16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

# FAST SUPER-RESOLUTION ON MOVING OBJECTS IN VIDEO SEQUENCES

*Antoine Létienne*[*], *Frédéric Champagnat*[*], *Caroline Kulcsár*[†], *Guy Le Besnerais*[*], *Patrick Viaris De Lesegno*[†]

([*]) Office National d'Etudes et de Recherches Aérospatiales (ONERA), DTIM/EVS
BP-72, 92322 Chatillon Cedex, France
*aletienn@onera.fr, fchamp@onera.fr, lebesner@onera.fr*
([†]) Laboratoire de Traitement et Transport de l'information, Université Paris 13
99 av. J.-B.Clément 93430, Villetaneuse, France
*kulcsar@l2ti.univ-paris13.fr, patrick.viaris@l2ti.univ-paris13.fr*

## ABSTRACT

We present a super-resolution (SR) color freeze frame of small moving objects in video sequences. In the last two decades, all SR methods except one focus on the case of rigid scene. We propose a fast and robust method that performs SR reconstruction on tracked objects. After an affine registration of the regions of interest of objects, a non-uniform interpolation in a high-resolution grid is performed, then, a restoration. Experiments on real data validate our choices, and confirm the robustness and the rapidity of our method.

## 1. INTRODUCTION

Super-resolution (SR) aims at improving the quality of an image by exploiting variations between images due to relative scene/sensor motion. SR is interesting in applications where the nominal image resolution is relatively low with respect to the expected use, such as medical imaging, video sequences processing in video surveillance or aerial observation using low-cost sensors.

In this paper, we deal with video sequences, possibly interlaced, from wide-field cameras like those used for video surveillance. We present a fast SR color freeze frame of small moving objects in mobile or still camera in applications such as automatic annotation of tracked objects. Our contribution is three-fold. First, we tackle the case of moving objects in SR. This case has previously been considered only once in [1], whereas most of the literature focuses on the case of rigid scene and moving camera [2]. Second, our method is tailored for fast processing and we are able to compute one SR still per second on a 2 GHz dual core PC. Third, our method is robust and tuned on real data.

On input, our SR reconstruction takes bounding boxes and segmentation masks of the region of interest (ROI) extracted from a video and containing objects to annotate. On output, we compute SR stills of the moving objects, each still using between ten and twenty successive ROI. As for each object a still is computed every second or so, the processing must be fast. In this context, we are interested in robust approaches with a good trade-off between image quality and computational cost.

The reconstruction step of the classical SR corresponds to the inversion of a large sparse matrix. In the translation motion case this inversion can be carried out using fast algorithms [3][4]. When the motion model is not a translation, this inversion can be performed only by iterative and costly techniques [5][6]. In order to deal with non-translation motion with fast algorithms, we opt for the alternative SR approach based on non uniform interpolation [7][8].

A flow diagram of the proposed approach is depicted in Fig. 1. The block **Detection & Tracking** extracts ROIs from low-resolution (LR) frames, giving the extracted objects. For this task we use the tracker provided within the VIVID project [9]. The SR block is built around several blocks. First, the block **Registration and Frame Selection** provides a subpixel registration between the LR ROIs and the reference ROI. The blocks **Non-uniform interpolation and Restoration** constitute the SR reconstruction step. A non-uniform interpolation provides a high-resolution (HR) grid. Then the SR frame is obtained after a restoration step.

Note that in Fig. 1 registration is embedded in the SR step, in contrast with the majority of the contributions in the litterature [2]. Indeed, registration and SR reconstruction are both considered in the present paper, as well as practical issues related to parameter tuning for both tasks.

Our method is validated on real data. We have observed empirically that the translation motion model was limiting the results quality. Moreover, registration turns out to be the longest step in the overall process.

Related works on registration methods and SR techniques are presented in section 2. The proposed SR process is described in section 3. The importance of the registration is shown in 3.1, in particular the necessity to choose a non-translational motion model. In section 4, we demonstrate the efficency of the method on real data from urban video surveillance. Particular attention is paid on the different trade-offs related to parameter tuning.

## 2. RELATED WORK

As far as we know, reference [1] is the only contribution on SR reconstruction of moving objects. Moreover, the present paper is the first one to report results on real data, whereas [1] is limited to synthetic data. The classical SR approach deals with the static background/ moving sensor context, and registration is overlooked most of the time. Although we consider that the two problems are related, they are presented separately in this section.

