Video Resolution Enhancement in the Presence of Moving Objects

Mahmood Amintoosi, Mahmood Fathy and Nasser Mozayani
Computer Engineering Department, Iran University of Science and Technology
Narmak, Tehran, Iran
{mAmintoosi,mahFathy,Mozayani}@iust.ac.ir

Abstract— *Moving objects is one of the main bottlenecks in many computer vision problems including video super resolution. Video Super-resolution algorithms reconstruct a high resolution video from a low resolution video. Recent advances in Super-Resolution techniques show trends towards model-based or example-based approaches. In this article a new method for video super-resolution problem based on our recent paper is proposed with at the same time handles moving objects efficiently. The overall method is based on this hypothesis, which some high-resolution(HR) images from the scene are available. Instead of the previous example based methods, which do blocking, here the whole training image is mapped to each frame. Proper mapping of the HR image onto LR frames has been done using a combination of feature-based and area-based registration methods. After creating a suitable mask for handling moving objects, each LR frame is fused with the registered HR image. The simulation results show the better performance of the proposed algorithm in compare with some other super-resolution methods.*

Keywords: Video Super-Resolution, Synthesis, Homography, SIFT, Registration.

1. Introduction

Nowadays digital cameras have been papular and taking films and movies became usual tasks. Many of low-price cameras – such as some mobile phones – can take low-resolution (LR) videos. Every one like to enhancement his/her LR videos or photos. Among the various image and video enhancement and restoration methods, Super-Resolution(SR) methods are the only ones which produce an output that has higher resolution than the input. The origin of the classic from of Super-Resolution known as Multiple Input Single Output(MISO) come back to work of Tsai and Huang [1] in 1984, motivated by the need to improve the resolution of images acquired by the Landsat 4 satellite [2]. The vast majority of the superresolution restoration algorithms – named as resonstruction methods – lie in this category, which use a short sequence of low-resolution input frames to produce a single superresolved high-resolution output frame.

The more general form of super-resolution problems known as Multiple Input Multiple Output(MIMO) is about to enhancement the spatial and/or temporal resolution of the videos. All of the MISO techniques may be applied to video restoration by using a shifting window of processed frames as illustrated in figure 1. For a given superresolution frame, a "sliding window" determines the subset of LR frames to be processed to produce a given super-resolution output frame. The window is moved forward in time to produce successive superresolved frames in the output sequence [2].

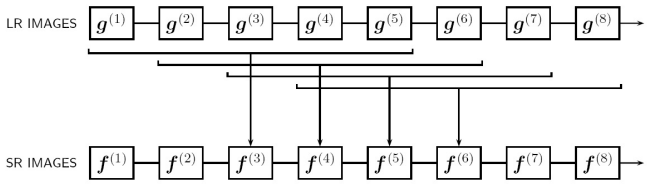


Fig. 1: "sliding window" technique for video super resolution [2].

The analysis performed by Lin and Shum[3], indicates that to achieve super resolution at large magnification factors, reconstruction based algorithms are not favorable and one should try other kinds of super resolution algorithms, such as model-based or example-based algorithms. The model-based approaches import plausible high-frequency textures from an image database into the low resolution image. These methods have gained significant interests in recent years because it promises to overcome the limit of reconstruction-based SR [4].

In previous example-based super-resolution algorithms [5], [6], [7] during the training phase, pairs of low-resolution(LR) and the corresponding high-resolution(HR) image patches are collected. Then, in the super-resolution phase, each patch of the given low-resolution image is compared to the stored low-resolution patches, and the high-resolution patch corresponding to the nearest low-resolution patch is selected as the output. Freeman *et. al.*[6] used a set of HR images as training data set. The super-resolution was performed by the Nearest Neighbor-based estimation of high-frequency patches based on the corresponding patches of input low-frequency image. The corresponding high frequencies patch of the best match has been selected for enhancing the resolution of the LR patch. To ensure that the high-frequency prediction is compatible with its neighbors, the pixels in the low-frequency patch and the high frequency overlap are concatenated to form a search vector. The method of Pham [4], is similar to Freeman's method but

in Discrete Cosine Transform (DCT) domain. His method produces better result than spatial methods specially with the compressed videos.

Although the mentioned methods has already shown an impressive performance, there is still room for improvement if we do not restrict ourself to small patches. In this paper a different approach for enhancement of video resolution is proposed which is based on this extra assumption that some high resolution images of the same scene is available. Here instead of taking small training patches from HR images, we considered the whole part of HR image for increasing the resolution of the input low resolution image. It is supposed that the HR image may be different with the LR image from the following aspects:

- **View Point**, due to camera movement,
- **Illumination**, due to different of exposure time or taking photos in distinctive times.
- **Resolution**, due to unequal zooming or changing the resolution setting of camera, or using different devices for image capturing.

