Available online at www.sciencedirect.com
ScienceDirect

Pattern Recognition Letters 27 (2006) 1768–1775

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Hallucinating multiple occluded face images of different resolutions

Kui Jia *, Shaogang Gong

Department of Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK

Available online 19 April 2006

Abstract

Learning-based super-resolution has recently been proposed for enhancing human face images, known as "face hallucination". In this paper, we propose a novel algorithm to super-resolve face images given multiple partially occluded inputs at different lower resolutions. By integrating hierarchical patch-wise alignment and inter-frame constraints into a Bayesian framework, we can probabilistically align multiple input images at different resolutions and recursively infer the high-resolution face image. We address the problem of fusing partial imagery information through multiple frames and discuss the new algorithm's effectiveness when encountering occluded low-resolution face images. We show promising results compared to those of existing face hallucination methods from both simulated facial database and live video sequences.

Keywords: Super-resolution (Hallucination); Bayesian framework; Image alignment

1. Introduction

Super-resolution is a technique to generate a higher resolution image given a single or a set of multiple low-resolution input images. The computation requires the recovering of lost high-frequency information occurring during the image formation process. Super-resolution can be performed using two different approaches: reconstruction-based (Elad and Feuer, 1997; Irani and Peleg, 1991; Schulz and Stevenson, 1996; Hardie et al., 1997), and learning-based (Freeman and Pasztor, 1999; Baker and Kanade, 2000a; Capel and Zisserman, 2001; Liu et al., 2001; Dedeoglu et al., 2004; Sun et al., 2003). The reconstruction-based approach inherits limitations when the magnification factor increases. In this paper, we focus on learning-based super-resolution, when applied to the human face, also commonly known as "hallucination" (Baker and Kanade, 2000b).

Capel and Zisserman (2001) used eigenfaces from a training face database as model prior to constrain and super-resolve low-resolution face images. To further improve the performance, they divided the human face into six unrelated parts and applied PCA on them separately. Combined with a MAP estimator, they can recover the result from a high-resolution eigenface space. A similar method was proposed by Baker and Kanade (2000a). Rather than using the whole or parts of a face, they established the prior based on a set of training face images pixel by pixel using Gaussian, Laplacian and feature pyramids. Freeman and Pasztor (1999) took a different approach for learning-based super-resolution. Specifically, they tried to recover the lost high-frequency information from low-level image primitives, which were learnt from several general training images. They broke the images and scenes into a Markov network, and learned the parameters of the network from the training data. To find the best scene explanation given new image data, they applied belief propagation in the Markov network. A very similar image hallucination approach was also introduced in (Sun et al., 2003). They used the primal sketch as the prior to recover the smoothed high-frequency information. Liu et al. (2001) combined the PCA model-based approach and Freeman's image primitive technique. They developed a mixture model combing a global parametric model called "global

* Corresponding author. Fax: +44 20 89806533.
E-mail addresses: chrisjia@dcs.qmul.ac.uk (K. Jia), sgg@dcs.qmul.ac.uk (S. Gong).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.02.009

face image" carrying common facial properties, and a local nonparametric model called "local feature image" recording local individualities. The high-resolution face image was naturally a composition of both.

More recently learning-based techniques have also been extended to video. In (Bishop et al., 2003), a direct application of (Freeman and Pasztor, 1999) to video sequences was attempted, but severe video artifacts were found. As a remedy, an ad hoc solution was proposed. It consisted of re-using high-resolution solutions for achieving more coherent videos. In (Dedeoglu et al., 2004), the authors extended the work of (Baker and Kanade, 2000a) to super-resolve a single human face video, using different videos of the face of the same person as training data. By exploiting Bayesian framework and spatial-temporal constraints, they reported an extremely high face video magnification factor. However, all existing techniques have not addressed the problem of variable resolutions of partially occluded inputs often encountered in video.

In this paper, we extend the work in (Jia and Gong, 2005) and propose a novel algorithm to super-resolve face images given multiple partially occluded inputs at different lower resolutions. By integrating hierarchical patch-wise alignment and inter-frame constraints into a Bayesian framework, we can probabilistically align multiple inputs at different resolutions and recursively infer the high-resolution face image. We address the problem of fusing partial imagery information through multiple frames and discuss the new algorithm's effectiveness when encountering occluded low-resolution face images. We show promising results compared to those of existing face hallucination methods from both simulated facial database and live video sequences.

The paper is organized as follows. In Section 2, we define the problem of hallucinating multiple partially occluded face images at different lower resolutions, and present a novel algorithm to probabilistically both align and infer high-resolution face images from a Bayesian framework perspective. Section 3 extends the new algorithm to cope with occluded face images in super-resolution. Experimental results are presented in Section 4 before conclusions are drawn in Section 5.

2. Hallucinating multiple images of different resolutions

2.1. Problem definition

In a surveillance video, a sequence or some snapshots of a human face can be captured, where their resolutions are often too small and vary significantly over time. The images can also be partially occluded. Such conditions make the images less useful for automatic verification or identification. Existing techniques have not considered hallucinating a high-resolution face image under these conditions.