The input of our SR reconstruction is made up of the set of bounding boxes and segmentation masks provided by the tools developped within the VIVID project [9]. This process is robust, but the registration accuracy is not sufficient for SR reconstruction [2][6][5]. A registration step is therefore performed to reach the required accuracy.

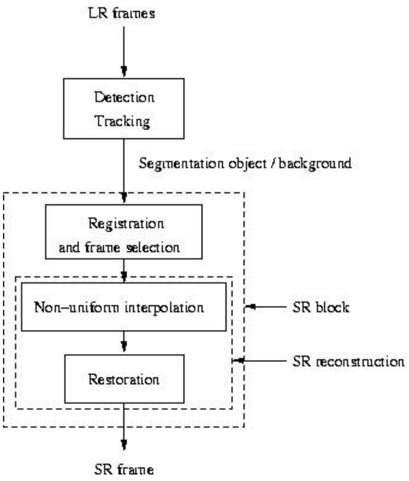

Figure 1: Flow diagram of the proposed SR reconstruction method.

## 2.1 Registration techniques

We have examined several registration techniques in order to select fast algorithms. We have observed experimentally that a translational model is not appropriate even on short time for modeling small 3D effects. Algorithms are broken down into two large families: parametric and non-parametric. Parametric methods are themselves divided into two classes: direct and feature-based. On one hand, one finds methods using a parametric model, such as translational, affine and homographic models. In this category, direct methods [10][11] minimize iteratively a criterion on pixel intensity. Feature-based methods [12] are based on matching Harris-type characteristic points and estimate model parameters by robust fit. On the other hand, non-parametric optical flow – dense – methods [13] provide a motion vector for each pixel. Feature-based techniques are fast, but fail to provide accurate results when the number of features is too low: this is often the case for small objects such as motorbikes, see Fig. 5 . The advantage of dense methods are their ability to register complex motion, but the computational cost is too high for fast implementation. In order to realize a trade-off between the processing speed and results quality, we choose a direct parametric method described in section 3.1.2.

## 2.2 SR reconstruction techniques

The classical SR approach [2][6][14] can be described as follows. After the registration step, a large size linear system is built, then regularized inversion is performed. The translational motion case may yield fast implementations, such as [3] which takes advantage of the FFT.

Non-uniform interpolation [2] is a more intuitive direct method, it is used in the case of the translational model [15][16] and also for more general motion models [7][17][8]. LR images are registered to form an irregular sampling of the scene. An iterative or direct non-uniform interpolation process provides an estimation of the intensity at any HR grid point [7][17]. Then restoration step is applied to reduce noise and blur effects.

Using this method with non-translational motions implies an approximation. As explained in [14], this approximation is valid only in the case of the rotation motion and isotropic sensor point-spread function (PSF). Thus, the optimality of the overall reconstruction algorithm is often not guaranteed.

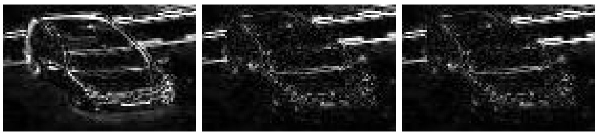

Figure 2: Registration residual for translation (left), affine (middle) and homographic (right) models. The mean-square error is equal to 26.1 for the translation, 19.5 for the affinity and 19.5 for the homography.

# 3. THE PROPOSED METHOD

## 3.1 Registration

The principles of our SR reconstruction method are applied separately on each RGB channel, then outputs are merged to get a color SR image. The registration block does not need color images and works with grayscale images.

### 3.1.1 *Motion model*

We have considered three motion models, translation, affine and homographic. Except in case of fronto-parallel motion, translational model does not offer a good accuracy, as can be seen in Fig. 2. Conversely on this example, affine and homographic models provide comparable performances. But fitting an homographic model requires more computational burden and is less stable than fitting an affine model.

As we deal with small objects in short sequences, the affine model seems the most appropriate while meeting the strong requirement of a fast registration. This choice is confirmed in section 4.

