MOTION COMPENSATED SUPER-RESOLUTION OF VIDEO BY LEVEL SETS EVOLUTION

Carlos Vázquez[†]*, Hussein Aly*[‡]*, Eric Dubois*[‡] *and Amar Mitiche*[†]

[†]INRS-EMT, Place Bonaventure, Suite 6900,
800 de la Gauchetière West, Montréal, Québec, H5A 1K6, Canada
e-mail: [vazquez]@inrs-emt.uquebec.ca
[‡]School of Information Technology and Engineering (SITE)
University of Ottawa,
161 Louis Pasteur, Ottawa, Ontario, K1N 6N5, Canada

ABSTRACT

In this contribution, we present an algorithm for motion compensated video super-resolution that combines the models we have developed for regularized image up-sampling and image reconstruction from irregularly-spaced samples. The information from several frames of a video sequence is used to obtain a high-resolution version of one of them. The problem is formulated in a variational framework as the minimization of an energy functional containing two kinds of terms. One relates to the error resulting from the approximation of the low-resolution images from the high-resolution one. A novel image formation model that accounts for the characteristics of the capturing and display devices is used for this term. The other term is a regularization functional measuring the correspondence of the HR image to a model related to our *a priori* knowledge of image characteristics. For this purpose we use a total variation (TV) prior. A continuous-space spline function models the high-resolution image, allowing to exactly solve a continuous-space defined problem in a discrete framework. The energy functional is minimized using the level-set representation. Preliminary results show the validity of the formulation and of the selected models.

1. INTRODUCTION

Super-resolution of video sequences is currently one of the most studied topics in video processing. The conversion from SD to HD video and scene interpretation from low-resolution video are examples of the vast number of applications a super-resolution algorithm can serve. The super-resolution problem consists in recovering a high-resolution (HR) image by combining multiple low-resolution (LR) ones. The common assumption is to consider some amount of motion in the sequence.

The main sequence of works on super-resolution build on the constraint that the LR images must be obtained from the HR one by appropriately applying the image formation model. Since this is an ill-posed problem, some kind of regularization is applied in order to find a unique solution. The main differences are found in the image formation model and the representation of the motion. In [1], a dense motion field is used in conjunction with a moving average filter to define the LR image formation model. A unique motion vector describes the motion of all the pixels in the HR image that contribute to a given pixel in the LR image, an important simplification that affects the quality of the results. In [2] a weighted average filter is combined with a parametric motion model which allows to better model the formation of the LR images. Global translational motion is also commonly considered since the solution is greatly simplified [3]. Recently a new kind of algorithm that relies on adding high-frequency features extracted from a training set of images has been proposed [4, 5].

In this contribution we follow the classical approach. Our image formation model, however, is somewhat different. We use an optimized blurring filter that takes into account the characteristics of the real camera that captures the LR images and the display that serves to visualize the HR image. This is combined with a continuous-space spline model of the HR image and a dense motion field in the HR grid with sub-pixel accuracy to model the formation of the LR image. The regularization prior is also different from previous works: we propose to use a total variation (TV) prior that preserves borders. We minimize the resulting energy functional with a level-set algorithm that ensures numerical stability.

The rest of the paper is organized as follows. In the next section we state the problem in a formal way. In section 3 we introduce the variational formulation and define the data fidelity and regularization functionals. Section 4 describes the level set minimization of the energy functional. Following, in section 5, we show experimental results and in section 6 we conclude.

2. PROBLEM STATEMENT

Let $\mathbf{S} = \{\mathbf{I}_t, t = 0, \ldots, T\}$ be an observed, sampled, LR video sequence composed of $T + 1$ images captured by a physical camera from a time-varying 3-D natural scene. The physical camera is modeled by a continuous-space linear shift-invariant filter with impulse response h_l followed by an ideal sampling on a regular sampling grid Γ. We will consider the temporal sampling rate fast enough (or the motion slow enough) to guarantee that essentially the same content is present in all images of the sequence. This is essential if we want to reconstruct a HR image from several LR frames of a video sequence and is a condition usually respected in natural video sequences. We will also consider in our presentation that there are no occlusions or newly exposed areas in the video

This work was supported by the Natural Sciences and Engineering Research Council of Canada under Strategic Grant OGP 0004234.

sequence. We are currently working on an improvement of our algorithm to handle these situations.

