Super-resolution of facial images in forensics scenarios

Satiro, Joao; Nasrollahi, Kamal; Correia, Paulo; Moeslund, Thomas B.

Published in:
IEEE 5th International Conference on Image Processing Theory, Tools and Applications

DOI (link to publication from Publisher):
10.1109/IPTA.2015.7367096

Publication date:
2015

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Satiro, J., Nasrollahi, K., Correia, P., & Moeslund, T. B. (2015). Super-resolution of facial images in forensics scenarios. In IEEE 5th International Conference on Image Processing Theory, Tools and Applications (pp. 55-60). IEEE Signal Processing Society. DOI: 10.1109/IPTA.2015.7367096

Super-resolution of facial images in forensics scenarios

João Satiro[1], Kamal Nasrollahi[2], Paulo L. Correia[1], and Thomas B. Moeslund[2]
[1] Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal
e-mail: joao.satiro@tecnico.ulisboa.pt, plc@lx.it.pt
[2] Visual Analysis of People (VAP) laboratory, Aalborg University, Denmark
e-mail: {kn,tbm}@create.aau.dk

Abstract— **Forensics facial images are usually provided by surveillance cameras and are therefore of poor quality and resolution. Simple upsampling algorithms can not produce artifact-free higher resolution images from such low-resolution (LR) images. To deal with that, reconstruction-based super-resolution (SR) algorithms might be used. But, the problem with these algorithms is that they mostly require motion estimation between LR and low-quality images which is not always practical. To deal with this, we first simply interpolate the LR input images and then perform motion estimation. The estimated motion parameters are then used in a non-local mean-based SR algorithm to produce a higher quality image. This image is further fused with the interpolated version of the reference image via an alpha-blending approach. The experimental results on benchmark datasets and locally collected videos from surveillance cameras, show the outperformance of the proposed system over similar ones.**

Keywords— **super-resolution, reconstruction, forensics, facial images**

I. Introduction

The face, as one of the most common biometrics, is of great importance in many real-world applications, like human-computer interaction, human identification, access control, border control, to name a few. Facial images are of critical importance in forensics scenarios as well. The main difference when considering forensics scenarios is that facial images are taken by surveillance cameras, if there is any in the scene, and subjects of interest are not cooperative with the system nor are the imaging conditions as controlled. This makes it very challenging to work with facial images in forensics scenarios.

The problems with facial images in forensic scenarios are: 1) they are usually very small and 2) they are of poor quality. The former one is a result of the distance between surveillance cameras and subjects of interest while the latter can be a result of, among others, not facing the camera, facial expression, bad illumination, and blur [1]. These problems make it very hard to use such facial images especially in automatic systems, like face recognition. One solution for dealing with these problems (mainly the small sizes of the images) is to use upsampling algorithms, like interpolation methods. The problem, however, with simple interpolation approaches is that they can't produce artifact-free higher resolution images from lower resolution input images. To produce higher resolution images that are less affected by artifacts, super-resolution (SR) algorithms have been used.

SR algorithms are generally divided into two groups: 1) reconstruction-based, applied when multiple input images of the same person are available; and 2) hallucination-based, usually applied when there is a single input image [2]. In the hallucination-based approaches, [3], [4], [5], [6], and [7], to name a few, there is usually a training step in which the relationship between the low-resolution (LR) images (or their patches) and their high-resolution (HR) counterparts are learned. In the testing step, this learned relationship is then used to predict or hallucinate missing HR details of an input LR image. Though recent works of this group, based on deep learning, like [4], or on dictionary learning, like [6], produce images of good quality with improvement factors quite larger than two, they are not suitable for forensics applications for a critical reason: having an input LR image, the learning algorithm behind these systems actually teach them to hallucinate missing HR details [7]. This means, that if, for example, some skin texture is missing in the LR image it might be hallucinated based on the data used for learning. This may result in, for instance, an eventual loss of a characteristic birthmark. This might not be that problematic with general computer vision applications, but for forensics scenarios where legal issues are of top priority for law enforcement, such hallucination techniques will not be acceptable to the court of law. This rules out the single image-based hallucination algorithms and any other upsampling algorithms that somehow involve these algorithms, like [8] and [9], for forensics applications. This leaves us with the reconstruction-based SR algorithms.

