In Proc. IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS'05), September 2005

Hallucinating Multiple Occluded CCTV Face Images of Different Resolutions

Kui Jia

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

Shaogang Gong

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

Abstract

In this paper, we present a novel learning-based algorithm to super-resolve multiple partially occluded CCTV low-resolution face images. 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 that of existing face hallucination methods.

1. Introduction

Super-resolution aims 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. In this paper, we focus on learning-based super-resolution, when applied to the human face, also commonly known as "hallucination"[3].

Capel and Zisserman [5] used eigenface from a training face database as model prior to constrain and super-resolve low-resolution face images. A similar method was proposed by Baker and Kanade [2], they established the prior based on a set of training face images pixel by pixel using Gaussian, Laplacian and feature pyramids. Freeman and Pasztor [4] 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. Liu and Shum [6] combined the PCA model-based approach and Freeman's image primitive technique. In [8], the authors extended the work of [2] to super-resolve a single human face video, using different videos of the face of the same person as training data. However, all existing techniques have not addressed the problem of variable resolutions of partially occluded inputs often encountered in video.

Figure 1: Low resolution and partially occluded face image patches detected in realistic CCTV video.

In a surveillant 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. Fig. 1 shows low-resolution face image patches of different sizes automatically detected in a CCTV video. We wish to perform super-resolution when some or all of the low-resolution inputs are occluded in the face detection process as shown in Fig. 2. The underlying problems we aim to address are three folds: (1) how to align multiple inputs at different lower resolutions, (2) how to cross-refer and recover missing pixels due to occlusion, and (3) a unified algorithm to perform alignment and super-resolution of multiple low-resolution inputs.

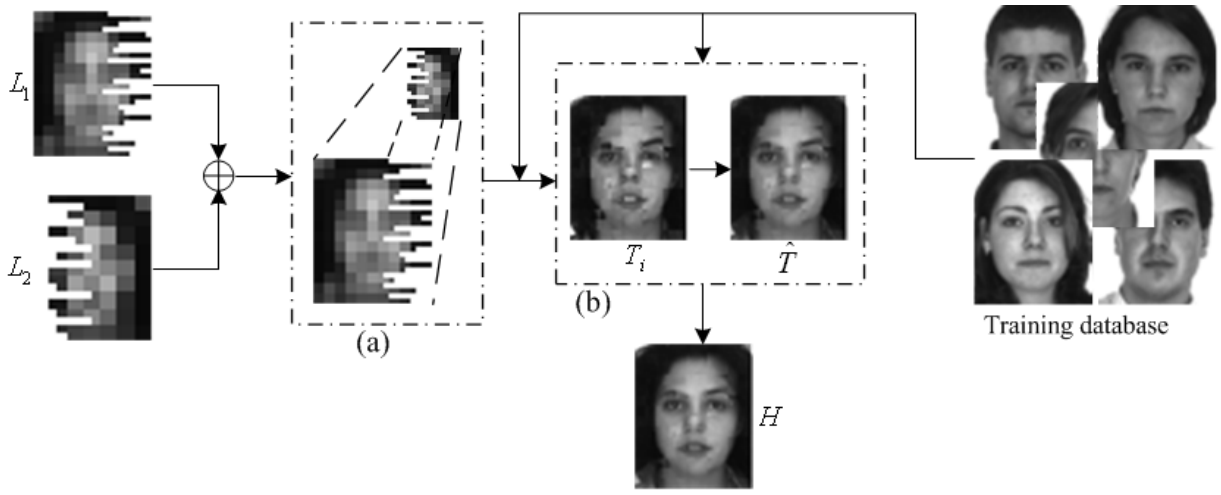

Figure 2: An illustration of our CCTV face image super-resolution process: L_1 and L_2 are occluded low-resolution inputs, T_i and $\hat{T}$ are intermediate templates, H is the final hallucination result of a higher resolution. (a) is the hierarchical image aligning process, and (b) is the process of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

2 CCTV Face Images Super Resolution

We formulate our problem of face image super-resolution of multiple low-resolution CCTV inputs of variable sizes 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:

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

Furthermore, we define T as an unknown intermediate template and I as the aligning parameter between low-resolution inputs L_1 and L_2. We can marginalize $P(H|L_1, L_2)$ over these unknown parameters as:

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

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

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

Assuming aligning parameter I will peak at the true value $I^\star$, which gives $P(I|L_1, L_2) = \delta(I - I^\star)$. Using Bayesian rule we have

$$\begin{aligned} &\sum_i P(H|I^\star, T_i, L_1, L_2) P(T_i|I^\star, L_1, L_2) \\ &= \sum_i \frac{P(L_1, L_2|H, I^\star, T_i) P(H|I^\star, T_i)}{P(L_1, L_2|I^\star, T_i)} P(T_i|I^\star, L_1, L_2) \end{aligned} \tag{3}$$

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 denominator as a constant C, $P(H|L_1, L_2)$ can be rewritten as

$$C \sum_i P(L_1|H) P(L_2|H) P(H|I^\star, T_i) P(T_i|I^\star, L_1, L_2) \tag{4}$$

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 $\hat{T}$, and compute template $\hat{T}$ by means of hierarchical low-level vision, similar to that of [2, 8]. Then with (1) and (4), we maximize the following cost function for H_{MAP}:

$$\begin{aligned} &\log P(L_1|H) + \log P(L_2|H) + \log P(H|I^\star, \hat{T}) \\ &\qquad + \log P(\hat{T}|I^\star, L_1, L_2) \end{aligned} \tag{5}$$

This resulting cost function is easily generalized from two to S inputs.

2.1 Finding the Intermediate Template

The basic idea for finding the intermediate template comes from [8]. As in (5), to find the best $\hat{T}$, we need to maximize the probability $P(\hat{T}|I^\star, L_1, L_2)$. By the Bayes rule we have

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

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:

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

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:

$$\prod_{p=1}^{N} P(L_p^1, L_p^2|\hat{T}_p)P(\hat{T}) = \prod_{p=1}^{N} P(L_p^1|\hat{T}_p)P(L_p^2|\hat{T}_p)P(\hat{T}) \tag{6}$$

where the L_p^2 stands for the hierarchically corresponding patch in L_2 with regard to L_p^1. This expression can also be easily generalized to S low-resolution inputs. The first $2N$ terms in right-hand side of (6) 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 [1] as local feature structure to strengthen these constraints. The prior $P(\hat{T})$ finally provides a spatial dependency constraint to refine the generated template.

Aligning Multiple Low-Resolution Inputs. For determining the aligning parameter $I^\star$, let us first consider the two inputs L_1 and L_2 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 $\bar{L_1}$. To compute the aligning parameter $I^\star$, we need to maximize the likelihood function $P(I|\bar{L_1}, L_2)$. Assume patches in $\bar{L_1}$ are mutually independent, by applying the Bayes rule we yield

$$P(I|\bar{L_1}, L_2) = P(\bar{L_1}, L_2|I)P(I) = \prod_i P(\bar{L_i^1}, L_i^2|I)P(I) \tag{7}$$

where $\bar{L_i^1}$ and L_i^2 have similar meanings as L_p^1 and L_p^2 in (6). Given any aligning parameter estimation, we define the above probability density function as

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

where $F_{\bar{L_i^1}}$ and $F_{L_i^2}$ are local patch feature vectors to be defined. The value $I^\star$ that maximizes the cost function (7) gives the optimal aligning parameter. Similarly we can generalize the two inputs case to that of S inputs.

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 [8].

Suppose L_p^s is an image patch in low-resolution input L_s, and $\bar{T_p}$ 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 [1] as their feature vectors, which stacks together local intensity, gradient and Laplacian image values at multiple scales. To each of the term $P(L_p^s|\bar{T_p})$, we define the probability density function as

$$P(L_p^s|\bar{T_p}) \propto \exp\left(-\|F_{L_p^s} - F_{\bar{T_p}}\|^2\right)$$

where $F_{L_p^s}$ and $F_{\bar{T_p}}$ are the feature vectors for L_p^s and $\bar{T_p}$. The final intermediate template $\hat{T}$ is estimated as:

$$\arg\max_T \prod_{p=1}^{N} P(L_p^1, L_p^2, \ldots, L_p^S|T_p) \prod_{m,n}\phi(T_m, T_n) \tag{8}$$

2.2 Inferring the High-Resolution Image

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

$$P(L_s|H) = \prod_u \frac{1}{\sigma_{L_s}\sqrt{2\pi}} \exp\left(-\frac{(L_s(u) - (A_s H)(u))^2}{2\sigma_{L_s}^2}\right) \tag{9}$$