In this paper, we define the problem of face hallucination in video as how to super-resolve a face image with multiple partially occluded inputs of different resolutions. As shown in Fig. 1, low-resolution face images are automatically detected as patches within rectangular regions in a video. We are interested in developing an algorithm that take into account these multiple inputs of different resolutions in order to yield an optimal higher resolution image. Furthermore, we also wish to perform super-resolution when some or all of the low-resolution inputs are occluded in the face detection process. An example of a hallucinated complete face image from multiple occluded inputs is also shown in Fig. 2.

The underlying challenges we aim to address are therefore three-fold. Firstly, how to align multiple inputs at different lower resolutions. Secondly, how to cross-refer and recover missing pixel information in different image frames due to occlusion. And finally, a unified algorithm to perform aligning, inferring missing information and super-resolving a high-resolution face image given multiple sources.

2.2. A Bayesian formulation

In this section, we formulate our problem of hallucinating multiple inputs of different resolutions by means of a Bayesian framework. Assuming H is the high-resolution image needs to be constructed, $L_1, L_2, \ldots, L_S$ are the low-resolution inputs with different resolutions. The task comes as finding the Maximum A Posterior (MAP) estimation of H given $L_1, L_2, \ldots, L_S$. Let us first consider the problem of only two low-resolution inputs L_1, L_2 (see Fig. 2), which can be formulated as

$$H_{\mathrm{MAP}} = \arg\max_{H} \log P(H|L_1, L_2) \tag{1}$$

Furthermore, we define T as an unknown intermediate template and I as the aligning parameter between low-resolution inputs L_1 and L_2 (how to determine parameter I and generate template T will be presented in Section 2.3). We can marginalize $P(H|L_1, L_2)$ over these unknown parameters as

$$P(H|L_1, L_2) = \sum_i \sum_j P(H, T_i, I_j | L_1, L_2)$$

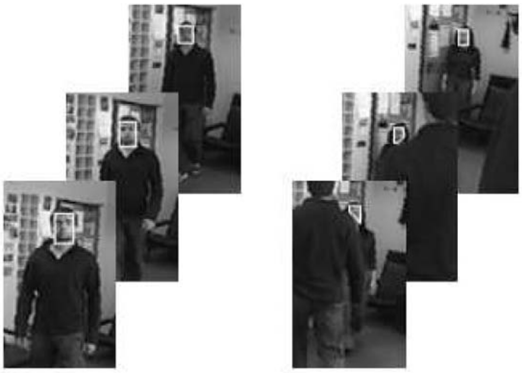

Fig. 1. A realistic face detection environment.

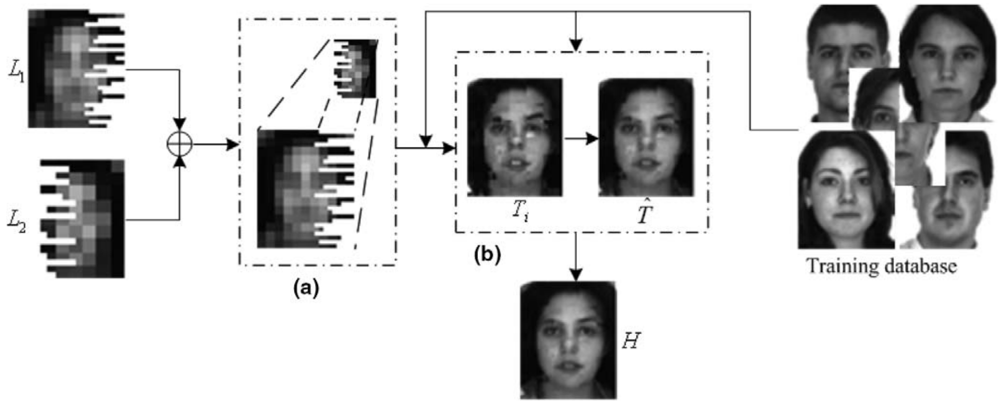

Fig. 2. An illustration of our hallucination process given multiple occluded input images: L_1 and L_2 are occluded low-resolution inputs, T_i and $\widehat{T}$ are intermediate templates, H is the final hallucination result of a higher resolution: (a) is the hierarchical image aligning process, and (b) is the process of patch learning and inter-frame constraint for estimating optimal intermediate template $\widehat{T}$.

where i and j are possible choices for T and I respectively. By applying the Bayes rule twice, the above becomes

$$\begin{aligned}&\sum_i \sum_j P(H|I_j,T_i,L_1,L_2)P(I_j,T_i|L_1,L_2)\\&\quad=\sum_i \sum_j P(H|T_i,I_j,L_1,L_2)P(T_i|I_j,L_1,L_2)P(I_j|L_1,L_2)\end{aligned}\tag{2}$$

Assuming aligning parameter I will peak at the true value $I^\star$, gives

$$P(I|L_1,L_2)=\delta(I-I^\star)\tag{3}$$

By substituting (3) into (2) and using the Bayesian rule, we have

$$\begin{aligned}&\sum_i P(H|I^\star,T_i,L_1,L_2)P(T_i|I^\star,L_1,L_2)\\&\quad=\sum_i \frac{P(L_1,L_2|H,I^\star,T_i)P(H|I^\star,T_i)}{P(L_1,L_2|I^\star,T_i)}P(T_i|I^\star,L_1,L_2)\end{aligned}\tag{4}$$