Let $\mathbf{I}_t = \{I[\mathbf{n}, t], \mathbf{n} = (n_1, n_2)^T \in \Gamma\}$ be the image at time $t \in [0, T]$, sampled on $\Gamma \subset \mathbb{R}^2$. Without loss of generality, we consider image $\mathbf{I}$ composed of a single component (luminance). Let $\mathbf{U}_t$ be the corresponding underlying continuous image at time t which, after filtering and sampling, produces the LR image $\mathbf{I}_t$.

We are interested in the recovery of a HR image $\mathbf{J}_0$, at time instant $t = 0$ and sampled on a regular grid $\Lambda \subset \mathbb{R}^2$ much denser than Γ, from the LR video sequence $\mathbf{S}$. The image $\mathbf{J}_0$ would be obtained by a virtual camera modeled by a continuous-space linear shift-invariant filter with impulse response h_h followed by an ideal sampling on Λ. The estimation of the optimal filter with impulse response h that serves to obtain $\mathbf{I}_0$ from $\mathbf{J}_0$ is the subject of our work [6]. We will use this approach to model the process of obtaining the LR images $\mathbf{I}_t$ from their HR versions $\mathbf{J}_t$. Fig. 1 illustrates the proposed image formation model.

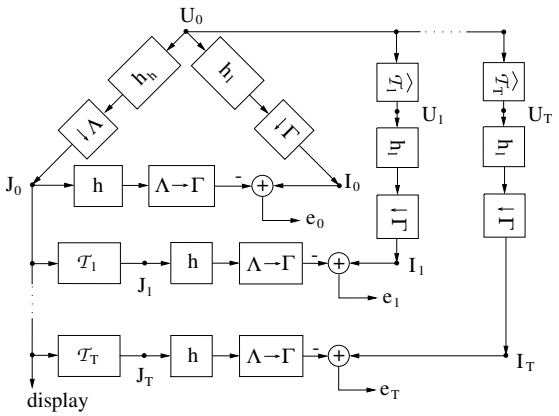


Fig. 1. Image formation model

Given that essentially the same content is present throughout the sequence, and assuming the constant intensity constraint along motion trajectories, the images in the sequence are related by geometrical transformations (motion). Let $\widehat{\mathcal{T}}_t$ be the real continuous-space transformation relating the reference continuous image $\mathbf{U}_0$ to the continuous image $\mathbf{U}_t$ and let $\mathcal{T}_t : \Lambda \rightarrow \mathbb{R}^2$ be its sampled version on Λ (motion vector field) that relates $\mathbf{J}_t$, the HR image at time t to our reference image $\mathbf{J}_0$ at time $t = 0$ through:

$$\mathbf{J}_t[\mathbf{m}] = \mathbf{J}_0(\mathcal{T}_t[\mathbf{m}]), \;\; \mathbf{m} \in \Lambda \tag{1}$$

In this equation we have somehow abused the notation by evaluating $\mathbf{J}_0$ at points $\mathbf{x} \in \mathbb{R}^2$ while $\mathbf{J}_0$ is an image defined on the discrete sampling grid Λ. Image values at points $\mathbf{x} \in \mathbb{R}^2$ must be found by spatial interpolation as is common in motion compensation algorithms. Moreover, to model the HR image $\mathbf{J}_0$, we use a continuous-space function which serves the interpolation process. In the following, we assume knowledge of the geometric transformations $\mathcal{T} = \{\mathcal{T}_t, t = 1, \ldots, T\}$, each described by a dense motion vector field ($\mathcal{T}_t = \{(u_t[\mathbf{m}], v_t[\mathbf{m}]), \mathbf{m} \in \Lambda\}$). The estimation of the transformations from the LR sequence is a research problem by itself in which we are currently working.

The problem can then be stated as follow: *Given the LR video sequence* $\mathbf{S}$ *and the set of transformations* $\mathcal{T}_t, t = 1, \ldots, T$*, recover the best estimate of the HR image* $\mathbf{J}_0$.

