Super Resolution for Automated Target Recognition

Maksim Levental

***Abstract*—Super resolution is the process of producing high-resolution images from low-resolution images while preserving ground truth about the subject matter of the images and potentially inferring more such truth. Algorithms that successfully carry out such a process are broadly useful in all circumstances where high-resolution imagery is either difficult or impossible to obtain. In particular we look towards super resolving images collected using longwave infrared cameras since high resolution sensors for such cameras do not currently exist. We present an exposition of motivations and concepts of super resolution in general and current techniques, with a qualitative comparison of such techniques. Finally we suggest directions for future research.**

1 Introduction

Super-resolution (SR) is a collection of methods[1] that augment the resolving power of an imaging system. Here, and in the forthcoming, by resolving power we mean the ability of an imaging device to distinguish distinct but proximal objects in a scene. If such objects are modeled as point sources of light then the resolving power of the imaging system is defined by Rayleigh's criterion: two point sources are considered *resolved* when the first diffraction maximum[2] of one point source (at most) coincides with the first minimum of the other (see figure 1).

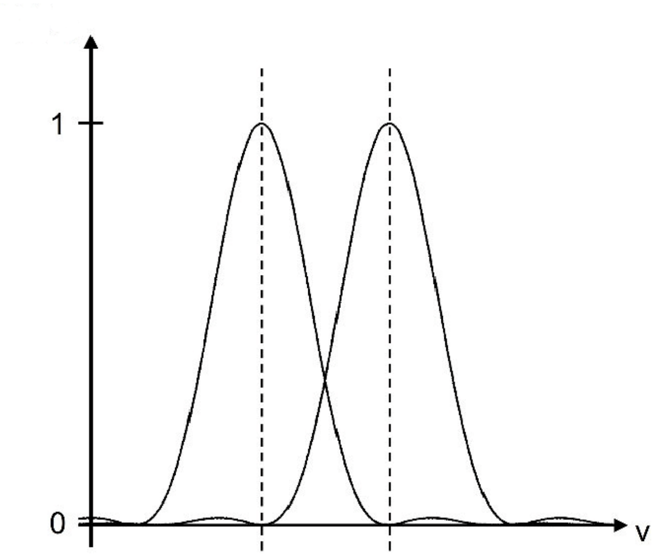

Fig. 1: Rayleigh's criterion[2]

Fig. 2: Multiple image super resolution[3]

SR techniques yield high-resolution (HR) images from one or more observed low-resolution (LR) images by restoring lost fine details and reversing degradations produced by imperfect imaging systems. In the case where a single LR source image is used to construct the HR correspondent, the techniques are referred to as single-image-super-resolution (SISR) techniques. These techniques typically operate by either learning some mapping from low resolution chips (uniform partitions of the image, e.g. 3×3 pixels) to higher resolution chips that are highly similar (according to some metrics) and obey regularity constraints (e.g. agreement at edges). In the case when multiple LR source images are used to construct the single HR correspondent, the techniques are referred to as multiple-image-super-resolution (MISR) techniques. MISR techniques rely on non-redundant and yet pertinent information in multiple images of the same scene (see figure 2). Note that for such information to exist there should be sub-pixel[3] shifts in either the imaging system or the scene between consecutive images.

For typical imaging use-cases, high resolution images are preferable to low resolution images; higher resolutions are desirable in and of themselves and as inputs to later image processing transformations that can degrade image quality (e.g. by virtue of quantization or compression). In theory the resolving power of an imaging system is primarily determined by the number of independent sensor elements that comprise that imaging system (each of which collects a component of the ultimate image). Naturally then, a way to increase the resolution of such a system is to increase the density of such sensor elements per unit area. Unfortunately, and counter-intuitively, since the number of photons incident on each

[1] We will often use the verb form "to super resolve" in order to denote the use of one or more such methods.

[2] The amplitude of the diffraction pattern (known as the Airy pattern) of a monochromatic point source through a circular aperture is given by

$$I(\theta) = I_0 \left[\frac{2J_1(ka\sin\theta)}{ka\sin\theta} \right]^2$$

where I_0 is peak intensity (at the center), $k = \frac{2\pi}{\lambda}$ is the wave number of the light, θ is the angle of observation, and J_1 is the Bessel function of the first kind of order one[1]. It is maxima/minima of this function that Rayleigh's criterion concerns.

[3] For example when a point source wholly captured by one sensor element shifts to distributing energy equally amongst the same element and a direct adjacent.

sensor decreases as the sensor shrinks, shot noise[4] thwarts that idea. Furthermore, while sensor density is primary, secondary effects due to optics limit resolution as well; the point spread of a lens (distortion of a point source due to diffraction), chromatic aberrations (distortion due to differing indices of refraction for differing wavelengths of light), and motion blur all function to obscure or erase details from the image.

In domains such as satellite/aerial photography, medical imaging, and facial recognition, high-resolution reconstruction of low-resolution samples is eminently useful since ab-initio acquisition of high-resolution images is either logistically difficult or impossible due to aforementioned imaging apparatus limitations. For example in the instance of satellite imagery, acquisition of high-resolution imagery is primarily hampered by optics and physics[5]. In contrast, in the cases of medical imaging (where procedures are invasive and patient exposure time needs to be minimized[6]) and facial recognition (e.g. for purposes of surveillance) the primary challenge is logistics and access to repeat collection opportunities.

