Superresolution and blind deconvolution of video

Filip Šroubek, Jan Flusser, and Michal Šorel
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodarenskou vezi 4, Prague 8, 182 08, Czech Republic
{sroubekf, flusser, sorel}@utia.cas.cz

Abstract

In many real applications traditional superresolution methods fail to provide high-resolution images due to objectionable blur and inaccurate registration of input low-resolution images. In this paper, we present a method of superresolution and blind deconvolution of video sequences and address problems of misregistration, local motion and change of illumination. The method processes the video by applying temporal windows, masking out regions of misregistration, and minimizing a regularized energy function with respect to the high-resolution frame and blurs, where regularization is carried out in both the image and blur domains. Experiments on real video sequences illustrate robustness of the method.

1. Introduction

Imaging devices, such as camcorders or web cameras, have limited achievable resolution due to many theoretical and practical restrictions. An original scene, represented by a discrete high resolution (HR) image denoted as f, warps at a camera lens because of the scene motion and/or change of the camera position. In addition, several external effects blur images: atmospheric turbulence, camera defocus, lens abberations, relative camera-scene motion, etc. Most of these effects are unpredictable and of transitory behavior, yet we assume that we can model them by convolution with an unknown point spread function (PSF) h. Finally, the camera sensor discretizes the image and produces digitized noisy image g, which we refer to as a *low-resolution (LR) image*, since the spatial resolution is too low to capture all the details of the original scene. In video, f changes in time, which implies that observed LR frames g's are also time dependent. We assume that the changes are sufficiently slow with respect to the frame rate of the video, so that for the given time j there exists a temporal window $\mathcal{W}^j = \{j-L/2,\ldots,j,\ldots,j+L/2\}$ of length $(L+1)$, $L>0$, in which the original high-resolution (HR) frame f^j can be related to the LR frames g_i ($i \in \mathcal{W}^j$) by the following acquisition model in the vector-matrix format

$$\mathbf{g}_i = \mathbf{D}(\theta_i^j)\mathbf{H}_i^j\mathbf{f}^j + \mathbf{n}^j\,. \tag{1}$$

A schematic depiction of the relation is in Fig. 1. The superscript j indexes HR frames and the subscript i indexes LR frames. The noise vector $\mathbf{n}^j$ is assumed to be additive and independent of f^j. Matrix $\mathbf{D}(\theta_i^j)$ denotes a decimation operator, which performs warping (defined by a set of parameters θ_i^j), convolution with a sensor PSF (assumed to be known), and downsampling. This way it models the acquisition process of the camera and a geometric transformation between the i-th and j-th frame. Matrix $\mathbf{H}_i^j$ denotes convolution with an unknown PSF h_i^j.

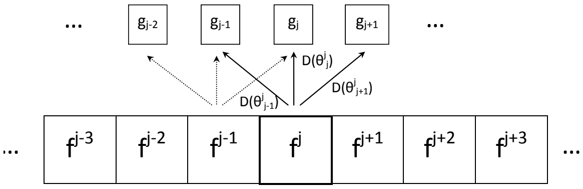

Figure 1. Acquisition model.

The above model is the state of the art as it takes all possible degradations into account and combines two fundamental problems of image processing: multiframe blind deconvolution to estimate h and superresolution (SR) to estimate HR frame f.

Current blind deconvolution techniques [5, 2] require no or very little prior information about the blurs, they are sufficiently robust to noise and provide satisfying results in most real applications. However, they can

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

hardly cope with the decimation operator, i.e. change of resolution, which violates the standard convolution model. On the contrary, state-of-the-art SR techniques [1] achieve remarkable results of resolution enhancement in the case of no blur. They accurately estimate the subpixel shift between images but lack any apparatus for calculating the blurs.

Recently in [4], we proposed a unifying method that simultaneously estimates the blurs and HR image from multiple LR images. The key idea was to determine subpixel shifts by calculating the blurs. As the blurs are estimated in the HR scale, positions of their centroids correspond to sub-pixel shifts. Therefore by estimating blurs we automatically estimate shifts with subpixel accuracy, which is essential for good performance of SR.