3. FORMULATION IN A VARIATIONAL FRAMEWORK

We formulate the problem in a variational framework to seek a function that minimizes an energy functional containing two terms: one related to the approximation error with respect to data (LR images) and a term related to the *a priori* knowledge of the characteristics of the HR image (regularization).

In order to formulate the problem in continuous space, we need a continuous model for the images. For this purpose we use, as in our previous work [7], a bi-cubic spline function:

$$f(\mathbf{x}) = \sum_{\mathbf{m} \in \Lambda} c[\mathbf{m}] \phi(\mathbf{x} - \mathbf{m}), \;\; \mathbf{x} \in \mathbb{R}^2, \;\; \mathbf{m} \in \Lambda, \tag{2}$$

where $\phi(\mathbf{x})$ is the 2-D bi-cubic B-Spline function and $\{c[\mathbf{m}] =< f, \mathring{\phi}_{\mathbf{m}} >, \mathbf{m} \in \Lambda\}$ are the spline coefficients. The function $\mathring{\phi}\,(\mathbf{x})$ is the dual function of ϕ and the subscript $\mathbf{m}$ denotes a spatial shift: $\mathring{\phi}_{\mathbf{m}}\,(\mathbf{x}) = \mathring{\phi}\,(\mathbf{x} - \mathbf{m})$.

The HR image $\mathbf{J}_0$ is, then, modeled as a sampled version of function f:

$$\mathbf{J}_0[\mathbf{m}] = f(\mathbf{m}), \;\; \mathbf{m} \in \Lambda. \tag{3}$$

Function f is fully described by the image of spline coefficients $\mathbf{C} = \{c[\mathbf{m}], \; \mathbf{m} \in \Lambda\}$ which, in turn, can be obtained from function samples (sampled image) $\{f(\mathbf{n}), \; \mathbf{n} \in \Lambda\}$ by fast digital filtering techniques as proposed in [8]. The recovery of function values from spline coefficients is achieved by using the same kind of algorithm. This model allow us to exactly solve a continuous problem in a discrete framework.

The problem we are faced with is then formulated as the solution to the minimization of an energy functional $\mathcal{E}$:

$$f(\mathbf{x}) = \arg\min_{\tilde{f}} \left(\mathcal{E}(\tilde{f}; \mathbf{S}, \mathcal{T})\right) \tag{4}$$

where $\mathbf{S}$ is the LR sequence and $\mathcal{T}$ is the sequence of HR motion vector fields. The energy functional $\mathcal{E}$ must contain a data fidelity term $\mathcal{E}_d$ that measures the correspondence of the solution to data and, since this term is generally not sufficient to define a unique solution, a regularization term measuring the correspondence of the solution to an image model that includes our *a priori* knowledge of the images:

$$\mathcal{E}(f; \mathbf{S}, \mathcal{T}) = \mathcal{E}_d(f; \mathbf{S}, \mathcal{T}) + \lambda \mathcal{E}_r(f). \tag{5}$$

The positive constant λ controls the tradeoff between approximation error and regularity of the solution.

3.1. Data fidelity functional

The data fidelity term is related to the approximation error:

$$\mathcal{E}_d(f; \mathbf{S}, \mathcal{T}) = \frac{1}{2} \sum_{t=1}^{T} \|e_t\|^2 \tag{6}$$

where e_t is the error function between the estimated LR image $\tilde{\mathbf{I}}_t$ obtained from $\mathbf{J}_0$ and the observed LR image $\mathbf{I}_t$. The error function e_t can be computed as:

$$e_t = \mathbf{H} \mathbf{\Phi}_t \mathring{\mathbf{\Phi}} \mathbf{f} - \mathbf{I}_t \tag{7}$$

where $\mathbf{H}$ is the $N \times M$ matrix (M, N being the number of samples in the HR and LR images respectively) which describes the