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

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

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

$$P(H|\hat{T}, I^\star) = \prod_v \frac{1}{\sigma_H\sqrt{2\pi}} \exp\left(-\frac{(H(v)-\hat{T}(v))^2}{2\sigma_H^2}\right) \quad (10)$$

Substitute (9) and (10) into the above objective function, we can finally infer high-resolution H by minimizing the following quadratic expression

$$\frac{\sigma_H^2}{\sigma_{L_1}^2}\|L_1 - A_1H\|^2 + \frac{\sigma_H^2}{\sigma_{L_2}^2}\|L_2 - A_2H\|^2 + \cdots + \frac{\sigma_H^2}{\sigma_{L_S}^2}\|L_S - A_SH\|^2 + \|\hat{T} - H\|^2 \quad (11)$$

2.3 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. Given occluded low-resolution inputs $L_1, L_2, \ldots, L_S$, the task here is to super-resolve the high-resolution H, even at 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 $\hat{T}$. Given $\hat{T}$, we can infer the final high-resolution H by minimizing (11).

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 (7). Given $I^\star$, we can simplify (8) as

$$\arg\max_T \prod_{p_i} P(L_{p_i}^1|T_p) \prod_{p_j} P(L_{p_j}^2|T_p) \prod_{p_k} P(L_{p_k}^1|T_p)P(L_{p_k}^2|T_p) \prod_{m,n} \phi(T_m, T_n) \quad (12)$$

from which $\hat{T}$ 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.

3 Experimental Results

We built our face image database from a subset of AR, FERET and Yale databases. Our database consists of 845 images of 169 different individuals (60 women and 109 men), in which each person has 5 different face images. Originally face images from these databases have different sizes, and also the area of the image occupied by 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 3 points: the centers of the eyeballs and the lower tip of the nose. These 3 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 testing 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 process 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. Those removed high-resolution images are later served as the ground truth images in the experiments on quantifying model error shown in Fig. 5.

3.1 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 (8). The high-resolution result was constructed by solving the quadratic cost function (11). Column (b) of Fig.3 shows some example high-resolution results.

To compare these results with hallucination using a single face image similar as in [2, 8], we performed experiments by taking only simulated 14×11 image as low-resolution input, example results are shown in column (c) of Fig.3. Comparing (b) to (c) in Fig.3, it suggests that the improvement of our hallucination results is not dramatic. This is because the largest low-resolution inputs already contains

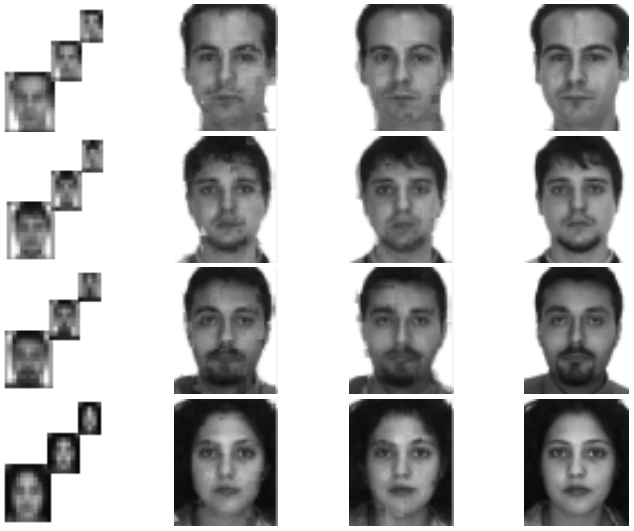

Figure 3: Comparing face hallucination using single and multiple inputs without occlusion or missing parts:(a) Multiple low-resolution inputs with frame resolution of 14×11, 9×7 and 7×5, (b) Results from our approach, (c) Results using 14×11 single image face hallucination, (d) Ground truth images with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