### 3.1.2 *Registration*

Let $I_k$ denote the ROI of frame $k$, to be registred w.r.t. the ROI $I$ of the reference frame. We implement the inverse compositional (IC) method of [18], based on the minimization of the criterion:

$$\sum_{\boldsymbol{x}\in\mathscr{M}} (I(\boldsymbol{x}) - I_k(W(\boldsymbol{x};\boldsymbol{p}))^2, \tag{1}$$

where $\boldsymbol{x}$ indexes a pixel of $I$ in the mask $\mathscr{M}$, $W(\boldsymbol{x};\boldsymbol{p})$ is the geometric warp and $\boldsymbol{p}$ is the coefficients of the transformation $W$. $I_k(W(\boldsymbol{x};\boldsymbol{p}))$ is the value of $I_k$ interpolated at point $W(\boldsymbol{x};\boldsymbol{p})$. The use of a segmentation mask on the object reduces the number of pixels to process, and makes the results more robust.

The classical Lucas-Kanade technique [11] is a Gauss-Newton (GN) descent on (1). The gradient of (1), involves in particular $I_k(W(\mathbf{x};\mathbf{p}))$ and $\frac{\partial}{\partial p}(I_k(W(\mathbf{x};\mathbf{p})))$, which vary at each iteration. The IC algorithm proposed by [18] rely on GN descent of the following sequence of criteria:

$$\sum_{\boldsymbol{x}\in\mathscr{M}} (I(W(\boldsymbol{x};\Delta\boldsymbol{p})) - I_k(W(\boldsymbol{x},\boldsymbol{p}^n)))^2, \tag{2}$$

where $\boldsymbol{p}^n$ is the estimate at iteration $n$, and $\Delta\boldsymbol{p}$ is the step increment of the parameter $\boldsymbol{p}$.

The gradient of (2) involves $I_k(W(\boldsymbol{x};\boldsymbol{p}^n))$ and $\frac{\partial}{\partial\Delta\boldsymbol{p}}(I(W(\boldsymbol{x};\Delta\boldsymbol{p}))$ computed at $\Delta\boldsymbol{p}=0$. The last quantity does not depend on $n$ and $I_k$, it is computed only once for all $I_k$ during initialization. Therefore IC approach is particularly attractive when multiple frames need to be registred w.r.t. a unique frame.

To limit the time for the registration step, the number of iterations must be controlled. We watch the evolution of maximum shift of the four corners of the ROI to register. Indeed, all pixels of the ROI have smaller motion than this maximum translation. Thus, the iterations are stopped as soon as the maximum shift falls below the threshold $RegThresh$ or when it exceeds the maximum number of iterations $maxIter$. Typical values of $RegThresh$ and $maxIter$ are 0.4 pixel and 5, respectively, and this saves about 50% iterations.

### 3.2 Thresholding on the registration

A wrong registration is very penalizing for the SR reconstruction, hence it is necessary to discard the uncorrectly registered images. Meaningful image are selected according to the registration quality. We perform a "$k\sigma$" adaptive threshold the registration error: $\varepsilon_i < k\sigma + \bar{\varepsilon}$, where $\varepsilon_i$ is the square root of the sum of squared error (1), $k$ is a parameter that tunes the selectivity, $\bar{\varepsilon}$ is the median of $\varepsilon_i$ for a set of $N$ images, $\sigma$ is the average of the absolute differences between $\varepsilon_i$ and $\bar{\varepsilon}$.

### 3.3 SR reconstruction

In contrast to [1], we chose a direct method, which is divided into two main steps. A non-uniform interpolation of LR pixels on a HR grid is performed. Then, a Wiener Filter is applied to reduce noise and blur effects.

#### *3.3.1 Non-uniform interpolation*

We have extended the shift & add method of [15] to non-translational motions. This process is depicted in Fig. 3. For a given SR factor, a HR grid is built based on the reference ROI. All ROIs are registered w.r.t. this HR grid to form a non-uniform grid. Each point is projected to the nearest neighbour of the HR grid, which corresponds to rounding coordinates. When several points fall on the same node of the grid, average is computed. Conversely, some points of the grid receive no value. A four nearest neighbors multipass interpolation is done to make for this lack of information. Interpolation of a single pixel has negligible influence on the quality of the HR image, while a contiguous set of pixels without value can deteriorate SR reconstruction, particularly in textured areas. An interesting quality index is the filling rate of the HR grid at the end of non-uniform interpolation.

#### *3.3.2 Wiener filtering*

The restoration step is performed by a Wiener Filter, the Discrete Fourier Tranform (DFT) of the restored image has the following form:

$$X(f) = \frac{H^*(f)Y(f)}{|H(f)|^2 + \alpha|f|^\beta}$$

where $Y$ is the DFT of the noisy HR image obtained as a result of the interpolation step , $X$ is the DFT of the restored image, $H$ is DFT of the PSF of the optical sensor, $\alpha$ is the regularization parameter of Wiener filtering, and $\beta$ is the spectral exponent of the prior image Power Spectral Density.