two steps of filtering with h and down-sampling from Λ to Γ. The $M \times M$ matrix $\mathbf{\Phi}_t$ has elements of the form $\phi_{ij}^{(t)} = \phi(\mathbf{x}_{\mathbf{m}_i,t} - \mathbf{m}_j)$ with $\mathbf{x}_{\mathbf{m}_i,t} = \mathcal{T}_t[\mathbf{m}_i]$ for $t = 0, \ldots, T$ (the transformation function for $t = 0$ is the identity: $\mathcal{T}_0[\mathbf{m}] = \mathbf{m}$). Matrix $\mathbf{\Phi}_t$ performs an interpolation step to compute the values of the function f at the (irregularly-spaced) motion-compensated locations from spline coefficients. The $M \times M$ matrix $\mathring{\mathring{\mathbf{\Phi}}}$ has elements of the form $\mathring{\phi}_{ij} = \mathring{\phi}(\mathbf{m}_i - \mathbf{m}_j)$ and is responsible for converting function samples to spline coefficients. The vector $\mathbf{f}$ is the lexicographic reading of the samples of function f modeling the HR image $\mathbf{J}_0$. The data fidelity term is defined in matrix form as:

$$\mathcal{E}_d(f; \mathbf{S}, \mathcal{T}) = \frac{1}{2} \sum_{t=1}^{T} \|\mathbf{H}\mathbf{\Phi}_t \mathring{\mathring{\mathbf{\Phi}}}\mathbf{f} - \mathbf{I}_t\|^2 \tag{8}$$

Minimization of $\mathcal{E}_d$ involves finding the function f that makes zero the functional derivative of the energy $\mathcal{E}_d$ with respect to f:

$$\frac{\partial \mathcal{E}_d}{\partial f} = \sum_{t=0}^{T} \mathring{\mathring{\mathbf{\Phi}}}^T \mathbf{\Phi}_t^T \mathbf{H}^T (\mathbf{H}\mathbf{\Phi}_t \mathring{\mathring{\mathbf{\Phi}}}\mathbf{f} - \mathbf{I}_t) \tag{9}$$

The transposed matrix $\mathbf{H}^T$ is implemented in a two step procedure: up-sampling from Γ to Λ and then filtering with $\bar{h} = h[-\mathbf{x}]$. Matrix $\mathbf{\Phi}_t^T$ can be saw, at least conceptually, as a three step operator: up-sampling to a 'continuous space', filtering with a continuous-space filter with impulse response $\bar{\phi}$, and down-sampling to Λ. In practice, since ϕ has a closed-form expression, the multiplication process $\mathbf{\Phi}_t^T \mathbf{b}$ is easily carried-out by evaluating the function $\varphi(\mathbf{x}) = \sum_{\mathbf{m} \in \Lambda} b[\mathbf{m}] \phi(\mathbf{x} - \mathcal{T}_t[\mathbf{m}])$ at points $\mathbf{x} \in \Lambda$. From the symmetry of $\mathring{\phi}$, we have $\mathring{\mathring{\mathbf{\Phi}}}^T = \mathring{\mathring{\mathbf{\Phi}}}$.

3.2. Regularization functional

Minimization of the data fidelity function $\mathcal{E}_d$ defined by (8) is an ill-posed problem, since matrix $\mathbf{H}$ is rank-deficient, resulting in an under-determined system of equations. This problem is usually solved through regularization in image processing applications.

We have chosen a Total Variation regularization for two reasons: First, this regularizer defines a class of piecewise regular functions, which is exactly what we are looking for. Second, it is a continuous-domain regularizer which fits our continuous model. The total variation regularizer functional is defined as:

$$\mathcal{E}_r(f) = \int_{\mathcal{W}} \|\nabla_{\mathbf{x}} f(\mathbf{x})\| \, d\mathcal{W} \tag{10}$$

on the continuous space domain $\mathcal{W}$ of support of f. In this equation, $\nabla_{\mathbf{x}}$ denotes spatial gradient.

Minimization of this functional involves the derivation with respect to the modeling function f:

$$\frac{\partial \mathcal{E}_r}{\partial f} = \operatorname{div}\left(\frac{\nabla_{\mathbf{x}} f}{\|\nabla_{\mathbf{x}} f\|}\right) \tag{11}$$

where div denotes divergence.

