Spatial Resolution Enhancement of Video Using Still Images

Guodong Guo
Charles Dyer

Technical Report #1502

October 2004

Spatial Resolution Enhancement of Video Using Still Images

Guodong Guo and Charles R. Dyer
Computer Sciences Department
University of Wisconsin-Madison

April 2, 2004

Abstract

Images captured by digital video cameras usually have lower spatial resolution than digital still cameras. This paper addresses the problem of combining images from digital still cameras and video cameras to generate a video sequence with higher resolution than the original video. A method is presented for accomplishing this goal and experimental results are shown that demonstrate its effectiveness.

1. Introduction

Visual information includes the dimensions of space, time, spectrum, and brightness [12]. However, a camera cannot capture all this information simultaneously. As a result, there are always trade-offs between the dimensions. For example, color cameras trade-off spatial resolution [12]. Among the multiple dimensions of images we are interested in the space-time interaction.

Digital still cameras capture the world at 5-10 times the spatial resolution of digital video cameras, while video cameras have denser temporal sampling. For example, the Kodak DCS-760 professional digital still camera has a resolution of 3032 $\times$ 2008 (6 megapixels), while the JVC JY-HD10U (high definition) digital video camera records frames of size 1280 $\times$ 720 (0.9 megapixels). For consumer products, 5 megapixel digital cameras (e.g. Canon Powershot G5) are common today, while most digital camcorders have 640 $\times$ 480 resolution (0.4 megapixels).

Why do digital still cameras and camcorders have such different spatial resolutions? One reason is the physical restriction. Charge-coupled devices (CCDs) are the most common image sensors used in digital cameras [4]. CCDs capture light in small photosites on their surface and the charge is read after an exposure. For example, charges on the last row are transferred to a read-out register. From there, the signals are fed to an amplifier and then on to an analog-to-digital converter. Once the row has been read, its charges in the read-out register row are deleted, the next row enters the read-out register, and all of the rows above march down one row. The charges on each row are "coupled" to those on the row above so when one moves down, the next moves down to fill its old space. In this way, each row can be read, one row at a time, as shown in Fig. 1. In digital video cameras, to capture 25 or more frames per second, there are a large quantity of charges to transfer per second. In order to keep the temporal sampling rate, the number of charges used for each frame has to be small enough. This is a space-time tradeoff.

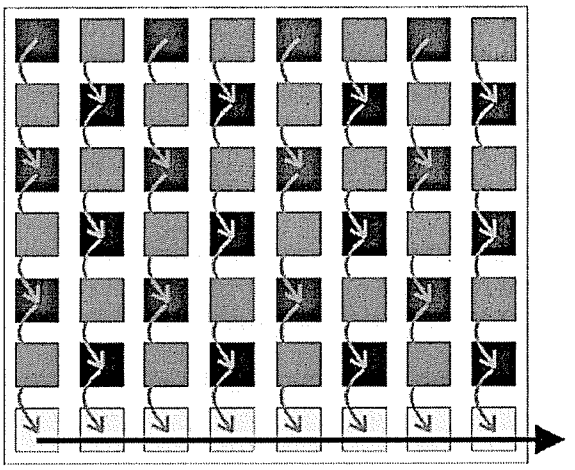

Figure 1: The CCD shifts one whole row at a time into the read-out register. The read-out register then shifts one pixel at a time to the output amplifier [4].

One way to break through this physical restriction is to use multiple cameras such as both digital still cameras and digital camcorders. Then combine the information from both kinds of cameras to enrich each other.

In practice, one may not need two or more cameras in order to reach this goal. Nowadays, many of the digital still cameras can capture short video segments and more and more digital camcorders can capture digital stills. Because of this property, one can use, for example, a single digital camera to capture high quality digital stills and low-resolution video sequences.

In this paper we consider the goal of combining the best qualities of each type of camera. Specifically, using high

resolution still images to enhance the spatial resolution of a video sequence. The framework of the approach is shown in Fig. 2. This problem is related to, but different from, existing super-resolution work that is based on signal reconstruction or example-based learning. In reconstruction-based super-resolution [10] [6] [16] [15] [3], multiple low-resolution images are registered to create a higher resolution image. See a review of classical approaches to super-resolution image reconstruction in [2]. In learning methods [8] [1], images and their size-reduced images are used as training pairs to learn high frequency information. Other recent work [13] aligns video sequences to increase resolution by assuming the video cameras have the same optical center.