Assuming H exists and based on the basic image observation model, the low-resolution inputs can be independently sub-sampled from H, then we have $P(L_1,L_2|H,I^\star,T_i)=P(L_1|H)P(L_2|H)$. By setting the denominator as a constant C, $P(H|L_1,L_2)$ can be rewritten as

$$C\sum_i P(L_1|H)P(L_2|H)P(H|I^\star,T_i)P(T_i|I^\star,L_1,L_2)\tag{5}$$

Although there could be many options for the intermediate template T_i, the one which is optimal maximizes the probability $P(T_i|I^\star,L_1,L_2)$. We define it as $\widehat{T}$, and compute template $\widehat{T}$ by means of hierarchical low-level vision, similar to that of (Baker and Kanade, 2000a; Dedeoglu et al., 2004) (more details in Section 2.3). Then with (1) and (5), we maximize the following cost function for H_{MAP}:

$$\begin{aligned}&\log P(L_1|H)+\log P(L_2|H)+\log P(H|I^\star,\widehat{T})\\&\quad+\log P(\widehat{T}|I^\star,L_1,L_2)\end{aligned}\tag{6}$$

This resulting cost function is easily generalized from 2 to S inputs as follows:

$$\begin{aligned}&\log P(L_1|H)+\log P(L_2|H)+\cdots+\log P(L_S|H)\\&\quad+\log P(H|I_1^\star,I_2^\star,\ldots,I_{S-1}^\star,\widehat{T})\\&\quad+\log P(\widehat{T}|I_1^\star,I_2^\star,\ldots,I_{S-1}^\star,L_1,L_2,\ldots,L_S)\end{aligned}\tag{7}$$

where $I_1^\star,I_2^\star,\ldots,I_{S-1}^\star$ are aligning parameters for $S-1$ low-resolution inputs with respect to the largest resolution.

The individual components in (7) can be interpreted as follows. The first S terms require the inferred high-resolution result H to satisfy the basic image observation model with respect to each of input L. The penultimate term assures $\widehat{T}$ to serve as a prior in this Bayesian framework. Finally, the last term provides an entrance to compute $\widehat{T}$.

2.3. Finding the intermediate template

The basic idea for finding the intermediate template comes from (Dedeoglu et al., 2004). As in (6), to find the best $\widehat{T}$, we need to maximize the probability $P(\widehat{T}|I^\star,L_1,L_2)$. By the Bayes rule we have

$$P(\widehat{T}|I^\star,L_1,L_2)\propto P(L_1,L_2|\widehat{T},I^\star)P(\widehat{T})$$

Assuming L_1 is the low-resolution input that aligning is based on, $I^\star$ defines the hierarchical patch-wise correspondence between L_1 and L_2, we factorize the low-resolution inputs into independent patches. The above likelihood can be derived as

$$\prod_{p=1}^{N}\left(\sum_{q=1}^{M}P(L_p^1,L_q^2|\widehat{T}_p,I^\star)P(\widehat{T})\right)$$

where L_p^1, L_q^2 refer to the local patches in L_1 and L_2, N and M are their patch numbers respectively. Regarding each patch p for L_1, there is only one matching q from 1 to M. Assuming $I^\star$ is known, we have the final likelihood function as

$$\prod_{p=1}^{N} P(L_p^1, L_p^2|\widehat{T}_p)P(\widehat{T}) = \prod_{p=1}^{N} P(L_p^1|\widehat{T}_p)P(L_p^2|\widehat{T}_p)P(\widehat{T}) \tag{8}$$

where the L_p^2 stands for the hierarchically corresponding patch in L_2 with regard to L_p^1. In other words, a patch in L_2 corresponds to multiple patches in L_1. This expression can also be generalized from two to S low-resolution inputs, and the likelihood expression becomes

$$\prod_{p=1}^{N} P(L_p^1|\widehat{T}_p)P(L_p^2|\widehat{T}_p)\cdots P(L_p^S|\widehat{T}_p)P(\widehat{T}) \tag{9}$$

The first $S \times N$ terms in (9) give the basic idea for how to generate the intermediate template from the hierarchical patch matching perspective. But their constraints are still too weak considering each low-resolution patch could be generated from many high-resolution database patches. One remedy to this problem is to pool contextual information among patches. To this end, we used parent vector (DeBonet and Viola, 1998) as a local feature structure to strengthen these constraints. The prior $P(\widehat{T})$ finally provides a spatial dependency constraint to refine the generated template.