This paper extends our previous work to video and presents remedies for two common problems in SR of video: change of illumination and local motion. Apart from robustness to misalignment, including estimation of blurs in the proposed method cancels effects of change of illumination. For warping (registration) of frames, we assume a homography model, which is mostly sufficient even for scenes with significant variations in depth, since change of camera position between neighboring video frames is relatively small. Nevertheless, homography cannot map regions that contain local motion. Thus discrepancies in preregistered images give us regions where local motion is highly probable. Masking out such regions in the decimation operator $\mathbf{D}$ and performing simultaneously blind deconvolution and SR, produces naturally looking HR frames. In regions, which are masked out in every frame, interpolation takes place, but in the rest precise SR can be calculated.

Section 2 outlines an alternating minimization approach to solve (1) and discusses each step of the proposed algorithm. Experiments on true web-camera sequences demonstrate performance of the proposed method in Section 3 and Section 4 concludes the paper.

2. Iterative restoration

In order to find the estimate of the HR video sequence $\{f^j\}$, we adopt an approach of minimizing a regularized energy function, which makes the method robust to noise and well posed. The energy function consists of three terms and takes the form

$$E(\mathbf{f}^j, \mathbf{h}^j, \theta^j) = \sum_{i \in \mathcal{W}^j} \|\mathbf{M}_i^j(\mathbf{D}(\theta_i^j)\mathbf{H}_i^j\mathbf{f}^j - \mathbf{g}_i)\|^2 + \\ + \alpha Q(\mathbf{f}^j) + \beta R(\mathbf{h}^j)\,, \quad (2)$$

where $\mathbf{h}^j$ denotes all blurs $\{h_i^j\}$ in the temporal window $\mathcal{W}^j$. Likewise, θ^j denotes a set of homography parameters $\{\theta_i^j\}$, $i \in \mathcal{W}^j$. Matrix $\mathbf{M}_i^j$ is a diagonal matrix, which performs masking of regions with local motion. The first term measures the fidelity to data and emanates from our acquisition model (1). The remaining two are regularization terms with weights α and β, which will be discussed later.

To find a minimizer, we perform alternating minimization of E over $\mathbf{f}^j$, $\mathbf{h}^j$ and θ^j. The proposed algorithm is outlined below and a more detail discussion of each step follows.

Algorithm

For each reference time j and associated frame sequence given by temporal window $\mathcal{W}^j$

1. Estimate homography $\{\theta_i^j\}$ between the reference frame g_j and each g_i for $i \in \mathcal{W}^j$. Calculate masks $\mathbf{M}_i^j$ and construct decimation operators $\mathbf{D}_i^j$. Initialize $\{\mathbf{h}_i^j\}$ with delta functions.

2. Find a new estimate of HR image

$$\mathbf{f}^j = \arg\min_{\mathbf{f}} E(\mathbf{f}, \mathbf{h}^j, \theta^j)\,. \quad (3)$$

3. Find a new estimate of PSFs

$$\mathbf{h}^j = \arg\min_{\mathbf{h}} E(\mathbf{f}^j, \mathbf{h}, \theta^j)\,. \quad (4)$$

4. Adjust homography parameters

$$\theta^j = \arg\min_{\theta} E(\mathbf{f}^j, \mathbf{h}^j, \theta) \quad (5)$$

   and update $\mathbf{D}_i^j$.

5. Repeat steps 2–4 until the image $\mathbf{f}^j$ meets a convergence criterion.

For superresolution purposes, the homography between the frames must be estimated with high precision. In addition, we need a global registration procedure that local motion does not disrupt. The following combination of phase correlation, minimization of least squares error between frames and RANSAC worked well for all video sequences we tested:

- Estimate shift by phase correlation. The frame being registered is shifted accordingly.

- If the phase correlation fails (difference after registration is larger than a threshold), apply one of standard homography estimation procedures based on a robust detector of control points and RANSAC [3].

- Adjust homography matrix by minimizing the least square error between the reference frame and the processed frame transformed using bicubic interpolation.