We present a recognition-based scheme to align high-resolution images with video sequences in Section 2, and robustly estimate the mapping between the images and videos in Section 3. Then we describe a factorization technique to rotate and correct the high-resolution images in Sections 4 and 5. Experimental results are shown in Section 6 and further issues are discussed in Section 7.

2. Image and Video Alignment via Recognition

In order to use high-resolution still images to enhance low-resolution video frames, one has to first establish the relationship between them. That is, align or register the images coming from different sources.

Video registration is a challenging problem [14]. Because of camera motion, the viewpoints of a video sequence may change continuously and be different from the digital still images' viewpoints. Furthermore, the illumination and camera automatic gain may also change. However, the biggest variation in our problem is the difference in resolution.

If two images to be matched have very different resolutions in addition to viewpoint and illumination changes, traditional direct methods using optical flow or local feature (e.g. corner) matching cannot be used because these features are used under the assumption that local image patches between two images do not change significantly in appearance. These features especially lack invariance to scale [11]. For example, corner features are usually computed using the same template size for two images to be matched. When two images have very different scales, the computed values will be different in the two images. In order to align still images with video sequences, we have to find some new matching techniques.

One possible way to deal with image matching with very different scales is to formulate it as a one-to-many matching problem [5]. The high-resolution image is size-reduced by various scales and some local features are extracted at each scale. Another way is to extract scale-invariant features. Lowe [11] proposed a scale-invariant feature transform (SIFT) operator and used it successfully for object recognition. Using the SIFT operator, scale information is automatically encoded in each extracted key point, and there is no need to extract features at various scales of the image. Here, we use SIFT feature matching as the first step for our super-resolution method, and show that the SIFT operator can deal with large resolution differences.

The SIFT operator first identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations by blurring image gradient locations. The resulting feature vectors are called SIFT keys. A nearest neighbor criterion is then used to find similar keys in both images. For more details on the SIFT operator, see [11].

3. Homography Estimation

After using the SIFT operator for feature extraction and the nearest-neighbor criterion for feature matching, there are usually a large number of incorrect feature correspondences. Robust methods such as RANSAC [7] [9] can be used to remove outlier matches and estimate the homography between the two images.

There are three cases in which a planar homography is appropriate [3] [9]: (1) images of a plane viewed under arbitrary camera motion, (2) images of an arbitrary 3D scene viewed by a camera rotating about its optical center and/or zooming, and (3) a freely moving camera viewing a very distant scene. To demonstrate our approach, in this paper we assume the scene is planar and so a planar homography is sufficient to describe the relation between a high-resolution image and a low-resolution image.

4. Making Image Planes Parallel

Assume $\mathbf{q} = H\mathbf{p}$, where $\mathbf{p} = (x, y, w)^T$ are the homogeneous coordinates of a point in the low-resolution image, and $\mathbf{q}$ is the corresponding point in the high-resolution image. H is a 3×3 matrix, mapping the low-resolution image to the high-resolution image. For super-resolution purposes, knowing only the mapping H is not enough. The goal is to obtain an image pattern in a high-resolution image with the same viewpoint and illumination as that in the low-resolution image, mimicking a virtual camera with only a spatial scale difference.

To accomplish this, the high-resolution image must first be rotated so that it is parallel to the low-resolution image, as shown in Fig. 3 where the high-resolution image B is rotated into B' so that B' is parallel to the low-resolution im-