4. LEVEL-SET MINIMIZATION

Level set methods are concerned with the evolution of fronts along the normal direction to the front. In the case of images, these fronts are the iso-intensity contours of the images. The general level set evolution equation can be written as:

$$\frac{\partial f(\mathbf{x}, \tau)}{\partial \tau} + \nu(\mathbf{x}, \tau) \|\nabla_{\mathbf{x}} f(\mathbf{x}, \tau)\| = 0 \tag{12}$$

where τ is algorithmic time of evolution, and $\nu(\mathbf{x}, \tau)$ is the speed of the front in the outward normal direction. Many evolution speed functions ν have been proposed for several applications of the level set method. In our algorithm we use a speed function that tends to minimize $\mathcal{E}(f; \mathbf{S}, \mathcal{T})$ composed of two terms corresponding to the two terms in the definition of the energy functional.

4.1. Data driven evolution speed component

The speed function component corresponding to the data fidelity term is defined as:

$$\nu_d(\tau) = \frac{\sum_{t=0}^{T} \left(\mathring{\mathring{\mathbf{\Phi}}} \mathbf{\Phi}_t^T \mathbf{H}^T (\mathbf{H}\mathbf{\Phi}_t \mathring{\mathring{\mathbf{\Phi}}}\mathbf{f} - \mathbf{I}_t) \right)}{\|\nabla_x f(\tau)\|} \tag{13}$$

In this equation, we have defined the speeds at positions on the regular grid Λ. Since f is a spline function, it is completely described by its samples on Λ.

4.2. Regularization driven evolution speed component

The second speed component, corresponding to the regularization term, is:

$$\nu_r(\tau) = -\lambda \operatorname{div}\left(\frac{\nabla_{\mathbf{x}} f}{\|\nabla_{\mathbf{x}} f\|}\right) = -\lambda \kappa(\tau) \tag{14}$$

where κ is the mean curvature function of the front (iso-intensity curve of f) at algorithmic time τ. Since f is modeled by a bicubic spline function described by its spline coefficients $c[\mathbf{m}]$, the mean curvature of the function is easily computed from the spline coefficients.

4.3. Final evolution equation

From (13) and (14), the final evolution equation for the solution of equation (4) is given by:

$$f^{(n+1)} = f^{(n)} + \Delta T \left(\lambda \kappa \|\nabla_{\mathbf{x}} f^{(n)}\| - \right. \tag{15}$$
$$\left. - \sum_{t=0}^{T} \left(\mathring{\mathring{\mathbf{\Phi}}} \mathbf{\Phi}_t^T \mathbf{H}^T (\mathbf{H}\mathbf{\Phi}_t \mathring{\mathring{\mathbf{\Phi}}}\mathbf{f}^{(n)} - \mathbf{I}_t) \right) \right)$$

where ΔT is the algorithmic time step.

5. EXPERIMENTAL RESULTS

We show the results of applying our algorithm for the super-resolution of an image from 8 LR images of a sequence with rotational motion. The up-sampling factor is 5 in each direction ($\times 25$ up-sampling rate) and the rotation angle between images is $2\pi/15$.

Fig. 2 shows the seven LR images used. The LR images have been obtained following the observation model, ie. by applying

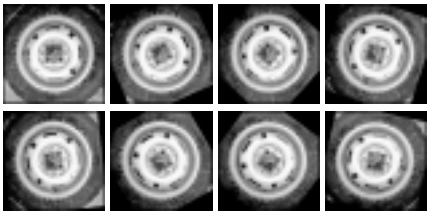
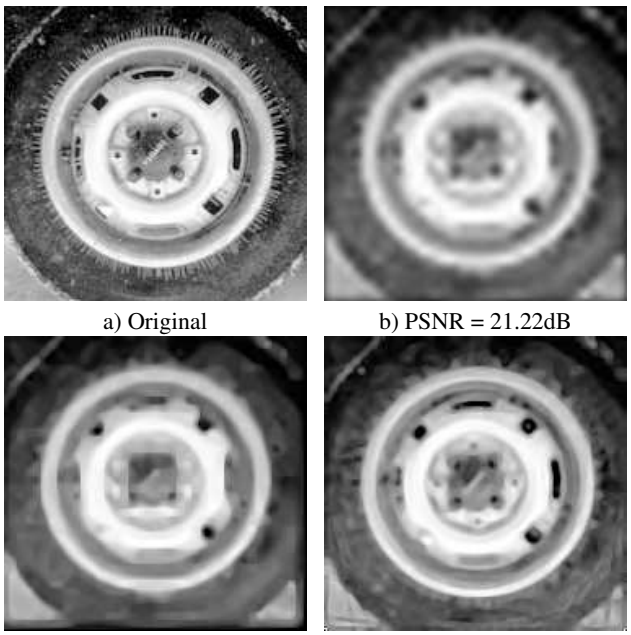