In many situations we encountered with the above conditions. Some cases, which some one may have LR and HR images from a specified scene are as follows:

- The owner of digital camera, takes photos with different resolution by manually setting the capturing resolution, because of storage capacity limitation.
- Reducing the camera resolution mistakenly by the owner or some other one.
- Having photos which are captured in different times or by different devices from a scene, that have some differences about view point, resolution and lighting.
- Having LR video frames and HR images, due to camera limitations.

In the above situations he/she likes to enhance his/her LR images using HR images. Sometimes our HR images could not cover the entire scene of LR image. This leads to resulting images with spatially varying resolution. Only the resolution of those parts of the LR image which have correspondence in HR images will be increased. In [8] we proposed the method for enhancing an still LR image with some HR images from the same scene without any objects. In this paper we extend our approach for LR video and at the same time handle the moving objects.

Moving objects in input images as the source of some errors is one of the major challenges in super-resolution domain. Eren *et. al.*[9] proposed a robust, object-based approach for SR problem using POCS framework. Their proposed method employs a validity map and a segmentation map. The validity map disables projections based on observations with inaccurate motion information for robust reconstruction in the presence of motion estimation errors. For their approach an accurate motion segmentation must be available, which is difficult to obtain in the presence of aliasing and noise. They assumed that objects of interest are marked on a reference low resolution frame interactively. Moreover they have to generate a spatially varying threshold for their validity map structure. Zomet *et. al.*[10] proposed a robust median estimator, which is combined in an iterative process to achieve a super resolution algorithm. Their method is interesting and is the basis or a part of some other methods [11], [4]. But their method suffers from ghost artifacts. Farsiu *et. al.*[12] proposed a Shift-and-Add method which is a simple fusion approach, but outlier effects are apparent in its output. The normalized convolution method presented by Pham [4] produces very good results, but it is a very high time consuming method.

The proposed method for dealing with moving objects is similar to Eren *et. al.*method[9]. We create a mask indicating the moving objects between registered HR training image and each magnified LR frame; and based on this mask, the high-frequencies information of the HR image is fused with LR frame.

The reminder of this paper is organized as follows: in section 2 the proposed method and in section 3 experimental results are provided. The last section describes concluding remarks.

2. The Proposed Method

The method proposed here is an extension of our recent work [8] for video. In [8] we proposed a method for single image super-resolution with mapping a HR training image to a LR image, which summarized as follows:

1) Resizing the LR image, for producing an LR image with desired number of pixels,
2) Finding interest points of this resized LR image and the HR image,
3) Removing outliers and estimating the transformation model,
4) Mapping HR image to LR image,
5) Producing a synthesized HR image with fusion of mapped HR image and resized LR image.

In [8] one LR image was enhanced with some HR training images. But here we have multiple LR frames and one HR training image. Moreover in the previous work we have not any objects. Here we have a video sequence containing moving objects. In addition in [8] because we have not applied a seamless blending approach, the border of the mapped regions are obvious; in this paper a seamless blending method has been used. Figure 2 shows the overall framework of the proposed method.

In the proposed method it is supposed that:

- the HR training image can be transformed with a planar projective model to each LR frame;
- the SIFT key-points of the static parts of scene are influencing the registration model, In the other hand, the moving objects are not so large, which affect the registration.