The reconstruction-based SR algorithms, also known as multi-frame algorithms, [10], [11], [12], [13], [14], to name a few, usually take a set of LR images of the same scene (here human face) and try to utilize the differences between them to reconstruct missing HR details by reversing the steps involved in the imaging model. The differences between LR images can be of different forms, e.g., like sub-pixel misalignment, depth, etc. These differences are usually compensated for in a registration step, where all the inputs are registered to a common frame. The registered images are then fused together to produce a HR image. The problem with the reconstruction-based SR algorithms is that they cannot provide very large improvement factors [2]. Furthermore, most of them need to have a good registration algorithm to be able to find the motion (or generally the differences) between different LR images. This is very challenging in forensics scenarios as the LR images are really of poor quality and explicit estimation of the motion is very error prone. Therefore, in this paper, we have used a modified version of [15] which does not need an

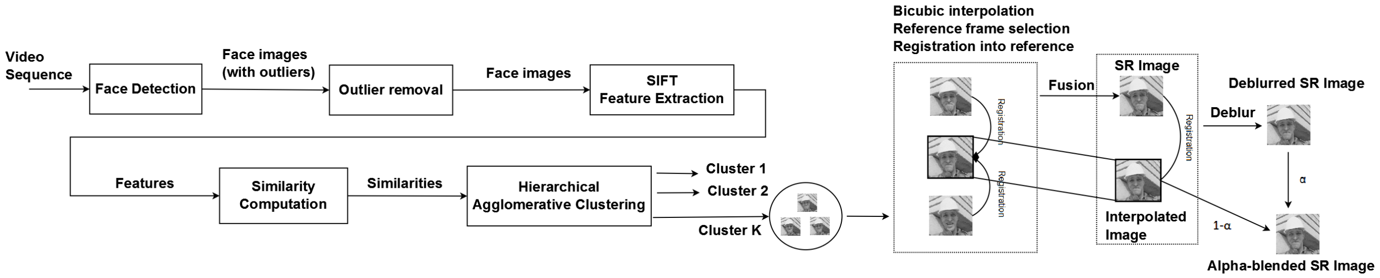

Fig. 1. The block diagram of the proposed system.

explicit motion estimation. We have shown that the proposed system can produce good quality results for typical images taken from surveillance cameras.

The rest of this paper is organized as follows: The proposed system is detailed in section II. Then, the experimental results are given in III. Finally, the paper is concluded in section IV.

II. The Proposed system

The block diagram of the proposed system is shown in Fig. 1. Following this diagram, faces are first detected from the input video sequence and then clustered based on some similarity measures. The clustered faces are then fed into a reconstruction-based SR algorithm. This algorithm combines information from different LR facial images into a HR one, which is of better quality than the interpolated LR input. Details of these steps are provided in the following subsections.

A. Face detection and clustering

For the present implementation the frontal face detection module of BioFoV[1] has been used which employs Haar like features of [16], [17]. This detector usually contains some outlier and background in the detected faces. To remove these, before clustering faces into groups a pre-processing step using skin-tone filtering is considered.

Outlier filtering is based on the assumption that human skin has a characteristic colour which may be distinguished from many other objects, although some will share the same colours, as those made from some types of wood. In addition, body parts other than faces will also pass the skin-tone filtering test, meaning that this filter cannot ascertain that a given image contains a face, but it is still able to exclude many of the outliers initially detected as faces. Explicit colour thresholding has been employed in the YCbCr colour space, following the analysis in [18], which has shown that when working with the HSV or YCbCr spaces, skin colour is concentrated in a small region of the space defined by the colour components. The selection of thresholds to be used was decided based on a set of tests performed on a database containing face and non-face images. The face images used for testing are from the Labeled Faces in the Wild database [19], which contains faces from people of different ethnicity acquired in unconstrained scenarios. Non-face images were collected from various Internet sites. From these preliminary tests, a selection of thresholds for skin-tone filtering was made by:

$$80 \leq Cb \leq 120, 133 \leq Cr \leq 173 \tag{1}$$

For the images selected as faces a set of features are extracted with the goal of discriminating faces of different users, or of the same user in different poses, and cluster them so as to provide input to the SR algorithm. SIFT features are used, since they are known for robustness to different scales, illuminations, orientations and affine transformations [20].