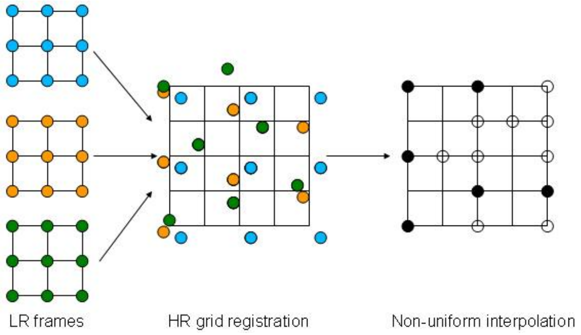

Figure 3: Non-uniform interpolation with nearest neighbor approximation. On the right grid nodes without circle have not received a value, ○ = 1 value and • = 2 values.

We use a box function PSF to model the sensor deformations. The Wiener Filter damps the amplitude of the frequencies with low signal-to-noise ratio, while reducing the blur effect due to the sensor. This method assumes uniform noise on the observed image. Then, after the reconstruction step, the HR image pixels are more or less averaged depending on the number of LR registered input pixels. The noise is not uniform in the image. Using a unique regularization parameter is therefore approximate.

## 4. EXPERIMENTAL RESULTS

### 4.1 Fast color super-resolution

To present our method, we have embedded the SR block of Fig. 1 into the tracking testbed of the VIVID project [9], which provides us with input tracks. In practice, the user delineates the selected object in the reference image using the VIVID GUI. In our evolution of VIVID, the user selects the SR parameters and clicks on the SR button. The object is then tracked backward and forward to obtain the corresponding ROIs which feed the SR block. A few tenth of second after, the SR still is displayed.

Fig. 4 illustrates the SR still obtained using one of the sequences of the VIVID database. On the left, a car is delineated in the scene. The SR reconstruction of this object is displayed on the right.

All the examples shown below come from an interlaced video aquired above a urban crossroad where many vehicules appear. We applied our process on vehicles of different sizes (50x50, 200x200) with a SR factor equal to 2, the computation times (CT) for the different examples are gathered in Table 1.

Fig. 5 shows zooms of results as well as the corresponding execution time, using 10 LR images. The various components of the vehicles are identified more easily, with a good restoration of details. The CT is growing with the size of ROIs. In addition, it should be noted that the proportion of time spent in registration and SR reconstruction is about 70% and 30% respectively. Currently, SR reconstruction reaches 12 to 20 fps (*frames-per-second*) depending on the size of objects on a 2 GHz dual core PC.

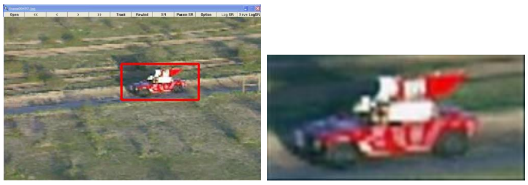

Figure 4: Selected object in the scene (left), and color SR reconstruction (right) with default settings proposed in section 4.3.3 and 20 images.

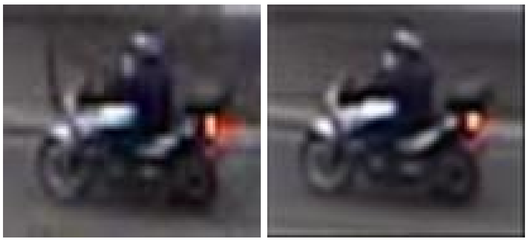

Figure 5: Original image of real size 51x46 pixels (left), after color SR processing of size 102x92 pixels (right) with default settings proposed in section 4.3.3.

### 4.2 Motion model and registration

We have applied the SR reconstruction on an object without registration, or with registration using a translation and an affinity. In Fig. 6 left, SR reconstruction without registration: results are useless. Warping of the front and the back of the van is not accurate enough with the translation (center), compared with the affinity (right). This result confirms that registration is mandatory and that the affine motion model is more appropriate.

### 4.3 Influence of parameters variation

The proposed fast method has several parameters: the images number, the registration parameters (*maxIter*, *RegThresh*) which influence the processing time, the interpolation parameter (the SR factor) which set the SR image size, and the restoration parameters ($\beta$, $\alpha$) which influence the quality of the final result.

