic recovery results even if recovered eyes are not much similar to his original eyes; e.g. the fourth image of the third row in Fig. 3.

However, the boundary region between occluded region and non-occluded region is not connected naturally. This effect can be caused by our assign rule for the chosen patches on the boundary region. On the boundary, the recovery algorithm selects one of the patches in the training set. Then, for the pixels in the occluded region, we assign pixel values of the selected patch, but for the pixels in the non-occluded region, we just assign values of the input image $\boldsymbol{o}$ via $\boldsymbol{s}^g$. As a consequence, the final recovery result has discontinuities on these boundaries.

4.1.3 Quantitative evaluation

To measure the recovery quality, the peak signal-to-noise ratio (PSNR) is used: Assume that all our images are 8-bit grayscale images. Given the recovered image $\boldsymbol{s}^*$ and the ground truth image $\boldsymbol{s}$, PSNR is defined as

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_I^2}{MSE}), \tag{27}$$

where MAX_I is the maximum possible pixel value of the image, 255 in our case, MSE is the mean squared error between two images defined as

$$MSE = \frac{1}{m}\sum_{i=0}^{m-1}(s_i - s_i^*)^2. \tag{28}$$

The PSNR comparison results are shown in Table 1. According to the table, ours gets the slightly higher PSNR values on average for the nose, mouth, and combined occlusion cases, but produces the lower value for the eyes occlusion. However, their differences are within 1.8 dB, and all standard deviation values are larger than 1.4 dB. Therefore, we can't say ours is better, but can say two methods are comparable.

We also compare the execution time of the algorithms. For every 100 additional occluded pixels, we measure the running time of two methods (Fig. 4). According to the results, the execution time of ours are almost constant compared with Lin and Tang's method [8]. Interestingly, ours sometimes terminates earlier even if occluded region is larger, where [8] monotonically increases against occluded region size. These are caused by the use of the different inference algorithms. Our one is the $\alpha-$expansion algorithm, which iteratively minimizes the total energy function until it converges, so the running time depends on the energy configuration of the graph. However, [8] uses greedy search algorithm using heuristically defined confidence information of pixels, where the algorithm visits each pixel in occluded region exactly once, so the running time monotonically increases.

This significant speed up of the computation time results from our modeling. Usually, the computation time of inference increases proportional to the number of vertices in MRF and the maximum clique size. When we use $r \times r$ patches, m pixels image can be modeled with roughly m/r^2 vertices which is much smaller than vertices in the pixel-based model in [8]. In addition, the maximum clique size is 2 for ours, where $r \times r$ in [8]; they compute the edge potential of i-th pixel using $r \times r$ neighborhood pixels. Even if our model is much simpler, the PSNR results are comparable, so we can claim that our model is more efficient. Or the use of the different inference algorithms can make these results possible, because our one, $\alpha-$expansion algorithm, is known as one of the best approximation algorithm from the view point of energy minimization among the various inference algorithms [12].

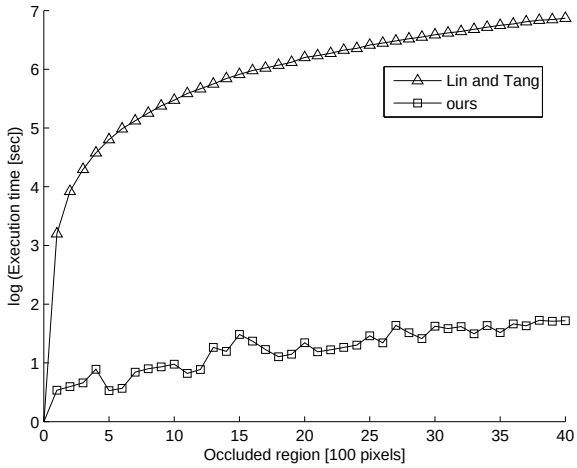

Figure 4: Execution time of the occluded face image recovery algorithms against variable occluded region size.

Figure 5: The training face images.

(a) (b) (c) (d)

Figure 6: The results of the occluded face image recovery algorithms for realistic test images. (a) The occluded image, (b) the detection result, (c) the recovery result of Lin and Tang's method [8], and (d) the result of our method.

4.2 Recovery with Real Occlusion

Now, we consider more realistic case, and also check our detection algorithm works. In real situation, face alignment, adjust the position of eyes, nose, and mouth, may not be an easy problem, especially when the face is occluded. In [15], they propose an unsupervised face alignment algorithm using non-rigid mapping of 2D-mesh, and distribute some of their

results publicly (Fig. 5). Some of aligned results are non-realistic (left two cases), and the others are relatively well aligned (right two cases). We manually select 214 relatively well aligned face images, and use them as a training set. Among whole images including poorly aligned face images, a few of them already have the real occlusion, sunglasses, so we choose six of them as a test set. For the test set, since we don't have the ground truth of non-occluded image, only qualitative evaluation is possible.

Some of the recovery results are shown in Fig. 6. For the detection process, the algorithm can't find the perfect occluded region even if we test with all possible parameter settings. Among six test images, the detection process is completely fail in two images, detects small redundant region in two images ((b), the first row), and detects large redundant region in two images ((b), the second row). For the recovery, our results are smoother on the boundary region than [8], but ours are more blurred. Although the recovery results are not as good as we expect for an automatic system, we can say that using our method, fully automatic occluded face image detection and recovery is possible even if the face alignment result is imperfect.

5 Conclusions

We have addressed an automatic procedure for face occlusion detection and recovery. We have presented two MRF models, one for occlusion detection and the other for occlusion recovery: (1) pixel-based MRF was used to model the probability of binary mask involving presence of occlusion at each pixel, assuming average face (computed using all face images in the training set) as a guess of true non-occluded face; (2) patch-based nonparametric MRF was used to refine the occlusion recovery reconstructed by PPCA, exploiting similarity between adjacent image patches as well as similarity between observed image patches and reconstructed image patches by PPCA in location-wise manner. The occlusion detection process was a binary label problem, so the global optimum was able to be found in polynomial time. The occlusion recovery process was a multiple label problem, so the approximate inference algorithm, α-expansion, was used. Qualitative and quantitative evaluation of recovery methods demonstrated that our proposed MRF models gained faster computation over the existing method, while the quality of recovery performance was as good as the existing one.

References

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of ACM SIGGRAPH*, 2000.

[2] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, 12(8):882–889, 2003.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

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

[6] N. Komodakis and G. Tziritas. Image completion using global optimization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[7] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 2 edition, 2001.

[8] D. Lin and X. Tang. Quality-driven face occlusion detection and recovery. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[9] C. Liu, H. Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

[10] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Proceedings of the Second International Conference on Audio and Video-Based Biometric Person Authentification*, pages 965–966, 1999.

[11] J. Sun, L. Yuan, J. Jia, and H. Y. Shum. Image completion with structure propagation. In *Proceedings of ACM SIGGRAPH*, 2005.

[12] 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*, 28:1068–1080, 2007.

[13] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 61(3):611–622, 1999.

[14] O. Veksler. *Efficient Graph-Based Energy Minimization*. PhD thesis, Cornell University, 1999.

[15] J. Zhu, L. Van Gool, and S. C. H. Hoi. Unsupervised face alignment by robust nonrigid mapping. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1265–1272, 2009.

Acknowledgments

This work was supported in part by Converging Research Center Program funded by the Ministry of Education, Science, and Technology (No. 2011K000673), NIPA ITRC support program (NIPA-2011-C1090-1131-0009), and NRF World Class University Program (R31-10100).


POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교