When computing the similarity between face images, we use the number of matching SIFT features. This global similarity measure is described in equation 2, where M_{ij} is the maximum number of keypoint matches found between the pairs of images i,j and j,i, and K_i and K_j are the number of keypoints found in images i and j, respectively.

$$S_G^{ij} = S_G^{ji} = \frac{M_{ij}}{min(K_i, K_j)} \tag{2}$$

However, when matching face images, we want to avoid that keypoints located in different areas of the face are matched. For instance, a feature close to an eye should not be matched to another close to the mouth. Therefore, the spatial distribution of keypoint matches is taken into account using the local similarity measure proposed in [21], and presented in Eq. 3.

$$S_L^{i,j} = \frac{1}{k} \sum_{n=1}^{k} (max(s(f_{in}^x, f_{jn}^y)) \times w_n), \ \forall_{x,y} \tag{3}$$

Where k is the number of face subregions, $s(f_1, f_2)$ denotes the cosine similarity between the feature descriptors f_1 and f_2, f_{in}^x is the feature x of image i in subregion n, w_n is the importance associated with the region n (the sum of all weights must be unitary), and x and y represent all the features in image i and j, respectively.

Hierarchical agglomerative clustering [22] is used in the present implementation, with an average linkage metric, where the average distance between the objects in the two clusters is considered, as it achieved the best results in the conducted clustering experiments.

[1] https://github.com/BioFoV/BioFoV

B. Super-resolution

As mentioned before, for forensics scenarios multi-frame SR techniques should be used. When trying to register images belonging to a given face cluster one may find complex motions, as the face is not a rigid body, and self-occlusions, with some face elements, such as the nose, occluding others. Therefore, the choice was to use a SR algorithm that does not need an explicit motion estimation [15]. This method involves: (i) choosing the reference image from the available LR images; (ii) register each image onto the reference; and (iii) fuse them into the HR space, using sub-pixel resolution. A deblurring filter is then applied.

1) Reference frame selection: In a forensics scenario, all LR images are typically of low quality. However, if one of them is of better resolution, illumination and/or sharpness, it should be selected as the reference. In the present implementation the image with the best illumination is chosen as reference.

2) Registration: The registration step finds the correspondences between pixels in the reference and all other LR images. Registration accuracy is crucial for the reconstruction success, determining pixel values and positions in the HR grid. Before registration all LR images are interpolated to the desired HR size using a bicubic interpolation.

Due to the complex nature of face images, a parametric registration using translation, rotation, scale or affine transformations (i.e., an explicit motion estimation) may not work well. Instead, an optical flow technique is used, finding the motion of every pixel between two images. In our implementation the optical flow was estimated according to Deqing Sun and Stefan Roth's implementation[2] of the Horn and Schunck's method [23], computing a motion vector for each pixel, which is then reversed for registration purposes. Motion vectors often correspond to sub-pixel displacements, not allowing a ”direct” registration. Two registration processes were considered: (i) rounding the motion vector to pixel resolution; (ii) performing a weighted average according:

$$R(i,j) = \sum_{rows}\sum_{cols} I(rows,cols) w_{row} w_{col} \tag{4}$$

Where R is the registered image, I the input image, i and j the pixel coordinates, u and v the motion vector components, $rows = [i + floor(u), i + ceil(u)]$, $cols = [j + floor(v), j + ceil(v)]$, and w_{row} and w_{col} weights associated with the elements of $rows$ and $cols$, respectively. For the first element of $rows$, $w_{row} = 1 - abs(round(u) - u)$, for the second element, $w_{row} = abs(round(u) - u)$. w_{col} is computed similarly, replacing u with v.