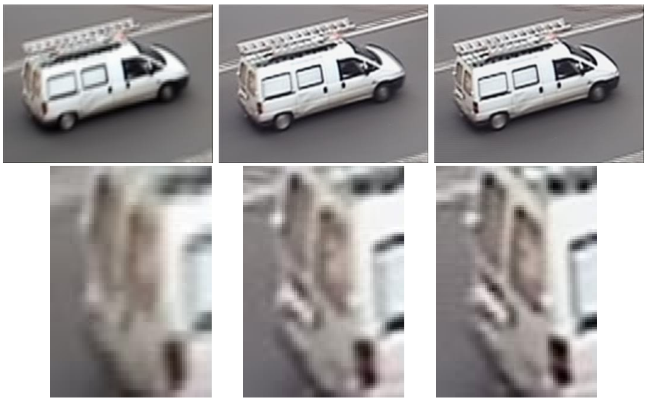

Figure 6: SR reconstructions with 10 images without registration (left), with translation (center) and affinity (right). Lower line: enlarged detail of each image

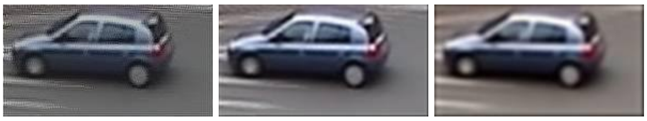

Figure 7: SR reconstruction under-regularized ($\alpha = 10$, left), over-regularized ($\alpha = 10^5$, right), and well regularized ($\alpha = 4000$, center), with default settings proposed in section 4.3.3.

#### 4.3.1 *Parameter $\alpha$ and $\beta$ of the restoration*

The regularization parameters $\alpha$ and $\beta$ of the Wiener Filter affects only the quality of the SR reconstruction, not the CT. Several values of $\beta$ were tested, and we could not notice much sensitivity. We have set $\beta = 4$, which amounts to a second-order Tikhonov regularization.

$\alpha$ is linked to the SR factor, and must be changed accordingly. If too low, the final image is grainy due to under-regularization. Conversely, if $\alpha$ is too high, the final image is blurred due to over-regularization. Fig. 7 presents three results illustrating the case with a SR factor of 2. On the left image, noise is very visible; the reconstruction is under-regularized with a too low value of $\alpha$, namely 10. The right image is blurred: it is over-regularized with too high value of $\alpha$ taken as $10^5$. A good result was found with $\alpha = 4000$, as on the central image. $\alpha$ should be increased with the SR factor.

#### 4.3.2 *Parameters maxIter and regThresh of the registration*

These two parameters affect both quality and the registration CT, and should thus be adjusted to provide a good trade-off, as the registration step represents 70% of the total SR CT. In addition, if this step fails globally on all images, final SR reconstruction is poor.

The processing time increases linearly w.r.t. number of iterations. Then, the parameter *maxIter* must be set to fix a limit of the registration step duration. However, if this parameter is too low, the results are not useless. Several values of *MaxIter* were tested, without significative variations for values greater than five. In this paper, the parameter *maxIter* is set to 5.

During the interpolation step, the motion of each LR pixel is rounded to the nearest HR pixel. Therefore it is sufficient to set the threshold *regThresh* near the HR pixel accuracy. Moreover, some registrations converge faster than others. We set *regThresh* the higher possible in order to limit the CT without degrading the result. It follows that for simple registrations the number of iterations is greatly reduced. For a SR factor equal to 2, The HR pixel size is equal to 0.5 LR pixel. We have scanned *regThresh* values between 0.1 and 0.9. We observed no sensitivity between 0.1 and 0.4, then the results get worse after 0.5. Thus *regThresh* is set to 0.4 LR pixel in present simulations. In case of higher SR factor, *regThresh* should be lower. This choice ensures a good registration accuracy, and reduces by about 50% the number of iterations.

#### 4.3.3 *Choice of images number*

Parameter tuning performed in the previous sections led to the following values: $MaxIter = 5$, $RegThresh = 0.4$, $\alpha = 4000$, and $\beta = 4$ for a SR factor equal to 2.

| Fig. | # pixel | CPU time(s) | Registration CPU time (%) |
|---|---|---|---|
| 4 | 3810 | 0.17 | 63 |
| 5 | 1050 | 0.27 | 94 |
| 6 | 7697 | 0.22-0.69-0.70 | 0-68-69 |
| 7 | 2036 | 0.31 | 85 |

Table 1: SR reconstruction CPU time for each example.