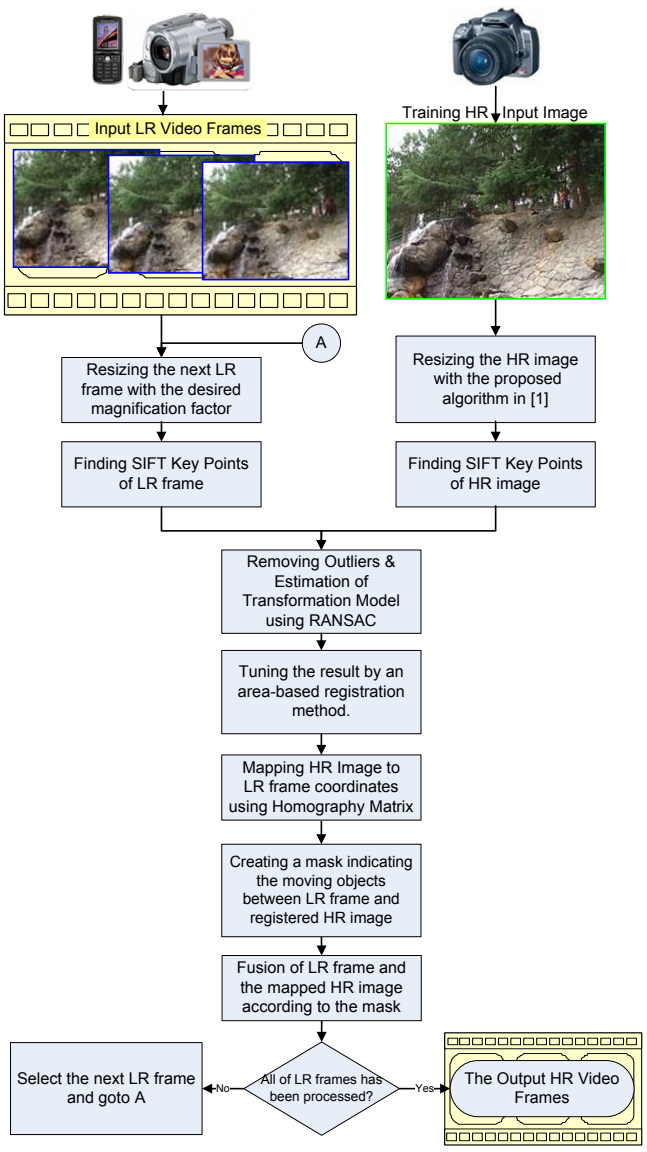


Fig. 2: The overall framework.

- all of the video frames are related to an specific scene; otherwise we need a shot detection algorithm for extracting shots from video and a HR training image for every shot.

The proposed method is shown in algorithm 1. In this algorithm $\mathbf{W}(\mathbf{x};\mathbf{p})$ denote the parameterized set of allowed warps, $\mathbf{p} = (p_1,\ldots,p_n)^T$ is a vector of parameters; $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$ is HR image T warped onto the coordinate frame of the video frame $g^{(i)}$ and $\mathbf{x} = (x,y)^T$ is a column vector containing the pixel coordinates[13]. The warp $\mathbf{W}(\mathbf{x};\mathbf{p})$ takes the pixel $\mathbf{x}$ in the coordinate frame of the HR image T and maps it to the sub-pixel location $\mathbf{W}(\mathbf{x};\mathbf{p})$ in the coordinate frame of the image $g^{(i)}$. The warp model may be any transformation model such as affine, homography or optical flow. But in this paper we concentrated on planar projective model.

It is supposed that magnification factor r is set by user.

Algorithm 1 Video Enhancement using HR images in the Persence of Moving Objects

Require: LR video frames $g^{(1)},\ldots,g^{(n)}$, HR training image T, magnification factor r.
Output: HR video frames $f^{(1)},\ldots,f^{(n)}$.

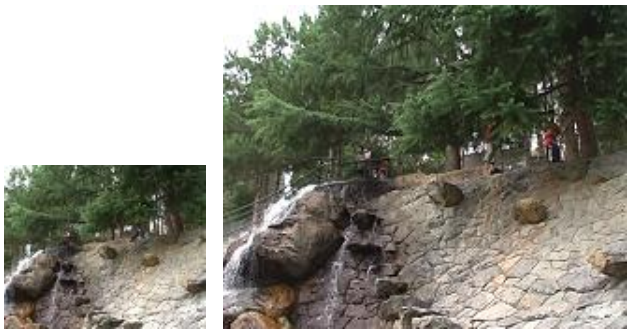
1: Find the SIFT key-points of HR training image.
2: **for** $i = 1$ to n **do**
3: Select the next LR frame $g^{(i)}$,
4: Resize $g^{(i)}$, with magnification factor r, for producing an LR image with desired number of pixels,
5: Find SIFT key-points of this resized LR image,
6: Remove outliers and estimate the transformation model $(\mathbf{W}(\mathbf{x};\mathbf{p}))$,
7: Tune the warp model by Lucas-Kanade registration algorithm [13].
8: Warp T based on $\mathbf{W}(\mathbf{x};\mathbf{p})$ onto coordinate frame of $g^{(i)}$, $(T(\mathbf{W}(\mathbf{x};\mathbf{p})))$,
9: Create mask M with thresholding of subtraction of $g^{(i)}$ and $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$ for detecting moving objects.
10: Produce $f^{(i)}$ by fusion of $g^{(i)}$ and $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$ according to inversion of M with multi-band blending approach [14].
11: **end for**

2.1 Handling Moving Objects

The usual methods for background and foreground detection such as [15], which are based on subtraction technique, can be used here, if we have not illumination changing. Instead of using such methods, we used here a simple subtraction method between each LR frame $(g^{(i)})$ and registered HR training image $(T(\mathbf{W}(\mathbf{x};\mathbf{p})))$. In line 9 of algorithm 1, mask M which illustrates the moving objects is build by thresholding the subtraction image. Lower value of this threshold leads to more sensitive to moving.