The decimation operator $\mathbf{D}(\theta_i^j)$ maps pixels of the estimated HR frame f^j to pixels of the observed LR frame g_i. The number of registration parameters θ_i^j depends on the type of geometric transform. In our case, we consider homography, i.e., 8 parameters for each i-j pair. To better model the acquisition process of the camera sensor, we include the sensor PSF (intrinsic blur of the camera) in to the decimation matrix. The sensor PSF is assumed to be of the Gaussian shape of known variance. The HR-LR mapping is done by associating with each row of $\mathbf{D}(\theta_i^j)$ a discrete sensor PSF, which is displaced and deformed according to the given homography. Finally, we mask out erroneous LR pixels by multiply the decimation operator by a diagonal matrix $\mathbf{M}_i^j$, which has zeros at locations of incorrect pixels and ones elsewhere. The erroneous pixels are located, e.g., in regions of local motion, where the global geometric transform does not hold. To determine $\mathbf{M}_i^j$, we perform the following. We take the difference of registered LR frames g_i and g_j and threshold its magnitude. Values below 10% of the intensity range of LR frames are considered as correctly registered and corresponding mask pixels are set to one; remaining mask pixels are zeroed. In order to attenuate the effect of misregistration errors, the morphological operator "closing" is then applied to the mask. Note that $\mathbf{M}_j^j$ will be always identity and therefore HR pixels of f^j in regions of local motion will be at least mapped to LR pixels of g_j. Depending on how many LR images map to the HR image, the restoration algorithm performs in each region tasks from simple interpolation up to well-posed super-resolution.

For the image regularization terms we use total variation, $Q(f) = \int |\nabla f|$. It seems to be a reasonable choice for common images taken by a standard camera. While in smooth areas it has the same isotropic behavior as the Laplacian operator, it also preserves edges in images. However, it is nonlinear and one must employ linearization techniques, such as half-quadratic algorithm. For the purpose of our discussion it suffices to state that after linearization and discretization we arrive at

$$Q(\mathbf{f}^j) = \mathbf{f}^{j\,\mathrm{T}}\mathbf{L}\mathbf{f}^j\,, \qquad (6)$$

where $\mathbf{L}$ is a symmetric block diagonal matrix constructed from values of the gradient of f and it is updated after every iteration of the algorithm.

The PSF regularization term $R(\mathbf{h}^j)$ is intrinsically multiframe as it utilizes relations between all the LR frames $\{g_i^j\}$ in $\mathcal{W}^j$. An exact derivation is given in [4]. Here, we leave the discussion by stating that the regularization term becomes

$$R(\mathbf{h}) = \mathbf{h}^{j\,\mathrm{T}}\mathbf{N}\mathbf{h}^j\,, \qquad (7)$$

where $\mathbf{N}$ is a symmetric matrix that depends solely on $\{g_i^j\}$, $i \in \mathcal{W}^j$.

Steps 2 and 3 both solve a system of linear equations, since each term of (2) is quadratic w.r.t. $\mathbf{f}$ and $\mathbf{h}$. Step 4 requires numeric approximation of derivatives, but since $\mathbf{h}^j$ compensate for misalignment, this step is often not necessary. Change of illumination results in change of contrast in frames, i.e., multiplication of image intensity values by a constant. If the estimated PSF energy ($\sum_x h_i^j(x)$) differs from 1, convolution with such PSF automatically adjusts contrast.

3. Experiments

The following two experiments demonstrate the ability of the proposed method to deal with real video sequences including elimination of artifacts in regions of local motion. We used a standard web camera to capture short (20s) video sequences with 30 fps and shutter speed $1/30$s. In both cases, we worked with central sections of the videos of size roughly 100×100 pixels.

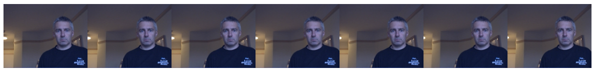

Figure 2. Input video sequence.

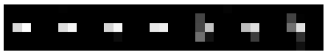

Figure 3. PSFs corresponding to the input video sequence.