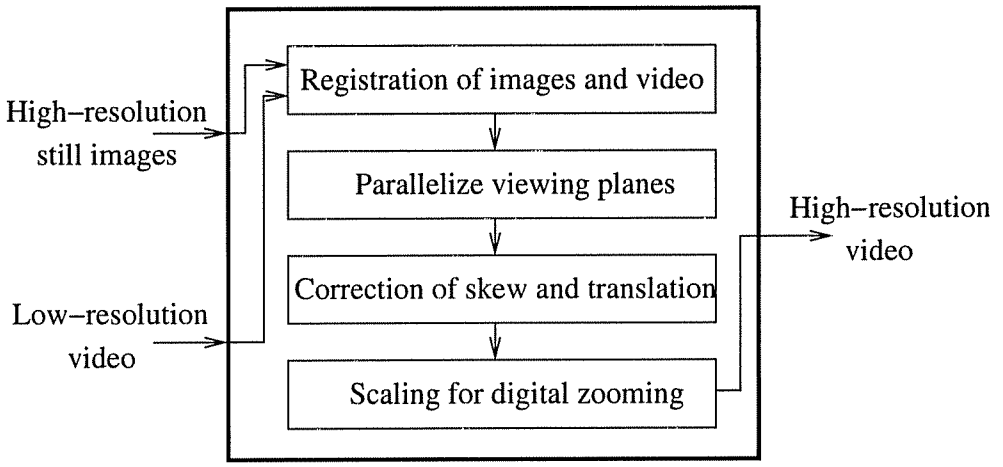

Figure 2: The framework of our approach. More details on each step can be found in the text.

age S. We use QR decomposition to estimate the required rotation.

4.1. QR Factorization

The 3×3 homography matrix H can be decomposed into two matrices via QR factorization,

$$H = R_1 U_1 \tag{1}$$

where R_1 is a rotation matrix, and U_1 is an upper triangular matrix. Then the inverse, H^{-1}, is defined as

$$H^{-1} = (R_1 U_1)^{-1} = U_1^{-1} R_1^{-1} = U_2 R_2 \tag{2}$$

where $R_2 = R_1^{-1}$ is also a rotation matrix, and $U_2 = U_1^{-1}$ is another upper triangular matrix.

From $\mathbf{p} = H^{-1}\mathbf{q}$ and Eq. (2), we get

$$\mathbf{p} = U_2 R_2 \mathbf{q} = U_2 \mathbf{q}' \tag{3}$$

where $\mathbf{q}' = R_2\mathbf{q}$ is the corresponding point in the rotated high-resolution image plane that is parallel to the low-resolution image frame. Point $\mathbf{p}$ in the low resolution image is mapped to point $\mathbf{q}'$ by

$$\mathbf{q}' = U_2^{-1}\mathbf{p} \tag{4}$$

and U_2^{-1} has the form

$$U_2^{-1} = \begin{bmatrix} \alpha_x & s & t_x \\ 0 & \alpha_y & t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

where s is the skew, α_x, α_y are scale factors in the x and y directions respectively, and t_x and t_y are translations. In practice, the skew, s, may or may not be zero. If $s \neq 0$, we need to decompose U_2^{-1} further by

$$U_2^{-1} = \begin{bmatrix} \alpha_x & 0 & t_x \\ 0 & \alpha_y & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{s}{\alpha_x} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = T_{st} T_k \tag{6}$$

where T_k is the skew transform matrix, and T_{st} is the transform of scale and translation. For the purpose of analyzing resolution difference, it is better to further decompose T_{st} as

$$T_{st} = \begin{bmatrix} \alpha_x & 0 & 0 \\ 0 & \alpha_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \frac{t_x}{\alpha_x} \\ 0 & 1 & \frac{t_y}{\alpha_y} \\ 0 & 0 & 1 \end{bmatrix} = T_s T_t \tag{7}$$

so we have $U_2^{-1} = T_s T_t T_k$. Letting $T_h = T_t T_k R_2$, one can apply T_h to the high resolution image by

$$\mathbf{q}'' = T_{tk}\mathbf{q}' = T_h \mathbf{q} \tag{8}$$

and apply T_s^{-1} to the low resolution image by

$$T_s^{-1}\mathbf{p} = \mathbf{q}'' \tag{9}$$

Eq. (8) warps the high-resolution image so that it is parallel to the low-resolution frame and has no skew or translation difference. The remaining difference between $\mathbf{q}''$ and $\mathbf{p}$ is just the scale factor, which is encoded in T_s. Eq. (9) is used to scale the low-resolution image and find the corresponding position in the rotated, skew-corrected, and translation-corrected high-resolution image for any point $\mathbf{p}$. Note that there is only a scale transformation, T_s^{-1}, between $\mathbf{p}$ and $\mathbf{q}''$. To summarize, all mappings are shown in Fig. 4.