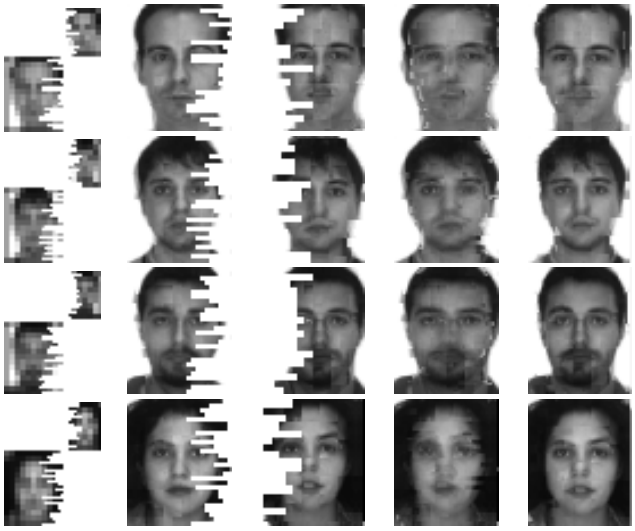

Figure 4: Hallucination with occluded faces: (a) Occluded face images with resolution of 14×11 and 7×5, (b) Results using the occluded 14×11 input only, (c) Results using the occluded 7×5 input only, (d) Results based on fusing the partial face images in column (b) and (c) by pixel averaging at overlapped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

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 are mostly redundant.

3.2 Comparison of Occluded Face Images

Significantly, the advantage of our multiple inputs 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 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 testing image with the frame resolution of 14×11. By the same token we could yield another testing image with frame resolution of 7×5. Some examples are shown in column (a) of Fig.4.

Based on the deduced objective function (12) in section 2.3, combined with equations (7) and (11), we can probabilistically infer a high-resolution reconstruction making

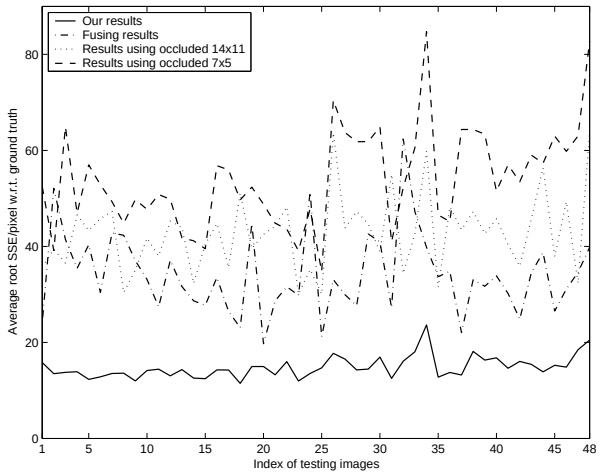

Figure 5: Average root Sum of Squared Error (SSE) per pixel w.r.t. ground truth of hallucination results. The solid line represents error from our hallucination results, the dotted line represents error from partially hallucinated parts using occluded 14×11 inputs, the dashed line represents error from 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.

use of all the information from the two occluded testing inputs. With existing learning-based super-resolution techniques, none of these 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 as in [2, 8] 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 performances 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 solid line) are the smallest compared to both those using the occluded 14×11 inputs (represented by dotted line) and those using the occluded 7×5 inputs (represented by dashed line). In Fig.5, it also suggests that the results based on fusing the partially hallucinated parts by pixel averaging (represented by 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 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 Conclusion

In summary, by introducing an intermediate template recursively estimated into a Bayesian framework, we present a novel model to super-resolve CCTV face images with multiple occluded inputs at different lower resolutions. The model in essence performs hierarchical patch-wise alignment and global Baysesian 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 have yet to conduct experiments on detected and tracked face images in live CCTV video, where face occlusions, motion between frames and illumination conditions may vary significantly. We did not consider pose variations either. In the future we will test our algorithm on those conditions. We will also extend our work on hallucinating automatically detected low-resolution face videos.

References

[1] J. S. DeBonet and P. A. Viola, "A non-parametric multi-scale statistical model for natural images", *Advances in Neural Information Processing Systems (NIPS)*, vol. 10, 1998.

[2] S. Baker and T. Kanade, "Limits on super-resolution and how to break them", *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*,June 2000.

[3] S. Baker and T. Kanade, "Hallucinnating Faces", *Proc. of IEEE Automatic Face and Gesture Recognition*, pp.83-90, March 2000

[4] W. Freeman and E. Pasztor, "Learning low-level vision", *7th International Conference on Computer Vision*, pp. 1182-1189,1999.

[5] D. P. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models", *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[6] C. Liu, H. Shum and C. Zhang, "A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model", *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp 192-198, 2001.

[7] C. M. Bishop, A. Blake, and B. Marthi, "Super-resolution enhancement of video", *Proceedings of Artificial Intelligence and Statistics*, Society for Artificial Intelligence and Statistics, 2003.

[8] G. Dedeoglu, T. Kanade, and J. August, "High-zoom video hallucination by exploiting spatio-temporal regularities", *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Vol.2, pp. 151-158, June 2004.