Fig. 2 shows the first seven frames of the original video sequence used for computation of the HR frame in the first experiment. Note that there is a significant translation component between frames. The fourth frame is used as a reference frame and three preceding and following frames are independently registered to it (temporal window of size 7). Then, we reconstruct the high resolution image as detailed in steps 2–5 of the algorithm. No local motion occurred in the tested video sequence and therefore masking matrices $\mathbf{M}_i^j$ approach identity. In Fig. 3 we can see the estimated PSFs, which correspond to LR frames in Fig. 2 and which model external blurring and help to eliminate inaccuracy in sub-

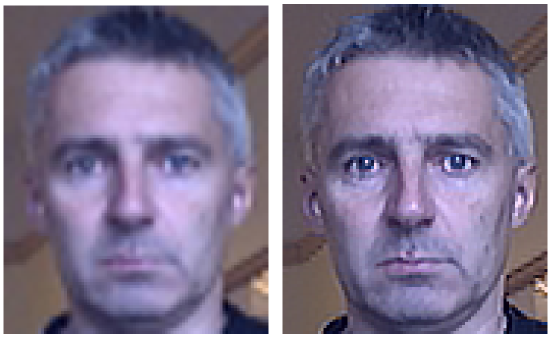

Figure 4. HR image computed by bicubic interpolation (left) and the proposed method (right).

pixel registration. Fig. 4 demonstrates that the new approach (right) is clearly superior to simple bicubic interpolation of the original frames (left).

The second experiment illustrates the advantage of masking if the video sequence contains local motion, such as a person waving a hand. The temporal window was set to 10 frames. An example of 5 (1, 3, 5 – reference, 7, 9) LR frames of one such temporal window is in Fig. 5 (top). Displacements of the waving hand are apparent. Registering the frames in the first step of the algorithm removes homography. The calculated masks in Fig. 5 (bottom) show that most of the erroneous pixels are around the waving hand. Note that during HR reconstruction only the middle frame, which is the reference one and does not have any mask, provides information about the pixels in the region of the waving hand. Comparison of estimating HR frames with and without masking is in Fig. 6. Ignoring masks results in heavy artifacts in the region of local motion. On the contrary, masking produces smooth results with the masked-out regions properly interpolated.

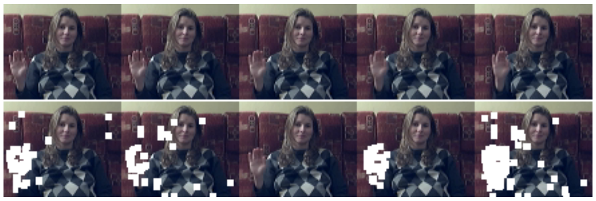

Figure 5. Input video sequence with motion. Examples of 5 frames (top) with regions masked out (bottom).

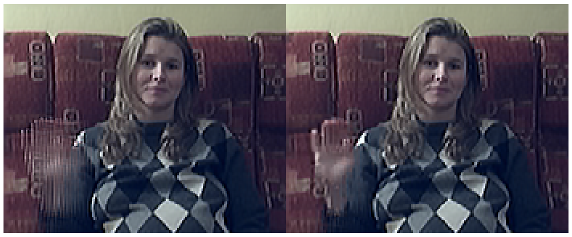

Figure 6. HR image using a traditional SR method (left) and our proposed method with masking (right).

4. Conclusion

The proposed algorithm performs simultaneously resolution enhancement and deblurring of video sequences. Introducing the deconvolution step renders the method less vulnerable to misregistration and change of illumination. Special attention is paid to local motion in video, which can produce heavy artifacts. Using masks we eliminate such artifacts at a price of performing only interpolation in the regions of local motion.

In future, we plan to research on predicting the quality of reconstructed HR frames based on the shape of PSF estimates, speeding up calculation by utilizing previous HR frames, and incorporating motion fields to segment the scene and perform segment-wise reconstruction.

References

[1] *Super-Resolution Enhancement of Digital Video*. EURASIP Journal on Advances in Signal Processing, 2007.

[2] S. Babacan, R. Molina, and A. Katsaggelos. Parameter estimation in TV image restoration using variational distribution approximation. *IEEE Trans. Image Processing*, 17(23):326–339, Mar. 2008.

[3] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University, Cambridge, 2nd edition, 2003.

[4] F. Šroubek, G. Cristóbal, and J. Flusser. A unified approach to superresolution and multichannel blind deconvolution. *IEEE Trans. Image Processing*, 16(9):2322–2332, Sept. 2007.

[5] F. Šroubek and J. Flusser. Multichannel blind deconvolution of spatially misaligned images. *IEEE Trans. Image Processing*, 14(7):874–883, July 2005.