4.2. Scale Coherence in Two Directions

The pixels in images may be square or non-square, which is determined by the physical CCDs. The *pixel aspect ratio*

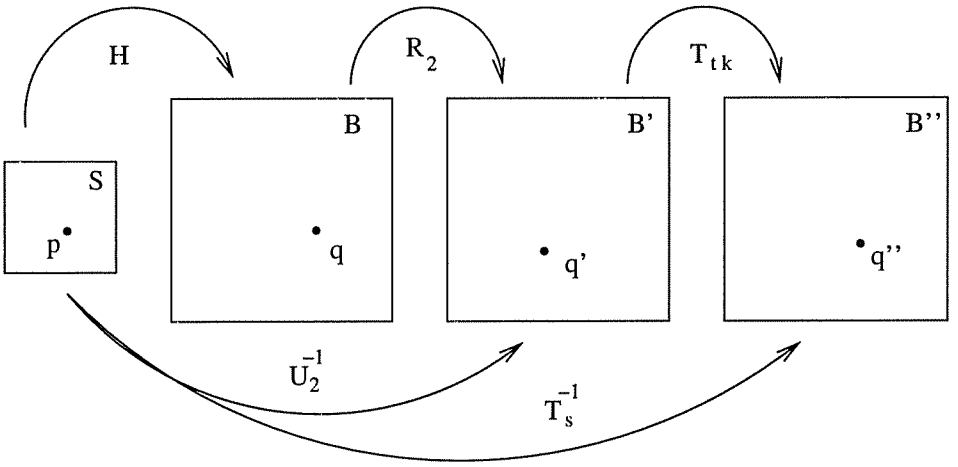

Figure 4: The relation between the low-resolution input image S, high-resolution input image B, rotated image B', and skew and translation corrected image B''. $\mathbf{p}$, $\mathbf{q}$, $\mathbf{q}'$, and $\mathbf{q}''$ are corresponding points in each image.

(AR) is the ratio of horizontal and vertical sizes of a pixel. This term also refers to an image's display resolution. For instance, an image with a 640×480 resolution has an aspect ratio of 4:3, while a 720×480 resolution has an AR of 3:2. The standard aspect ratio for traditional television sets and computer monitors is 4:3 while the aspect ratio for high-definition, wide-screen digital systems is 16:9. In our super-resolution work, the high-resolution still images may have a different AR than the low-resolution video frames when two different cameras are used. Different ARs may result in different scale factors in the x and y directions, i.e., $\alpha_x \neq \alpha_y$ in Eqs. (5) (6) and (7). While the goal is to enhance the spatial resolution of each video frame, it is not a good idea to change the aspect ratio of the low-resolution frames after enhancement. To avoid this, the two scale factors, α_x and α_y, should be normalized to a common value, analogous to digitally zooming the low-resolution images by a given percentage. Assuming $\alpha_x > \alpha_y$, T_s can be decomposed as

$$T_s = \begin{bmatrix} \alpha_x & 0 & 0 \\ 0 & \alpha_x & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\alpha_y}{\alpha_x} & 0 \\ 0 & 0 & 1 \end{bmatrix} = T_{ss} T_{sc} \quad (10)$$

Let $T_h^{'} = T_{sc} T_t T_k R_2$ and apply it to the high-resolution image, and only apply T_{ss}^{-1} to the low-resolution images. The scale factor between the low-resolution and high-resolution images is equal to the first element of T_{ss}^{-1}, i.e., $T_{ss}^{-1}(1,1)$, assuming the last element, $T_{ss}^{-1}(3,3)$, equals 1.

In practice, even if the aspect ratios of the two cameras are the same, or only one digital camera is used to capture both the high-resolution still images and low-resolution videos, the estimated scale factors, α_x and α_y, may still be different because of the image and video registration accuracy, and possibly the manufacturing precision. So, normalize the scale factors α_x and α_y to a common value in all cases.

4.3. Non-Uniqueness

QR decomposition is not unique. Thus when we use the computed R to warp the high-resolution image, it may result in an "invalid" rotation (e.g., the rotated points have negative coordinates). To prove the non-uniqueness of QR decomposition, let $H = RU = (RD)(D^{-1}U) = R'U'$, given that D is orthogonal with determinant 1 and $D \neq I$. Since both R and D are orthonormal, RD is also orthonormal, and $D^{-1}U$ is upper triangular.