2.2 Fusion

For fusion stage of registered HR image $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$ and LR frame $(g^{(i)})$, we used multi-band blending approach [14] as a powerfull image fusion technique. With this fusion method one can determine which regions of each image contributed in the final composite image by a mask. We produce the final HR frame $f^{(i)}$ by compositing the static part of the scene from registered HR image and regions related to moving objects of LR frame $g^{(i)}$. The multi-band blending approach [14] guaranties the smoothness of the transition between this parts, so we have a seam-less result.



(a) First LR frame($g^{(1)}$) (b) HR training image (T)

Fig. 3: First low resolution video frame and the high resolution training image.

(a) $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$ (b) Mask M

Fig. 4: Some Intermediate images, (a) HR training image T warped onto coordinate frame of $g^{(1)}$. (b) Mask M created based on subtraction of 4(a) and resized of 3(a)

3. Experimental Results

We applied our proposed method on a test video frames shown in figure 2 and compared its performance with some other algorithms.

We started from a high resolution video sequence (720×576 pixels), which is captured by Panasonic NV-GS75 handy-cam. In this 69 frames length sequence, two separate sources of motion were present. First, by shaking the camera in hand a motion was created for each individual frame. Second, children movement and waterfall. Each frame is first blurred and next downsampled by a factor two. The resulting sequence is compressed with Microsoft MPEG-4 Video Codec V2. The mentioned process has been done by VirtualDub[1]. This results in blurred and low resolution (360×288 pixels) images that can be used as input for our algorithm.

One of the original HR frames is considered as the HR training image [2].

Figure 3 shows the first LR frame and the HR training image.

[1]http://www.virtualdub.org/

[2]The LR video, training HR image and the output result can be downloaded from the following address:
http://webpages.iust.ac.ir/mamintoosi/DataSets/SR/IPCV09.zip

(a) Nearest (b) Bicubic (c) Freeman[6] (d) This Paper

Fig. 5: Close-up of replication (nearest) bicubic resizing method, Freeman's example-based method [6], and the proposed method for enhancing the first LR frame shown in figure 3(a) using training HR image 3(b).

Figure 4(a) shows $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$, HR image T transformed by appropriate warping model in line 8 of algorithm 1. Figure 4(b) shows the inversion of mask M, lines 9,10 of algorithm 1. The black regions in the mask indicate differences between registered HR image $T(\mathbf{W}(\mathbf{x};\mathbf{p}))$ and LR image:

- The right and bottom side black strips are due to shaking the camera by hand,
- The lower left black regions are due to waterfall,
- The upper left black region is about to trees,
- The middle right black regions are due to children movement, and
- The lower right black regions are not produce by moving objects; these are due to parallax effect.

Figure 5 shows a subjective comparison between different methods on a magnified portion of their results. Note the relative improvement in the quality of the proposed method (5(d)) as a result of fusion of HR training image and LR video frame.

Figures 6 and 7 show quantitative comparisons of the proposed method based on Mean Square Error (MSE) and Image Energy criteria [16]. Although based on the MSE comparison (figure 6) Freeman's method is better than our approach, but inspection of the results shown in figures 5 and 7 indicates that the proposed method is very better than the Freeman method, visually.

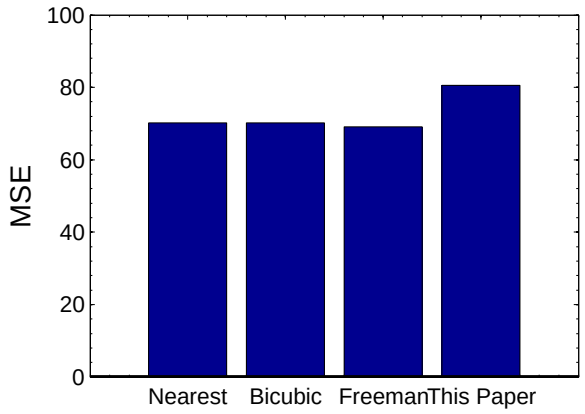


Fig. 6: MSE comparison.