If the aliasing is strong on the LR images, it is potentially interesting to use more frames in order to enhance the SR reconstruction quality. But, in the case of small 3D objects, the registration gets increasingly difficult as the number of images grows. The number of used images should increase with the SR factor, so as not to degrade SR reconstruction by a low filling rate of the HR grid. In the case of SR factor of 2, the SR recontruction with ten frames provides a good trade-off between the result quality and the CT.

## 5. CONCLUSION

We propose a fast and robust color SR reconstruction method, adapted to the original context of small moving objects.

In the context of video surveillance we made several interesting observations: first, the hypothesis of translational motion is too crude and the affinity seems to be a good trade-off. Second, the registration is the most expensive step. This last remark is surprising since the CT alloted to registration is never considered in the litterature. In the short term this result encourages us to optimize further the registration and to consider more carefully the registration errors impact on the SR quality. Moreover, we will consider ground-truth experiments in order to asses the merits of affine versus translation models, and the influence of the different parameters. Other potential evolutions relate to the simultaneous processing of color channels and robustness to illumination variations.

The present SR functionality is involved in automatic annotation of video. In this context, tracks of several hundred images may contain multiple aspects of the same object. A perspective of this work is to provide automatic HR stills for each aspect.

## REFERENCES

[1] A. W. M. van Eekeren, K. Schutte, J. Dijk, D. de Lange, and L. van Vliet, “Super-resolution on moving objects and background,” in *Proceedings of the IEEE International Conference on Image Processing*, Atlanta, October 2006, pp. 2709–2712.

[2] S. C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: A technical overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, May 2003.

[3] N. Nguyen, P. Milanfar, and G. Golub, “A computationnaly efficient superresolution image reconstruction algorithm,” *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 573–583, April 2001.

[4] 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. 1327–1343, October 2004.

[5] G. Rochefort, F. Champagnat, G. Le Besnerais, and J.-F. Giovannelli, “An improved observation model for super-resolution under affine motion,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3325–3337, November 2006.

[6] R. C. Hardie, K. J. Barnard, J. G. Bognar, E. E. Armstrong, and E. A. Watson, “High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system,” *Optical Engineering*, vol. 37, no. 1, pp. 247–260, January 1998.

[7] S. Lertrattanapanich and N. K. Bose, “High resolution image formation from low resolution frames using delaunay triangulation,” *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1427–1441, 2002.

[8] R. Schultz and M. Alford, “Multiframe integration via the projective transformation with automated block matching feature point selection,” in *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, vol. 6, March 1999, pp. 3265 – 3268.

[9] R. Collins, X. Zhou, and S. K. Teh, “An open source tracking testbed and evaluation web site,” in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005)*, January 2005.

[10] M. Irani and P. Anandan, “All about direct methods,” in *In W. Triggs, A. Zisserman, and R. Szeliski, editors, Vision Algorithms: Theory and practice. Springer-Verlag*, 1999.

[11] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Preoceedings DARPA Image Understanding Workshop*, April 1981, pp. 121–130.

[12] P. H. S. Torr and A. Zisserman, “Feature based methods for structure and motion estimation,” in *Workshop on Vision Algorithms*, 1999, pp. 278–294.

[13] J. Barron, D. Fleet, and Beauchemin, “Performance of optical flow techniques,” in *International Journal of Computer Vision*, vol. 12, no. 1, 1994, pp. 43–77.

[14] D. Capel and A. Zissermann, “Computer vision applied to super resolution,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 75–86, May 2003.

[15] M. Elad and Y. Hel-Or, “A fast super-resolution reconstruction algorithm for pure translationnal motion and common space-invariant blur,” *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1187–1193, October 2001.

[16] M. S. Alam, J. G. Bognar, R. C. Hardie, and B. J. Yasuda, “Infrared image registration and high-resolution reconstruction using multiple translationally shifted aliased video frames,” *IEEE Transactions on Instrumentation and Measurement*, vol. 49, no. 5, October 2000.

[17] M.-C. Chiang and T. E. Boult, “Efficient super-resolution via image warping,” *Image and Vision Computing*, vol. 18, pp. 761–771, 2000.

[18] S. Baker and I. Matthews, “Lucas-Kanade 20 years on: A unifying framework,” *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221 – 255, March 2004.