In practice, we can check if α_x and α_y (in Eq. (6)) are both negative. If yes, we can choose

$$D = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (11)$$

and use $H = R'U'$ instead of RU. Note that α_x and α_y can not have different signs because we cannot capture an image with positive scale in one dimension and negative scale in the other.

5. Photometric Correction

Besides the geometrical differences between the low and high resolution images, there may also be differences in the intensities between the images because of global illumination variation and/or camera automatic gain changes. To cope with photometric variation, we use a simple linear method to align the intensities of the warped high resolution image with the low resolution image,

$$E_{new} = \frac{E - B''_{min}}{B''_{max} - B''_{min}} (S_{max} - S_{min}) + S_{min} \quad (12)$$

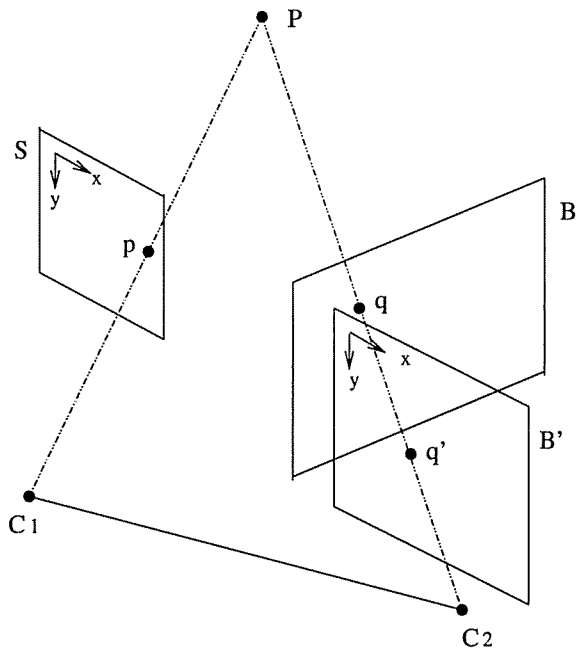

Figure 3: Two cameras (with centers $C1$ and $C2$ respectively) are used to capture the low-resolution image S and high-resolution image B which is rotated into B' so that the viewing plane B' is parallel to S. Note that this rotation is different from the traditional image rectification in stereo where both images are warped into parallel to the baseline $C1C2$.

where B''_{max} and B''_{min} are the maximum and minimum intensities in a region in the warped high-resolution image, S_{max} and S_{min} are the maximum and minimum intensities in the corresponding region in the low-resolution image, E is the given pixel's intensity in B'', and E_{new} is the photometrically-corrected value. Eq. (12) is applied for each pixel in each color channel separately.

The whole procedure presented in Sections 2 to 5 can be applied to each frame of the video sequence using each high-resolution still image.

6. Experiments

A Canon PowerShot A70 digital camera was used to capture both the high-resolution still images (of size 2048×1536) using the "auto mode," and the video sequences (each frame of size 320×240) with the "video mode." The scene is a rug containing many details. For display purposes only, the still images were reduced to 1280×960, which has no influence on demonstrating the basic idea.

In Fig. 5 one image extracted from the video sequence is shown at the top-left, and one high-resolution image is shown in the middle. Using the SIFT operator for feature detection, 5,834 points were extracted from the high-resolution image, and 1,457 points from the low-resolution image. Using nearest neighbor matching, 471 correspondences were found. However, there are many outliers (i.e., mismatches) there. Using RANSAC to estimate the homography, 173 inliers were selected, from which only 30 are displayed in both images (top-right and middle in Fig. 5) to avoid confusion in this visualization. The condition number of the 3×3 homography matrix H is large, but the estimate is accurate. We also used the normalization approach, but it did not improve significantly the results. QR factorization and related manipulations were performed, Eq. (8) was used to warp the high-resolution image parallel to the low-resolution image frame and to correct skew and translation. Eq. (9) was used to zoom in the low-resolution image. The scales were estimated using Eq. (10) and the scales in the x and y directions are the same without changing the aspect ratio of the low-resolution images. Photometric correction using Eq. (12) was then done. For the low-resolution image shown at the top-left in Fig. 5, its enhanced high-resolution image (of size 1392×1044) is shown at the bottom. The estimated scale difference is 4.35, which is bigger than the image size difference (four times in each direction) between the input high-resolution image (1280×960, middle in Fig. 5) and the low-resolution image (320×240).