2.3.1. Aligning multiple low-resolution inputs

For determining the aligning parameter $I^{\star}$, let us first consider the two inputs L_1 and L_2 case again. Assuming L_1 is the low-resolution input that aligning is based on, we sub-sample L_1 to the resolution of L_2, and it becomes $\overline{L}_1$. To compute the aligning parameter $I^{\star}$, we need to maximize the likelihood function $P(I|\overline{L}_1, L_2)$, which is consistent with the likelihood function $P(I|L_1, L_2)$ in (2) and (3). Assume patches in $\overline{L}_1$ are mutually independent, by applying Bayes rule we attain

$$P(I|\overline{L}_1, L_2) = P(\overline{L}_1, L_2|I)P(I) = \prod_i P(\overline{L}_i^1, L_i^2|I)P(I) \tag{10}$$

where $\overline{L}_i^1$ and L_i^2 have similar meanings as L_p^1 and L_p^2 in (8). Given any aligning parameter estimation, we define the above probability density function as

$$P(\overline{L}_i^1, L_i^2|I) \propto \exp\left(-\|F_{\overline{L}_i^1} - F_{L_i^2}\|^2\right)$$

where $F_{\overline{L}_i^1}$ and $F_{L_i^2}$ are local patch feature vectors to be defined in Section 2.3.2. The value $I^{\star}$ that maximizes the cost function (10) gives the optimal aligning parameter. Similarly we can generalize the two input case to that of S inputs.

2.3.2. Template prior and local feature structure

The Markov Random Field (MRF) model assigns a probability to each template patch configuration T, and according to the Hammersley–Clifford theorem, $P(T)$ is a product $\prod_{T_m,T_n}\phi(T_m, T_n)$ of comparability function $\phi(T_m, T_n)$ over all pairs of neighboring patches. The details as how to compute $P(T)$ can be found in (Dedeoglu et al., 2004).

Suppose L_p^s is an image patch in low-resolution input L_s, and $\overline{T}_p$ is a random patch from the training database which has already been sub-sampled to the resolution of L_p^s. For each of these patches, we adopt the parent vector (DeBonet and Viola, 1998) as their feature vectors, which stacks together local intensity, gradient and Laplacian image values at multiple scales. To each of the term $P(L_p^s|\overline{T}_p)$, we define the probability density function as

$$P(L_p^s|\overline{T}_p) \propto \exp\left(-\|F_{L_p^s} - F_{\overline{T}_p}\|^2\right)$$

where $F_{L_p^s}$ and $F_{\overline{T}_p}$ are the feature vectors for L_p^s and $\overline{T}_p$. Similarly to (7), the pdf of the first $S \times N$ terms in (9) is generalized as

$$\prod_{p=1}^{N} P(L_p^1|\widehat{T}_p)P(L_p^2|\widehat{T}_p)\cdots P(L_p^S|\widehat{T}_p) \propto \prod_{p=1}^{N} \exp\left(-\sum_{s=1}^{S}\|F_{L_p^s} - F_{\overline{T}_p}\|^2\right) \tag{11}$$

The final intermediate template $\widehat{T}$ is estimated as

$$\arg\max_{T} \prod_{p=1}^{N} P(L_p^1, L_p^2, \ldots, L_p^S|T_p) \prod_{m,n} \phi(T_m, T_n) \tag{12}$$

2.4. Inferring the high-resolution image

After obtaining the intermediate template $\widehat{T}$, we can use the first $S + 1$ terms of (7) as objective function to infer the final result H.

Suppose the acquisition of $L_1, L_2, \ldots, L_S$ should observe the image observation model by blurring and sub-sampling the high-resolution H, we approximate the process as

$$L_s = A_s H + \eta_{L_s}$$

where $s = 1, \ldots, S$, A_s is a sub-sampling model, and η_{L_s} is Gaussian noise. Assuming any L_s is pixel-wise independent, then we have

$$P(L_s|H) = \prod_u \frac{1}{\sigma_{L_s}\sqrt{2\pi}} \exp\left(-\frac{(L_s(u) - (A_s H)(u))^2}{2\sigma_{L_s}^2}\right) \tag{13}$$

The final inference of H should be coherent with the intermediate template $\widehat{T}$ with a probability of $P(H|I^{\star}, \widehat{T})$. We express the relationship as

$$H = \widehat{T} + \eta_H$$

Assuming noise η_H is pixel-wise independent and Gaussian, we have

$$P(H|\widehat{T}, I^{\star}) = \prod_v \frac{1}{\sigma_H\sqrt{2\pi}} \exp\left(-\frac{(H(v) - \widehat{T}(v))^2}{2\sigma_H^2}\right) \tag{14}$$

Substitute (13) and (14) into the above objective function, we can finally infer high-resolution H by minimizing the following quadratic expression:

$$\frac{\sigma_H^2}{\sigma_{L_1}^2}\|L_1 - A_1H\|^2 + \frac{\sigma_H^2}{\sigma_{L_2}^2}\|L_2 - A_2H\|^2 + \cdots + \frac{\sigma_H^2}{\sigma_{L_S}^2}\|L_S - A_SH\|^2 + \|\widehat{T} - H\|^2 \qquad (15)$$

3. Hallucinating multiple occluded face images

Another significant advantage of the Bayesian framework presented above is its ability in recovering missing data from occluded low-resolution face images. This capability is important because occlusion is very common when capturing faces in video, as illustrated in Fig. 1.