3) Fusion: Conventional fusion processes include a mean or median operation, assuming that registration is perfect. Considering this is not generally true, it is advantageous to adopt a fusion technique resilient to image registration imperfections, such as the one proposed by Elad and Protter [15]. This technique computes the SR image using weighted values from the neighborhood of each pixel.

The present implementation uses a different registration procedure, requiring the fusion technique to also be adapted. The first main difference is that in [15] fusion is performed directly from the LR to the HR space, with HR pixel weights depending on a neighborhood of the corresponding pixel in the LR images. But here images are interpolated to the desired HR size, with the computation of each HR pixel being done according to:

$$\hat{z}(i,j) = \frac{\sum\limits_{[k,l]\in N(i,j)} \sum\limits_{t=1}^{T} W_t[k,l] y_t[k,l] \mathcal{N}\{||(k,l)-(i,j)||_2, 0, \sigma_N\}}{\sum\limits_{[k,l]\in N(i,j)} \sum\limits_{t=1}^{T} W_t[k,l]} \tag{5}$$

Where $N(i,j)$ is a square neighborhood of pixel i,j; $W_t(k,l)$ the weight associated with pixel k,l from image t, y_t the t^{th} interpolated LR image, and $\mathcal{N}\{x,\mu,\sigma\}$ is the Gaussian distribution operator with mean μ, and standard deviation σ, evaluated at x. Using Eq. 5 each pixel is computed as a combination of the weighted values of its neighborhood, times a penalizing factor, which reduces the importance of pixels farther away from the neighborhood center. This penalizing factor was not consider in [15], being a contribution of this paper. The weight of a pixel k,l in the interpolated image t is computed using:

$$W_t[k,l] = exp\left\{-\frac{||P_{k,l}[OF_t^{-1} z - y_t]||_1}{2\sigma_P^2}\right\} \mathcal{N}\left\{|dk_t| + |dl_t|, 0, \sigma_D\right\} \tag{6}$$

Where z is the HR targeted image, dk_t and dl_t are the k,l values of the optical flow in image t, and OF_t^{-1} is the reversed optical flow registration information, which yields the simulated interpolated image when applied to z. Since z is not known, the interpolated reference image is used, which is a relatively good guess. Then, its difference to the real interpolated image y_t is found, providing an error image. A patch P is extracted around the k,l pixel from this error image, and the L1-norm is applied to it. In [15], the equivalent error (not the same because they simulate the LR image, instead of the interpolated one) was computed with the L2-norm. However, we used the L1-norm in order not to penalize in excess relatively small errors. Finally, a displacement penalty is also included, which reduces the importance of pixels that have large motions in the registration. Here we also use the L1 instead of the L2-norm to compute the displacement error.

4) Deblurring: The two most relevant kinds of blur are motion blur and the camera's blurring from its natural Point Spread Function (PSF). The former is implicitly dealt with by using several pictures and registering them to a reference. Therefore, the latter is the one which we need to address. However, we don't know the PSF from the camera. Also, considering a neighborhood of the pixel in the fusion process causes a kind of blur similar to the one from a low-pass filter. Therefore, we use a blind deconvolution algorithm which

[2] https://github.com/ahmadh84/occlusiontracking

deblurs the image and tries to find the PSF, simultaneously. The PSF and deblurred image are found using the Lucy-Richardson method of *deconvblind*.

5) Alpha-blending: For some parts of the image, where the gradient is small and the image values alter slowly, the interpolation of the image may be enough (or even perform better) than applying SR. On the contrary, SR is preferred when sub-pixel displacements from several LR images are needed to reconstruct the HR image. Therefore, blending the interpolated reference image with the super-resolved image may achieve better results. The blending is performed by:

$$I_{\alpha b} = \alpha I_{SR} + (1-\alpha) I_{Int}, \alpha \in [0,1] \tag{7}$$

where I_{SR} is the super-resolved image and I_{Int} is the interpolated image.