To see the result clearly, it is better to look closely at some selected regions in the images. A 100×100 window was cropped from the low-resolution image (at the top-right in Fig. 5) and shown in the top of Fig. 6. The small patch was re-scaled using bilinear interpolation (middle left) and bicubic interpolation (middle right) as shown in Fig. 6. Clearly, many details were lost and the image patch looks vague. Image interpolation does not add new information although the image size is bigger. The corresponding patch in the warped high resolution image is cropped and shown at the bottom-left in Fig. 6, which is much clearer. The flowers in the middle and the stripes at bottom-left can be seen clearly. Finally, photometric correction using Eq. (12) was performed and the new image is shown at the bottom-right in Fig. 6. From this experimental result we can see that the low-resolution image can be greatly enriched using the information from the input high-resolution image.

The input video used in the experiment contained 90 frames. The same procedure described above was executed independently for all frames in the video. The low-

resolution input video and high-resolution output video are not shown here. Instead, they will be shown at the authors' web pages. The experiments demonstrate that our approach for still-image-based video enhancement is promising.

7. Discussion

We have demonstrated an approach for using high-resolution digital still images to enhance low-resolution video sequences. There are several questions remaining to be answered: 1) How many high-resolution images are needed? Currently, we only use one high-resolution image to enhance the whole video sequence. Some regions in the low-resolution images cannot be "enhanced" because the corresponding parts do not exist in the high-resolution image. Hence more high-resolution images may be necessary. 2) How far apart can the viewpoints be when capturing the videos and high-resolution images? If they are too far apart, there will be distortions when warping the images. 3) How should the high-resolution images for a more general, non-planar, scene be warped? In our experiments, we assumed a 3×3 homography, which is not general enough to deal with all possible scenes. 4) How should photometric correction be done for more complex illumination conditions? We believe that all these problems deserve investigation based on the results here.

8. Conclusion

We have proposed enhancing the spatial resolution of video sequences using higher resolution digital still images. A recognition-based method using invariant features is presented to register the high-resolution images with the low-resolution video sequences. A simple, robust method based on QR factorization is used to warp the high-resolution images in order to mimic a digital "zooming" effect. The procedure realizes the basic idea of our still-image-based video enhancement framework. Many extensions of the method are possible in order to build a real system for practical use.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. In *IEEE PAMI*, volume 24, pages 1167–1183, 2002.

[2] S. Borman and R. L. Stevenson. Super-Resolution from Image Sequences - A Review. In *Proceedings of the 1998 Midwest Symposium on Circuits and Systems*, 1998.

[3] D. Capel and A. Zisserman. Computer vision applied to super resolution. In *IEEE Signal Processing Magazine*, pages 75–86, 2003.

[4] D. P. Curtin. *The Textbook of Digital Photography*. http://www.shortcourses.com/, 2003.

[5] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Proc. CVPR*, pages 612–618, 2000.

[6] M. Elad and A. Feuer. uper-resolution reconstruction of image sequences. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 21, pages 817–834, 1999.

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting wth application to image analysis and automated cartography. In *Comm. ACM*, volume 24, pages 381–395, 1981.

[8] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. In *IJCV*, volume 40, pages 25–47, 2000.

[9] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[10] M. Irani and S. Peleg. Improving resolution by image registration. In *Graphical Models and Image Processing*, volume 53, pages 231–139, 1991.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.

[12] S. K. Nayar and S. G. Narasimhan. Assorted pixels: Multi-sampled imaging with structural models. In *Proc. ECCV*, volume 3, pages 148–162, 2002.

[13] E. Shechtman, Y. Caspi, and M. Irani. Increasing space-time resolution in video. In *Proc. ECCV*, pages 753–768, 2002.

[14] R. Szeliski. Video registration: Key challenges. In M. Shah and R. Kumar, editors, *Video Registration*, pages 247–252, Boston, 2003. Kluwer Academic Publishers.

[15] M. Tipping and C. Bishop. Bayesian image super-resolution. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1303–1310, 2003.

[16] A. Zomet and S. Peleg. Super-resolution from multiple images having arbitrary mutual motion. In S. Chaudhuri, editor, *Super-Resolution Imaging*, pages 195–209. Kluwer Academic, 2001.