Given occluded low-resolution inputs $L_1, L_2, \ldots, L_S$, the task here is to super-resolve the high-resolution H, even in the extreme case that none of the inputs captures a complete face. Within our Bayesian framework, we need first to estimate the aligning parameter $I^\star$, and then compute the intermediate template $\widehat{T}$. Given $\widehat{T}$, we can infer the final high-resolution H by minimizing (15).

Assuming L_1 and L_2 are two partially occluded images, even though not all patches from both images are present (i.e. partially missing) for alignment, a $I^\star$ can still be estimated by maximizing (10). Given $I^\star$, we can simplify (12) as

$$\arg\max_T \prod_{p_i} P(L^1_{p_i}|T_p) \prod_{p_j} P(L^2_{p_j}|T_p) \times \prod_{p_k} P(L^1_{p_k}|T_p)P(L^2_{p_k}|T_p) \prod_{m,n} \phi(T_m, T_n) \qquad (16)$$

from which $\widehat{T}$ can be generated, where p_i stands for the patches in L_1 without corresponding patches in L_2, p_j stands for the patches in L_2 without corresponding patches in L_1, and p_k are those patches that are common in both L_2 and L_2. The remaining process follows details in the above section. Fig. 2 illustrates the entire process for hallucinating occluded face images.

4. Experimental results

4.1. Setup

Our face images come from a subset of AR, FERET and Yale databases. The database consists of 845 images of 169 different individuals (60 women and 109 men), in which each person has five different face images. Originally face images from these databases have different sizes, and also the area of the image occupied by the face varies considerably. To build up a standard training patch database, we need to align these face images manually. This alignment was performed by hand marking the location of three points: the centers of the eyeballs and the lower tip of the nose. These three points define an affine warp, which was used to warp the images into a canonical form. The canonical image has 56×46 pixels with the right eye at $(25, 31)$, the left eye at $(25, 16)$, and the lower tip of the nose at $(34, 24)$.

In our current experiments, instead of testing our algorithm on automatically detected face images in live video, we generated the test images as follows. We first blurred any given high-resolution image from this database with different filters to introduce different Point Spread Functions (PSF) accordingly, and then sub-sampled the blurred images to low-resolutions. We then added random translational motion to introduce a measurable degree of random misalignment resulting from most automatic face detection processes on live video feed. For selecting high-resolution face images to generate testing data, we used "leave-one-out" methodology: For any series of generated testing low-resolution face images, we removed their corresponding high-resolution source from the database, and the remaining high-resolution images serve as the learning database. The removed high-resolution images later served as the ground truth images in the experiments on quantifying model error as shown in Fig. 5.

4.2. Comparison of single face images without missing parts

One advantage of our framework is its ability to deal with face hallucination with multiple inputs at different resolutions. To evaluate its effectiveness, for any given 56×46 image from the database, we generated three low-resolution images at the sizes of 14×11, 9×7 and 7×5 using the above method. Given these three testing face images, we took the largest 14×11 one as the low-resolution input that alignment is to be based upon, and estimated the aligning parameters for the 9×7 and 7×5 images. Then we generated the intermediate template based on (12). The high-resolution result was constructed by solving the quadratic cost function (15). Column (b) of Fig. 3 shows some example high-resolution results.

To compare these results with hallucination using a single face image similar to those in (Baker and Kanade, 2000a; Dedeoglu et al., 2004), we performed experiments by taking only simulated 14×11 images as low-resolution input; example results are shown in column (c) of Fig. 3. Comparing (b) and (c) in Fig. 3 suggests that we improved the hallucination results. However, the improvement is not dramatic because the largest low-resolution inputs already contain most of the information that could also be contributed from the other low-resolution inputs. In other words, the information from the other low-resolution inputs is mostly redundant.

4.3. Comparison of occluded face images

Significantly, the advantage of our multiple input based approach over existing hallucination methods becomes dramatic when low-resolution input images are partially occluded with missing parts. Such input images are common when detecting and tracking face images of moving targets in live video. In other words, if many of the low-resolution inputs at different resolutions miss pixels due to