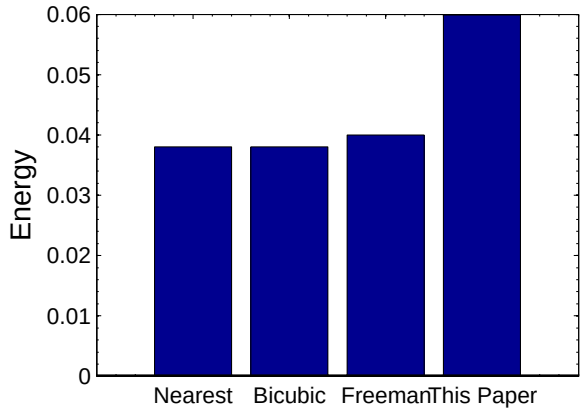


Fig. 7: Energy comparison.

4. Concluding Remarks

In this paper, we proposed a video super-resolution method using high resolution training image from the same scene. We accepted a few differences between video frames and training images including: slightly differ in view point, illumination and zooming differences. The high frequencies details of the input LR image are amplified by fusing a registered version of HR image to it.

The main contributions of the paper are as follows: (i) instead of previous works which used small blocks of training images for amplifying the high frequency information, here we used the whole registered form of the training image and (ii) we detect the regions corresponding to moving objects and remove them in fusion stage.

Although the proposed method does not enhance some parts of each LR frame, but since the most of this regions are due to moving objects, this is not obvious in the output video. Our experiments showed that this algorithm outperforms in quality some other methods.

5. Acknowledgment

The authors are indebted to Dr. Vandewalle [17] for his Super-Resolution package and Dr. D. Lowe for his SIFT key-points program[3]. We also thank to Dr. Peter Kovesi [18] and Dr. Simon Baker and his co-workers [13] for providing many useful MATLAB functions.

References

[1] R. Tsai and T. Huang, "Multiframe image restoration and registration," in *Advances in Computer Vision and Image Processing*, R. Y. Tsai and T. S. Huang, Eds., vol. 1. JAI Press Inc, 1984, pp. 317–339.

[2] S. Borman, "Topics in multiframe superresolution restoration," Ph.D. dissertation, University of Notre Dame, Notre Dame, IN, May 2004.

[3] Z. Lin and H. Shum, "Fundamental limits of reconstruction-based superresolution algorithms under local translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 83–97, 2004.

[4] T. Q. Pham, "Spatiotonal adaptivity in super-resolution of undersampled image sequences," Ph.D. dissertation, aan de Technische Universiteit Delft, 2006.

[5] S. Baker and T. Kanade, "Hallucinating faces," in *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*. Washington, DC, USA: IEEE Computer Society, 2000, p. 83.

[6] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, 2002.

[7] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *SIGGRAPH 2001, Computer Graphics Proceedings*, E. Fiume, Ed. ACM Press / ACM SIGGRAPH, 2001, pp. 327–340.

[8] M. Amintoosi, M. Fathy, and N. Mozayani, "Regional varying image super-resolution," in *IEEE International Joint Conference on Computational Sciences and Optimization*, vol. 1, Sanya, China, April 23-26 2009, pp. 913–917.

[9] P. E. Eren, M. I. Sezan, and A. Tekalp, "Robust, object- based high-resolution image reconstruction from low-resolution video," *IEEE Trans. Image Processing*, vol. 6, no. 10, pp. 1446–1451, 1997.

[10] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust super resolution," in *Proceedings of the Int. Conf. on Computer Vision and Patern Recognition (CVPR)*, Dec 2001, pp. 645–650.

[11] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[12] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to super-resolution," in *Proc. of the 2003 SPIE Conf. on Applications of Digital Signal and Image Processing*, Aug 2003, pp. 121–130.

[13] S. Baker, R. Gross, and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, pp. 221–255, 2004.

[14] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Trans. Graph.*, vol. 2, no. 4, pp. 217–236, 1983.

[15] M. Amintoosi, F. Farbiz, and M. Fathy, "A QR Decomposition based mixture model algorithm for background modeling," in *ICICS2007, Sixth International Conference on Information, Communication and Signal Processing*, Singapore, December 2007, pp. 1–5.

[16] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, p. 10, 2007.

[17] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A Frequency Domain Approach to Registration of Aliased Images with Application to Super-Resolution," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 71 459, 14 pages, 2006.

[18] P. D. Kovesi, "MATLAB and Octave functions for computer vision and image processing," School of Computer Science & Software Engineering, The University of Western Australia, available from: http://www.csse.uwa.edu.au/∼pk/research/matlabfns/.

[3] Available online at: http://www.cs.ubc.ca/∼lowe/keypoints/