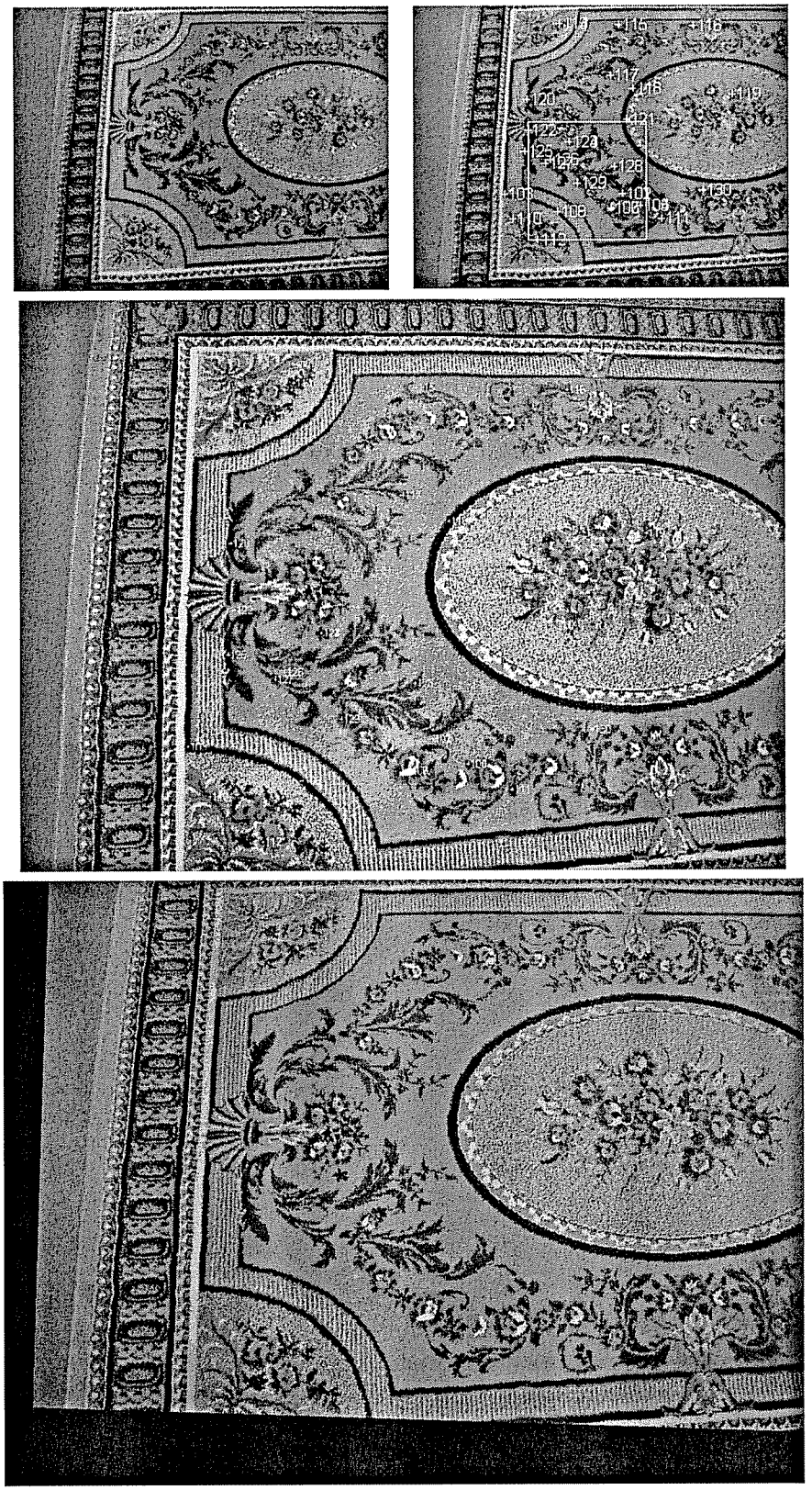

Figure 5: Top Left: One frame from a video sequence of size 320 × 240; Top right: partial features detected by the SIFT operator, with the square of size 100 × 100; Middle: A high resolution still image of size 1280 × 960, 16 times larger than the video frame. Some corresponding points are labelled in this image. Bottom: The resolution-enhanced image of size 1392 × 1044. The black region is because there is no corresponding information to obtain from the input high-resolution image.