III. Experimental Results

In this section, the experiments performed to evaluate both the registration and the SR techniques are presented. We assess the results using two measures. The first one is the Peak Signal to Noise Ratio (PSNR) which is defined as:

$$PSNR = 10 log_{10}(\frac{peakval^2}{MSE}) \tag{8}$$

where $peakval$ is the maximum value that the data can take, which is 255 for images, and MSE is the Mean Squared Error. The second used measures is the Structure Similarity Index (SSIM), defined by:

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

in which μ_x, μ_y, σ_x, σ_y and σ_{xy} are, respectively, the mean, standard deviation and cross-variance for images x and y. C_1 and C_2 are constants.

The experiments were performed on benchmark video sequences for SR, namely ”Foreman” and ”Suzie”, and on a cluster of 9 face images extracted from a simulated surveillance scenario.

A. Registration Experiments

Five different registration techniques were implemented and compared. Two based on simple geometric transforms (affine and rigid body), two based on the optical flow algorithms described in Section II-B.2, and one hybrid technique, which consists of applying a rigid body transform to align the images before registering with the optical flow from equation 4.

The comparison here shown, was made with the ”Foreman” video sequence, having as a reference the first frame, and registering the next 8 frames onto that one. The results are shown in Fig. 2, where ”OF” and ”OF Prob” are the Optical Flow algorithms (i) and (ii) described in section II-B.2.

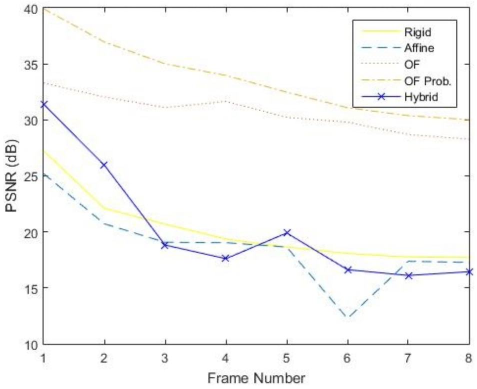

Fig. 2. Comparison of five different registration algorithms

Fig. 2 shows that the optical flow registration methods perform much better than the geometric transformation techniques, which is due to the complex motions of the human face, like local movements, which can’t be handled by simple geometric transformations. Therefore, as expected, the larger the facial movement, the larger is the difference between the optical flow and the geometric transformations performances. The ”probabilistic” optical flow performs better than the ”direct” one, because it takes into account all pixels involved in the motion area. The hybrid registration has a worse performance than the other optical flow techniques because the rigid body transformation applied in the beginning doesn’t help aligning the images, only distorts and creates noise before the optical flow is applied.

A visual comparison of the performance of the several registration techniques applied on the 5^{th} input frame is shown in Fig. 3.

Due to the above obtained results, it was decided to use the probabilistic optical flow as the registration technique of the proposed SR algorithm.

B. SR Experiments

In this section, the results obtained by applying the SR algorithm described in section II-B are presented. The five first frames of the two aforementioned video sequences were used. The shown results were obtained using the parameters that yielded the best performance metrics: Patch size=$3x3p$, $\sigma_N = 1$, $\sigma_P = 4$ and $\sigma_D = 2.5$. It is worth mentioning that these parameters may not be perfect, because their optimum value changes for every different sequence, and for different resolutions. If the optimal relationship between the parameters and the image sequences is found, the results may improve significantly.