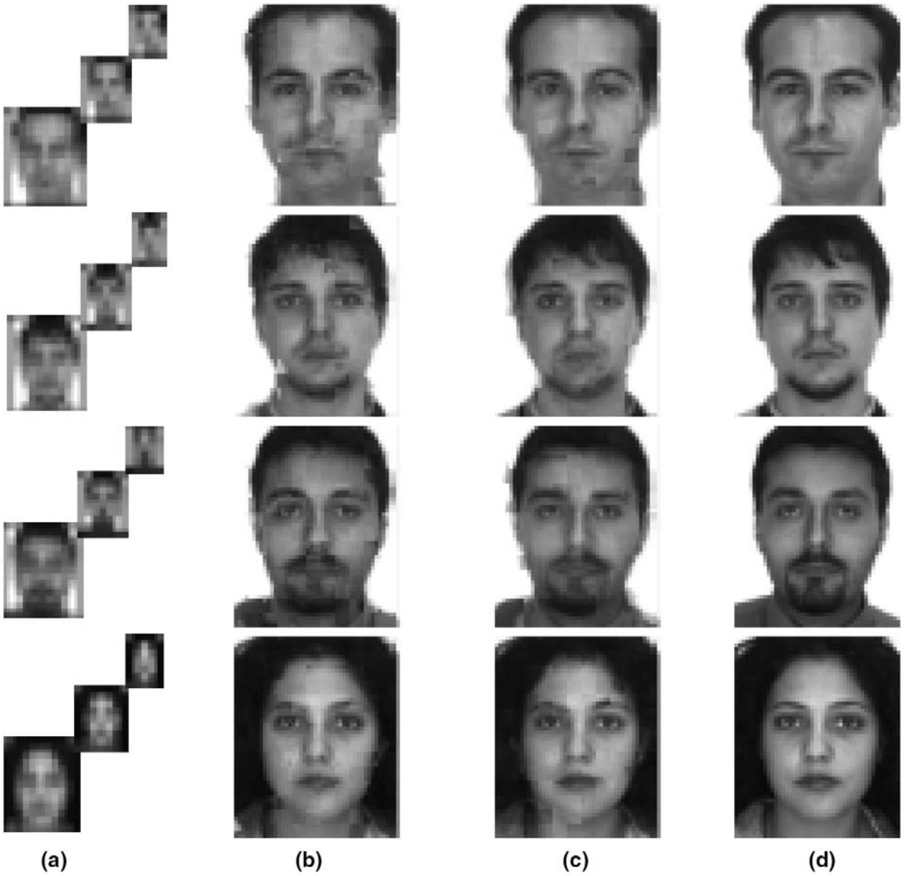

Fig. 3. Comparing face hallucination using single and multiple inputs without occlusion or missing parts: (a) multiple low-resolution inputs with frame resolution of 14 × 11, 9 × 7 and 7 × 5, (b) results from our approach, (c) results using 14 × 11 single image face hallucination and (d) ground truth images with resolution of 56 × 46.

occlusion (or pool lighting and viewpoint), it becomes essential to align them before super-resolving a high-resolution image takes place. Different from early experimental settings, given any 56 × 46 image in this experiment, we first randomly removed part of it to simulate the face image being partially occluded, and then generated the first occluded test image with the frame resolution of 14 × 11. By the same token we could yield another test image with frame resolution of 7 × 5. Some examples are shown in column (a) of Fig. 4.

Based on the deduced objective function (16) in Section 3, combined with Eqs. (10) and (15) in Section 2, we can probabilistically infer a high-resolution reconstruction making use of all the information from the two occluded test inputs. With existing learning-based super-resolution techniques, neither of two partially occluded low-resolution input images can provide sufficient information for recovering a complete face image at a higher resolution. Fig. 4 shows example results using single-image face hallucination technique similar in (Baker and Kanade, 2000a; Dedeoglu et al., 2004) given partially occluded low-resolution input images of 14 × 11 (b) and 7 × 5 (c) respectively. As expected, only part of a face was recovered at the higher resolution of 56 × 46. Furthermore, we show results in column (d) based on fusing the partially hallucinated face images from columns (b) and (c) of Fig. 4. It shows clearly that motion and illumination variations between different occluded input images at different lower resolutions make simple fusing a poor solution. On the other hand, our results shown in (e) improve significantly those of either (b) or (c) at the resolution of 56 × 46. It is also worth pointing out that given that our inputs were partially occluded with significant missing parts at the resolutions of 7 × 5 and 14 × 11, our magnification factor is effectively over 8 × 8 which goes beyond the existing 4 × 4 limit (to obtain a *desired* high-resolution result) for the current hallucination techniques.

To quantify the performance of different techniques, we measured the average root Sum of Squared Error (SSE) per pixel w.r.t. the original high-resolution image ground truth, as shown in Fig. 5. Consistent to Fig. 4, the average root SSE/pixel from our results (represented by the solid line) are the smallest compared to both those using the occluded 14 × 11 inputs (represented by the dotted line) and those using the occluded 7 × 5 inputs (represented by the dashed line). Fig. 5 also suggests that the results based on fusing the partially hallucinated parts by pixel averaging (represented by the dash-dot line) are much worse than our results. To explain this, we should notice that, although the partial face images in columns (b) and (c) of Fig. 4 (corresponding to the dotted and dashed lines in Fig. 5) are