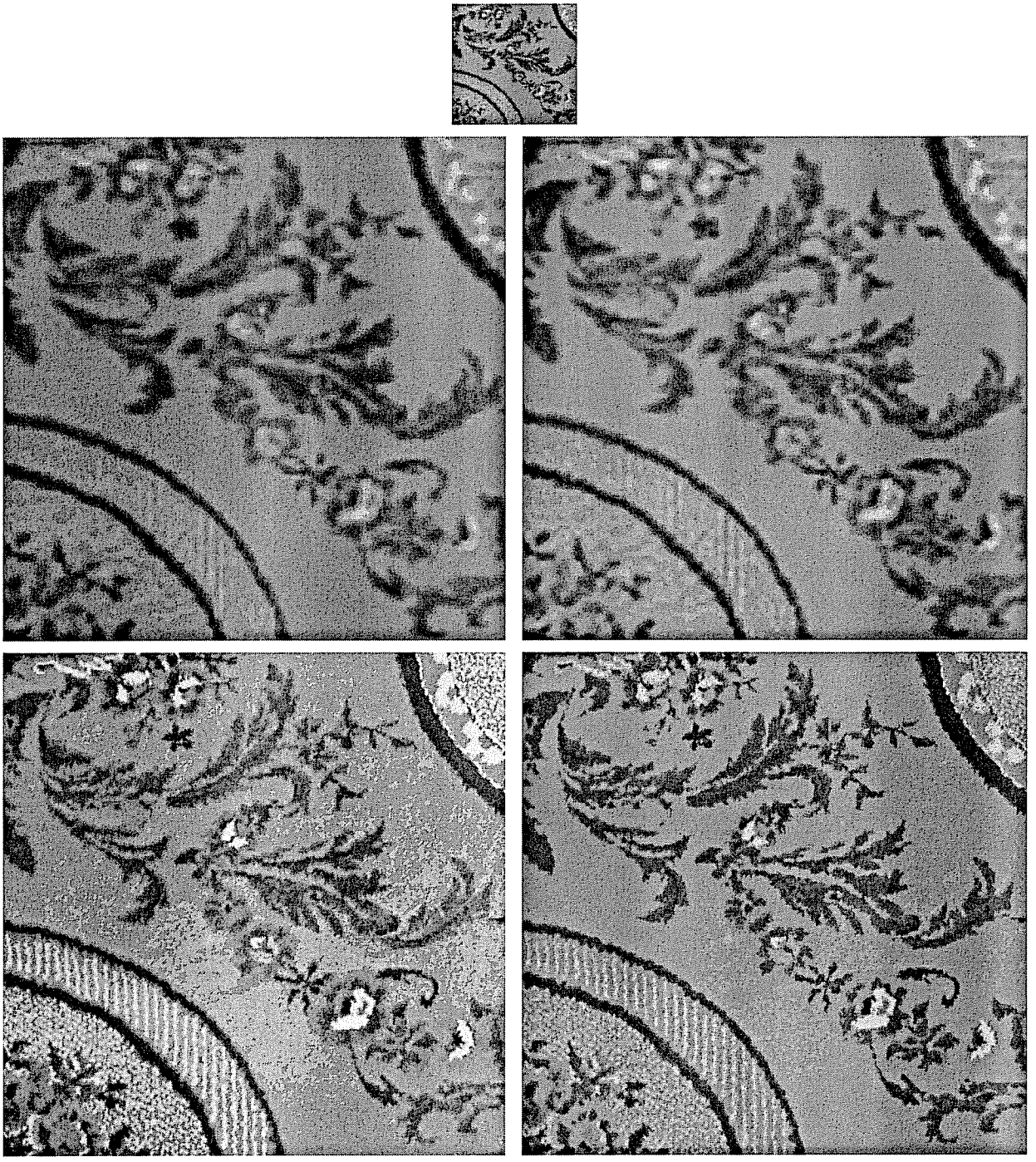

Figure 6: Top row: The image block of size 100×100 cropped from the square shown in the top right image of Fig. 5; Middle-left: Cropped square enlarged using bilinear interpolation with the estimated scale 4.35; Middle-right: Enlarged using bicubic interpolation; Bottom-left: Corresponding high resolution block extracted and warped from the bottom image in Fig. 5; Bottom-right: Photometrically corrected image of the bottom-left image.