Because usually face images from surveillance videos are very small, we downsized the benchmark video sequences to half, yielding sizes of $144 \times 156p$ and $120 \times 176p$, for the ”Foreman” and ”Suzie” sequences, respectively. The results are summarized in tables 1 and 2, and they regard only a

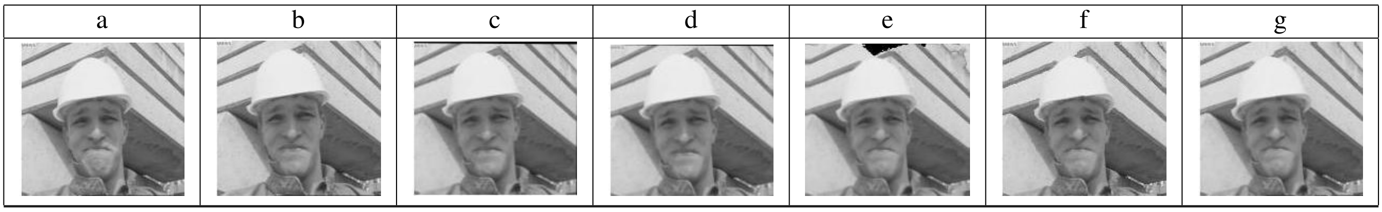

Fig. 3. a is the reference image, b is the image to be registered, c, d, e, f and g are the registered images with rigid body transformation, affine transformation, direct optical flow and probabilistic optical flow, respectively.

Sequence	a	b	c	d	e	f [24]	g [25]
Foreman	33.85	30.81	32.35	34.07	34.42 ($\alpha = 0.6$)	32.82	34.01
Suzie	32.35	31.17	31.61	32.91	33.02 ($\alpha = 0.7$)	32.97	33.09
Surveillance	29.55	30.33	-	30.69	30.70 ($\alpha = 0.9$)	-	-

Table 1. PSNR results of proposed SR technique in dB's. Column a contains the interpolated image results, b the simple super-resolved image, c the results of b registered into a, d results of c after deblurring (or b if c isn't computed), e alpha-blending of d and a, f the results from [24] (generalized non-local means), and g results from [25] (non-local kernel regression).

Sequence	a	b	c	d	e	f [24]	g [25]
Foreman	0.9411	0.8944	0.9119	0.9291	0.9422 ($\alpha = 0.2$)	0.9025	0.9120
Suzie	0.9252	0.8983	0.9020	0.9236	0.9285 ($\alpha = 0.5$)	0.8797	0.8671
Surveillance	0.8864	0.8819	-	0.8971	0.8998 ($\alpha = 0.7$)	-	-

Table 2. SSIM results of proposed SR technique. Column a contains the interpolated image results, b the simple super-resolved image, c the results of b registered into a, d results of c after deblurring (or b if c isn't computed), e alpha-blending of d and a, f the results from [24], and g results from [25].

region of interest (the face), as the outsides of the image aren't important in our context.

The SR results are compared to the bicubic interpolation of the reference image. In column "SR Simple" are the results of the algorithm described in section II-B with no other changes. Then, if this algorithm provides worse results (in terms of PSNR) than the ones from the interpolated image (cases of "Foreman" and "Suzie" sequences), the super-resolved image is again registered against the latter. Then, a blind deconvolution deblurring is applied on this registered image. Finally, an alpha-blending technique is applied, blending the interpolated and the deblurred images. In the cases that it improves the performance, the metrics are presented.

By analysing the results, we can verify that the proposed technique outperforms the interpolated image by about $1dB$ in terms of PSNR, while in terms of the SSIM, there are improvements by the orders of 10^{-3} to 10^{-2}. Before comparing the results with other state of the art works, like [24] and [25], it is important to notice that, although the used video sequences were the same, the experiments were performed under different conditions: first, unlike the other works we downsampled the images to half their size before the algorithm was applied; second, our values only regard a region of interest (around the face), while in [24] and [25] the whole image is considered. Having this in mind, we can verify that in terms of PSNR, the results from our work outperform or are comparable to the state of the art. In terms of SSIM, which has into account the way that the human eye perceives an image, the proposed method provides the best results (Fig. 5).

IV. Conclusion and future works

In this paper we presented a way to detect and cluster face images from a surveillance video, to use them as input to a proposed multi-frame SR algorithm, in order to obtain a higher resolution image from an individual for forensic purposes. We addressed several registration techniques, comparing their performance on images of the "Foreman" video sequence. We chose the registration technique which yielded the best results (probabilistic optical flow) for the SR algorithm. This algorithm is a direct one, which contains an image fusion step. This step was based on the state of the art work from [15], but adapted to our optical flow registration. The results from the proposed algorithm outperforms the bicubic interpolation, and are comparable to the state of the art. As for future work, improvements to the proposed SR technique should be made, namely in finding the optimal relation between the input images and the algorithm parameters. Also, experiments with benchmark face recognition databases can be performed, in order to validate the use of the proposed work in a forensic scenario.