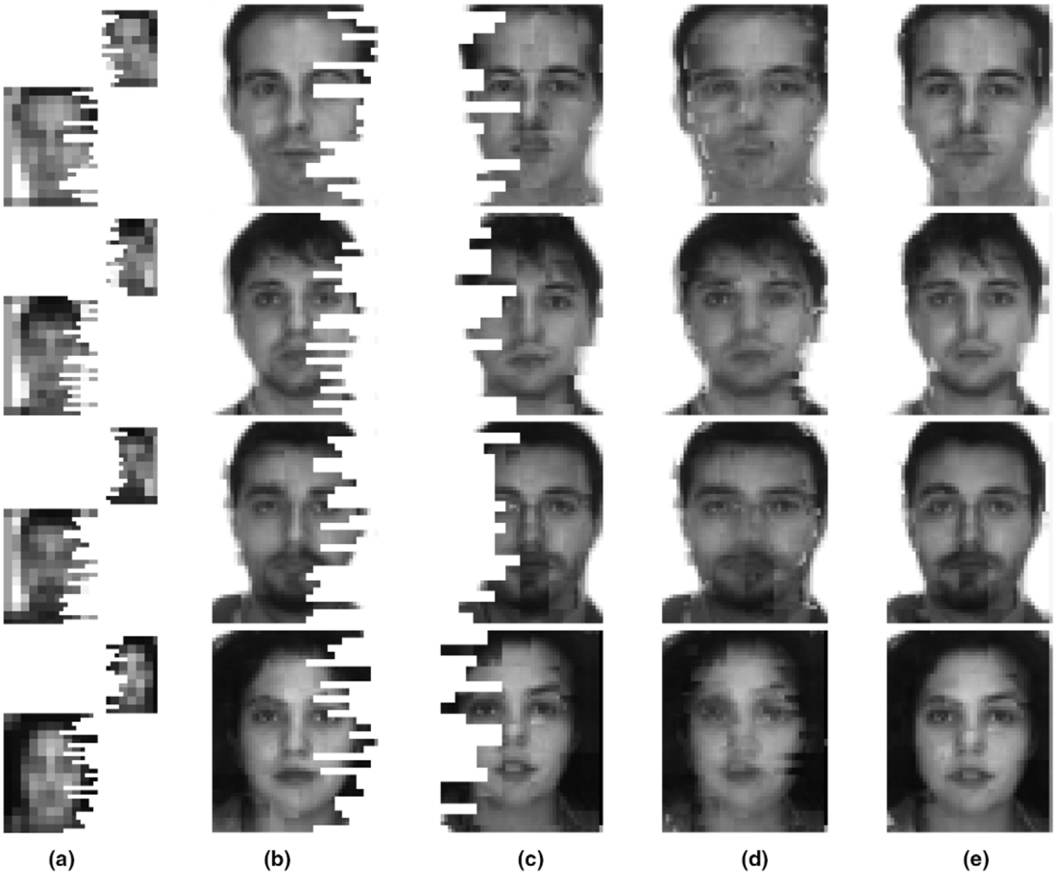

Fig. 4. Hallucination with occluded faces: (a) occluded face images with resolution of 14 × 11 and 7 × 5, (b) results using the occluded 14 × 11 input only, (c) results using the occluded 7 × 5 input only, (d) results based on fusing the partial face images in column (b) and (c) by pixel averaging at overlapped parts, (e) our hallucination results. Ground truth images are the same as in column (d) of Fig. 3.

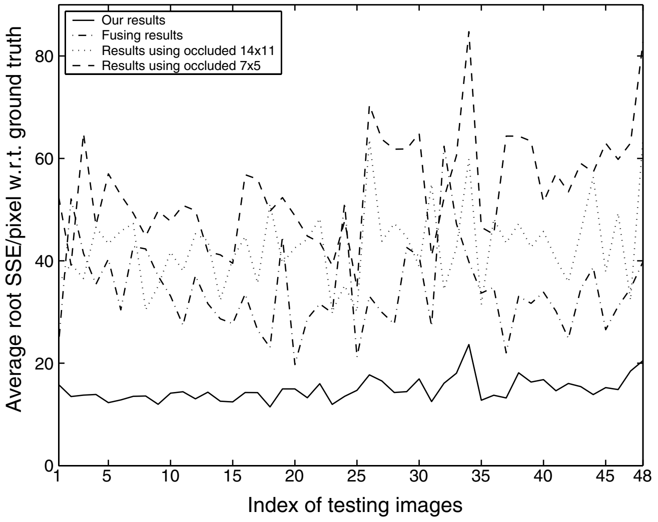

Fig. 5. Average root Sum of Squared Error (SSE) per pixel w.r.t. ground truth of hallucination results. The solid line represents error from our hallucination results, the dotted line represents error from partially hallucinated parts using occluded 14 × 11 inputs, the dashed line represents error from partially hallucinated parts using occluded 7 × 5 inputs, and the dash-dot line shows error from fusing the partially hallucinated results using occluded 14 × 11 and 7 × 5 input images respectively by pixel averaging at overlapped parts.

already independently aligned into general face frames with reference to the training database, they are essentially pixel-wise uncorrelated. The occluded low-resolution inputs were respectively super-resolved into partial high-resolution face images without considering the motion and illumination variations between them. Indeed it is these variations at low-resolution that make aligning and fusing at high-resolution fail. Only by utilizing a hierarchical and recursive formulation of an intermediate template as proposed in our approach, we are able to align and super-resolve across occluded inputs of different resolutions.

