Super-resolution of facial images in forensics scenarios

João Satiro¹, Kamal Nasrollahi², Paulo L. Correia¹, and Thomas B. Moeslund²

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.au.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
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. De-
tails 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\textsuperscript{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:

\begin{align}
80 \leq Cb \leq 120, 133 \leq Cr \leq 173
\end{align}

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,
iluminations, 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.

\begin{equation}
S_G^{ij} = S_G^{ji} = \frac{M_{ij}}{\min(K_i,K_j)}
\end{equation}

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.

\begin{equation}
S_L^{ij} = \frac{1}{k} \sum_{n=1}^{k} (max(s(f_{in}^x,f_{jn}^y)) \times w_n), \forall x,y
\end{equation}

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.

\textsuperscript{1}https://github.com/BioFoV/BioFoV

Fig. 1. The block diagram of the proposed system.
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 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. The 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, \( rows = [i + \text{floor}(u), i + \text{ceil}(u)] \), \( cols = [j + \text{floor}(v), j + \text{ceil}(v)] \), and \( w_{row} \) and \( w_{col} \) weights associated with the elements of \( rows \) and \( cols \), respectively. For the first element of \( rows \), \( w_{row} = 1 - \text{abs}(\text{round}(u) - u) \), for the second element, \( w_{row} = \text{abs}(\text{round}(u) - u) \). \( w_{col} \) 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:

\[ \hat{z}(i, j) = \frac{\sum_{[k,l]\in N(i,j)} W_t[k,l] y_t[k,l] |N([k,l])−(i,j)||z.,0,\sigma_N} \sum_{[k,l]\in N(i,j)} W_t[k,l] \tag{5} \]

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 \( N(x, \mu, \sigma) \) is the Gaussian distribution operator with mean \( \mu \), and standard deviation \( \sigma \), 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:

\[ W_t[k,l] = \exp \left\{ -\frac{||P_{t} OF_{t}^{-1} x - y_t||_1}{2\sigma_t^2} \right\} N \{ |dk| + |dl|, 0, \sigma_D \} \tag{6} \]

Where \( z \) is the HR targeted image, \( dk \) and \( dl \) 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
to Noise Ratio (PSNR) which is defined as: the results using two measures. The first one is the Peak Signal
Error. The second used measures is the Structure Similarity
Index (SSIM), defined by:

\[ \text{PSNR} = 10 \log_{10} \left( \frac{\text{peakval}^2}{\text{MSE}} \right) \]  
\[ \text{SSIM} = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \]

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:

where \( \mu_x, \mu_y, \sigma_x, \sigma_y \) and \( \sigma_{xy} \) 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.

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:

\[ I_{ab} = \alpha I_{SR} + (1 - \alpha) I_{Int}, \alpha \in [0, 1] \]  

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

III. EXPERIMENTAL RESULTS

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.

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
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 downsamled 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

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 deblurring, 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 deblurring fusion, d is the result of SR after deblurring, and f is the results of alpha-blending of d and b.