References

[1] K. Nasrollahi, T.B. Moeslund, and M. Rahmati. Summarization of surveillance video sequences using face quality assessment. *International Journal of Image and Graphics (IJIG)*, 11 (2): 207-233, 2011.

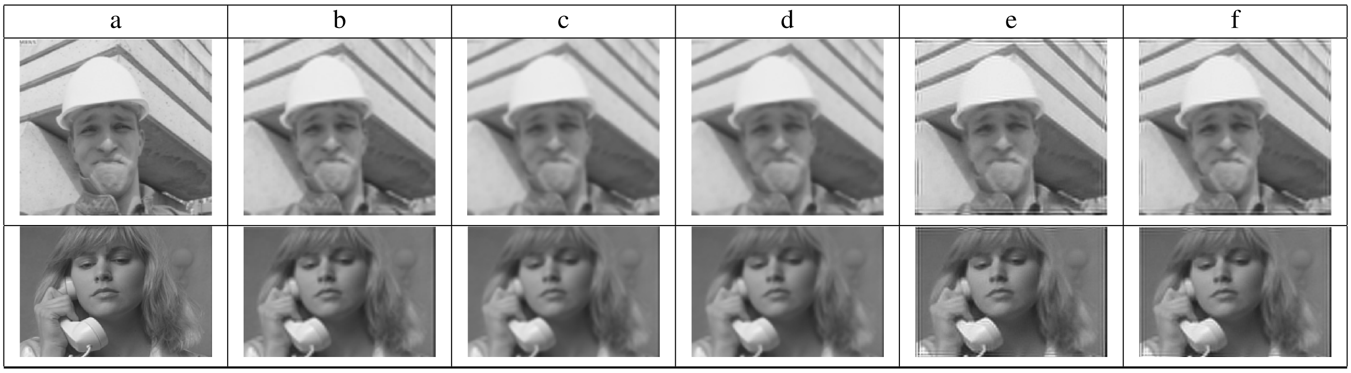

Fig. 4. Benchmark video sequences results. a is the reference image, b is interpolated (bicubic) image, c is the super-resolved image with probabilistic fusion, d is the super-resolved registered into the interpolated one, e is the result of SR after debluring, and f is the results of alpha-blending of e and b.

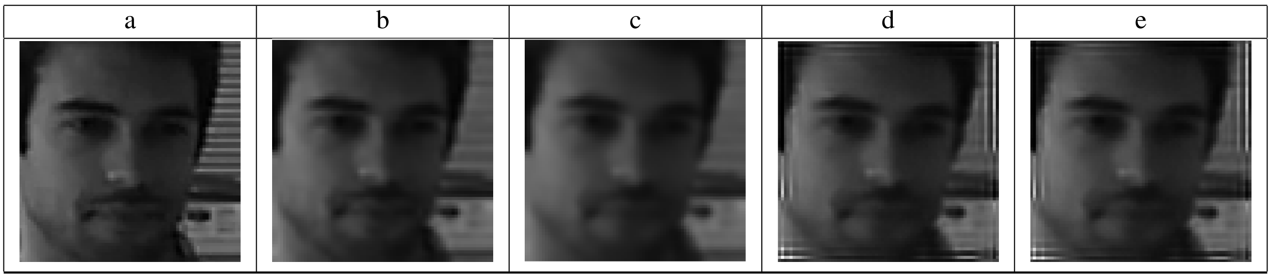

Fig. 5. Simulated surveillance video results. a is the reference image, b is interpolated (bicubic) image, c is the super-resolved image with probabilistic fusion, d is the result of SR after deblurring, and f is the results of alpha-blending of d and b.