Fig. 2. Low-resolution images

the corresponding transformation to the original image "Tire" followed by filtering with h_l and down-sampling on Γ.

Fig. 3 shows the up-sampled versions of the first image in the sequence. In order to offer a point of comparison we show in Fig. 3a the original HR image, in Fig. 3b the result of up-sampling the same image with bi-cubic splines, a well known interpolation algorithm. Fig. 3c shows the result of up-sampling the first image with the total variation algorithm proposed in [6] and Fig. 3d shows the result of applying the proposed method.



a) Original b) PSNR = 21.22dB

c) PSNR = 21.25dB d) PSNR = 23.54dB

Fig. 3. HR images: a) Original, b) Cubic splines interpolation of first image, c) Total variation up-sampling from first image, d) Proposed method with 8 LR images

The image obtained with our algorithm is much better defined. The blur has been greatly reduced and the edges are sharp. There is no ringing in regular regions and the whole image is more pleasant to observe. We can also observe the introduction of new features which are not visible in the spatial interpolation of the first image alone. This is a direct result of the use of several LR images. The PSNR measure with respect to the original image is also better. The advantage of using the TV prior is appreciated from the images in Fig. 3c and d, even if the PSNR for the TV up-sampling and the spline interpolation are almost equals, the TV up-sampled image is much more pleasant to observe.

6. CONCLUSIONS

We have presented an algorithm to construct a motion compensated HR image from a LR video sequence and the corresponding HR motion vector fields. The problem of estimating the HR motion vectors from the LR video sequence is under study and will be addressed in future contributions. We have formulated the problem as the minimization of an energy functional containing two kinds of terms. One relates to the error resulting from the approximation of the LR images from the HR one. The other is a regularization term measuring the correspondence of the HR image to a model related to our *a priori* knowledge of image characteristics. A regularization parameter is used in order to control the tradeoff between approximation and regularization. A novel image formation model that accounts for the characteristics of the capturing and display devices is introduced. A Total Variation prior is used as regularizer in order to preserve borders and eliminate the staircase effect introduced by most interpolation algorithms. The resulting objective function is minimized by means of the level-set formalism to provide a stable solution. The HR image is modeled by a spline function which allows to solve a continuous-space problem in a discrete framework. The preliminary experimental results are encouraging and show the validity of the proposed model and the applicability of the algorithm. We are currently working on an improved algorithm that will take into account occluded and motion exposed areas.

7. REFERENCES

[1] R. Schultz and R. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Processing*, vol. 5, no. 6, pp. 996–1011, June 1996.

[2] R. Hardie, K. Barnard, and E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Processing*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997.

[3] N. Nguyen, P. Milinfar, and G. Golub, "A computationally efficient superresolution image reconstruction algorithm," *IEEE Trans. Image Processing*, vol. 10, no. 4, pp. 573–583, Apr. 2001.

[4] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 9, pp. 1167–1183, Sept. 2002.

[5] C. M. Bishop, A. Blake, and B. Marthi, "Super-resolution enhancement of video," in *Proc. Artificial Intelligence and Statistics*, C. M. Bishop and B. Frey, Eds., Key West, FL, USA, Jan. 2003.

[6] H. Aly and E. Dubois, "Regularized image up-sampling using a new observation model and the level set method," in *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Spain, Sept. 2003, vol. 3, pp. 665–668.

[7] C. Vázquez, E. Dubois, and J. Konrad, "Reconstruction of irregularly-sampled images in spline spaces," *IEEE Trans. Image Processing*, Jan. 2004, accepted.

[8] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Parts–I and II," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 821–848, Feb. 1993.