4.4. Experiments on live video data

For testing the robustness of our approach when applied on real data, we captured video sequences from a corridor surveillance camera. Since it is unrealistic to assume that partially occluded face images could be accurately detected, especially in case of low-resolution, in this experiment we selected frames where complete faces at different lower resolutions existed, and manually cropped out these face images. After that we randomly removed partial faces

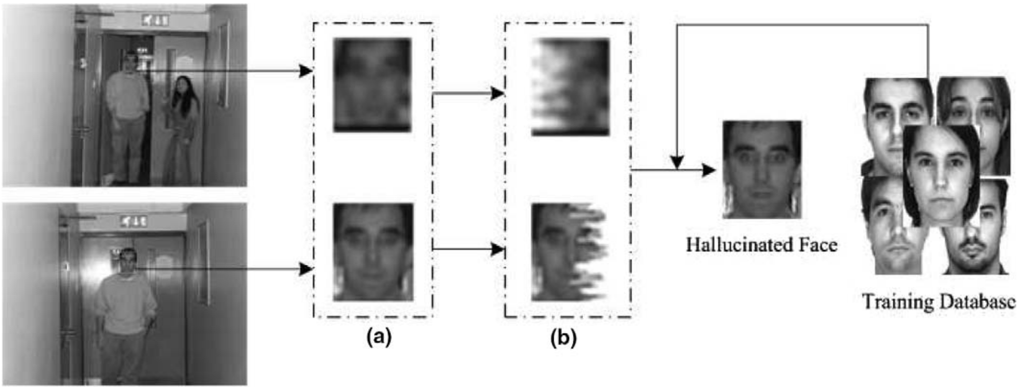

Fig. 6. Example of hallucinating multiple occluded face images using live video data: (a) low-resolution facial images in live video were located and segmented first and (b) partial faces were randomly removed to simulate the occlusion conditions.

and used the left parts as testing inputs. The high-resolution training images are from a subset of AR face database. Fig. 6 demonstrates the illustration example.

In the live sequence experiments, we have no ground truth images which could be collected, to verify the likeness of hallucinated results. But the resulting images do demonstrate the robustness of our approach. As suggested in Fig. 6, the quality of hallucinated face is as good as, if not better than, those in simulated database experiments.

5. Conclusion

In summary, by introducing an intermediate template recursively estimated into a Bayesian framework, we present a novel model to super-resolve face images with multiple occluded inputs at different lower resolutions. The model in essence performs hierarchical patch-wise alignment and global Bayesian inference. Beyond the classic face hallucination algorithms, we both consider the spatial constraints and exploit the inter-frame constraints across multiple face images of different resolutions. As a consequence, the new algorithm is more effective for dealing with occluded low-resolution face images. We showed significantly improved results over existing face hallucination methods.

In this work, we manually conducted experiments on live video sequences, in which the pose and illumination variations were solved by manual alignment and normalization. In the future we will extend our work on hallucinating automatically detected low-resolution face videos.

References

Baker, S., Kanade, T., 2000a. Limits on super-resolution and how to break them. In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition. South Carolina, vol. 2, pp. 372–379.

Baker, S., Kanade, T., 2000b. Hallucinating faces. In: Proc. IEEE Automatic Face and Gesture Recognition. Grenoble, France, pp. 83–90.

Bishop, C.M., Blake, A., Marthi, B., 2003. Super-resolution enhancement of video. In: Proc Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics. Key West, Florida, USA.

Capel, D.P., Zisserman, A., 2001. Super-resolution from multiple views using learnt image models. In: Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition. Kauai, HI, USA, vol. 2, pp. 627–634.

DeBonet, J.S., Viola, P.A., 1998. A non-parametric multi-scale statistical model for natural images. In: Advances in Neural Information Processing Systems (NIPS). Denver, USA, vol. 10, pp. 773–779.

Dedeoglu, G., Kanade, T., August, J., 2004. High-zoom video hallucination by exploiting spatio-temporal regularities. In: Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition. Washington, DC, USA, vol. 2, pp. 151–158.

Elad, M., Feuer, A., 1997. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. IEEE Trans. on Image Processing 6 (12), 1646–1658.

Freeman, W., Pasztor, E., 1999. Learning low-level vision. In: 7th Internat. Conf. on Computer Vision. Kerkyra, Greece, pp. 1182–1189.

Hardie, R.C., Barnard, K.J., Armstrong, E.E., 1997. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. IEEE Trans. Image Process. 6, 1621–1633.

Irani, M., Peleg, S., 1991. Improving resolution by image registration. CVGIP: Graphical Models Image Proc. 53, 231–239.

Jia, K., Gong, S., 2005. CCTV face hallucination under occlusion with motion blur. In: Proc. IEE Internat. Symposium on Imaging for Crime Detection and Prevention. London, UK, pp. 85–88.

Liu, C., Shum, H., Zhang, C., 2001. A two-step approach to hallucinating faces: Global parametric model and local nonparametric model. In: Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition. Kauai, HI, USA, pp. 192–198.

Schulz, R.R., Stevenson, R.L., 1996. Extraction of high-resolution frames from video sequences. IEEE Trans. Image Process. 5, 996–1011.

Sun, J., Zhang, N., Tao, H., Shum, H., 2003. Image hallucination with primal sketch priors. In: Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition. Madison, USA, vol. 2, pp. 729–736.