[2] K. Nasrollahi and T.B. Moeslund. Super-resolution: a comprehensive survey *Machine Vision and Applications (MVAP)*, 25 (6): 1423-1468, 2014.

[3] J.B. Huang, A. Singh, and N. Ahuja. Single image super-resolution using transformed self-exemplars. *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2015.

[4] C. Dong, C.C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. *European Conference on Computer Vision (ECCV)*, 2014.

[5] Y. Zhu, Y. Zhang, and A.L. Yuille. Single image super-resolution using deformable patches. *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2917-2924, 2014.

[6] C.Y. Yang and M.H. Yang. Fast Direct super-resolution by Simple Functions. *Computer Vision (ICCV), IEEE International Conference on*, 561-568, 2013.

[7] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. *Computer Vision (ICCV), IEEE International Conference on*, 349-356, 2009.

[8] K. Nasrollahi, T.B. Moeslund. Hybrid Super-resolution using refined face logs. *Image Processing Theory Tools and Applications (IPTA), 2nd International Conference on*, 435-440, 2010.

[9] K. Nasrollahi, T.B. Moeslund. Finding and improving the key-frames of long video sequences for face recognition. *Biometrics: Theory Applications and Systems (BTAS), 4th IEEE International Conference on*, 1-6, 2010.

[10] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu. Handling motion blur in multi-frame super-resolution. *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2015.

[11] G. Polatkan, M. Zhou, L. Carin, D. Blei, and I. Daubechies. A Bayesian nonparametric approach to image super-resolution. *Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on*, 37(2): 346-358, 2015.

[12] E. Turgay and G.B. Akar. Texture and edge preserving multiframe super-resolution. Image Processing, IET, 8(9): 499-508, 2014.

[13] C. Jin, J.L. Nunez-Yanez, and A. Achim. Bayesian video super-resolution with heavy-tailed prior models. Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on, 24(6): 905-914, 2014.

[14] E. Quevedo, J. de La Cruz, G. Callico, F. Tobajas, and R. Sarmiento. Video enhancement using spatial and temporal super-resolution from a multi-camera system. Consumer Electronics, IEEE Transactions on, 60(3): 420-428, 2014.

[15] M. Protter, M. Elad Super Resolution With Probabilistic Motion Estimation *Image Processing, IEEE Transactions on* Vol:18, Issue: 8, 1899-1904, 2009.

[16] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on*, 1: 511-518, 2001.

[17] R. Lienhart, J. Maydt. An Extended Set of Haar-like Features for Rapid Object Detection *Image Processing, IEEE International Conference on*, 2002.

[18] S. Kayal. Improved Hierarchical Clustering for Face Images in Videos: Integrating positional and temporal information with HAC. *Multimedia Retrieval, ACM International Conference on*, 2014.

[19] Gary B. Huang, Manu Ramesh, Tamara Berg, Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments *University of Massachusetts, Amherst, Technical Report*, 07-49, 2007.

[20] D.G. Lowe. Object recognition from local scale-invariant features *Proceedings of the International Conference on Computer Vision*, 1150-1157, 1999.

[21] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, B.-L. Lu. Person-specific SIFT features for face recognition *Acoustics, Speech and Signal Processing, IEEE International Conference on*, 2007.

[22] S. Kayal. Improved Hierarchical Clustering for Face Images in Videos: Integrating positional and temporal information with HAC *Proceedings of International Conference on Multimedia Retrieval*, 2014.

[23] B.K.P. Horn, B.G. Schunck. Determining Optical Flow *Artificial Intelligence*, 17: 1-3, 185-203, 1981.

[24] M. Protter, M. Elad, H. Takeda, P. Milanfar Generalizing the Non-Local-Means to Super-resolution Reconstruction *IEEE TIP* 36-51, 2009.

[25] H. Zhang, J. Yang, Y. Zhang, T. S. Huang Image and Video Restorations via Nonlocal Kernel Regression *Cybernetics, IEEE Transactions on* 43: 3, 1035-1046, 2013.