The benefits of enhancing images using SR techniques include not only more pleasing or more readily interpretable images for human consumption but higher quality inputs for automated learning systems as well. In particular object detection systems trained on super-resolved images outperform those trained on the low resolution originals[7]. Indeed this is our ultimate goal — not super-resolution per se but super-resolution in the service of improved object detection performance for longwave-infrared (LWIR) imagery. Note that while practically speaking, there exist hardware solutions for increasing the resolution of an imaging system, we discount the value of such propositions. We instead take low resolution images as given and seek techniques that allow for ex post facto reconstruction or inference of precise details. This necessarily constrains techniques under consideration to be algorithmic in nature and software in practice.

The rest of this survey is outlined as follows: Section 2 introduces imaging systems, notation, and the model of imaging that will be the mathematical framework for the proceeding sections, Section 3 surveys classical techniques (those that do not employ neural networks), Section 4 surveys neural-network techniques with heavy emphasis on deep learning (i.e. deep networks), Section 5 discusses the scope and goals of the author's research program, and Section 6 summarizes.

2 Background

2.1 Imaging systems

We begin with a practical discussion of imaging systems. An imaging sensor is a device that converts an optical image into a digital signal. Charge-coupled devices (CCD) and complementary metal-oxide-semiconductor (CMOS) devices are the most common imaging sensors; CCDs have better performance while CMOS devices are newer and less expensive. A third type that's of particular interest to us is the microbolometer, which is used as a sensor in thermal cameras.

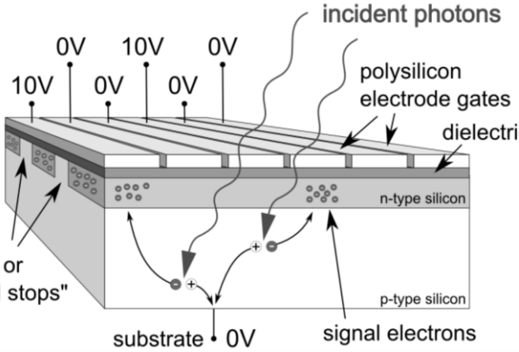

Fig. 3: CCD buried channel MOS capacitor[8]

CCDs consist of densely packed two-dimensional arrays of buried channel[6] MOS capacitors (see figure 3) with an individual MOS capacitor being the fundamental photon detecting element. Individual MOS capacitors are biased by a gate voltage such that a potential well is produced in the n-type silicon (referred to as the n-channel). This potential well acts as a storage system for charge induced by the inner photoelectric effect[7]. When photons are incident on a MOS capacitor some of the photons are absorbed, some are scattered, and some are transmitted. Those photons that are transmitted interact with electrons in the valence band of the silicon exciting them into the conduction band, and thereby create electron-hole pairs that either diffuse or recombine. For high-quality silicon, the lifetime of such a pair is several milliseconds (before recombination)[10]. The electrons of the electron-hole pairs that do not recombine diffuse into the potential well, while the holes migrate to the grounded substrate (i.e. out of the sensor). Electrons created in this way are called *photoelectrons*.

CCD arrays consist of two sub-arrays: an image section and a readout section (see figure 4). The image section is arranged with every third stripe of electrode tied electrically to form three sets of equipotentials. In figure 4 these equipotentials are labeled $\Phi 1, \Phi 2, \Phi 3$, and taken together constitute a vertical register (VR). They function to move the collected photoelectrons down one electrode line at a time, using charge coupling, while the channel stops function to prevent diffusion of charge across channels. The VR mechanism that shifts collected charge operates as follows:

1) Suppose initially there's a collection of photoelectrons on each channel at $\Phi 1$ and only $\Phi 1$. Note this means $\Phi 2, \Phi 3$ are at 0v (again just as in figure 3).

[4] TODO

[5] Rayleigh's criterion implies that the angular resolution R of a telescope with optical diameter $D = 2.4\text{m}$ observing visible light ($\sim$500nm) is approximately[4]

$$R \approx 1.220 \frac{\lambda}{D} = 1.220 \frac{500\text{nm}}{2.4\text{m}} \approx 0.06\text{arcsec}$$

From an altitude of 250 km this corresponds to a ground sample distance of 6cm. This loss of resolving power is further exacerbated by refraction through turbulent atmosphere[5].

[6] Buried channel as distinct from surface channel. In surface channel MOS capacitors signal charge is stored at the Si-SiO$_2$ interface, which can lead to charge trapping during the charge transfer process[9].

[7] The photoelectric effect is the emission of electrons when light hits a material. The inner photoelectric effect is that phenomenon in the bulk matter of semiconductors.

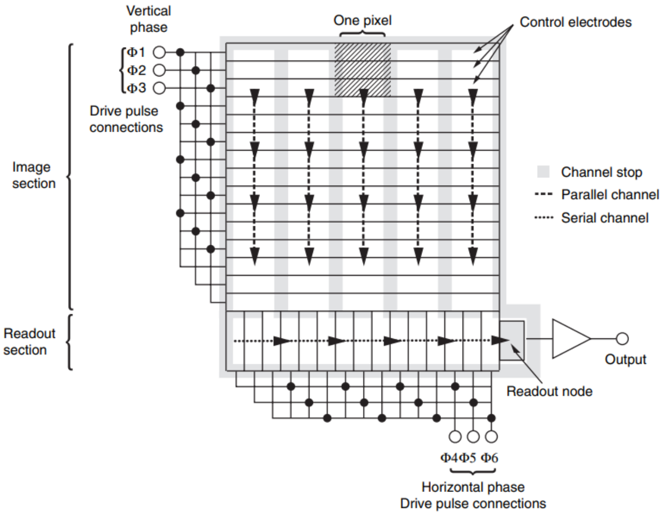

Fig. 4: CCD array[11]

2) $\Phi2$ is positively biased to 10V. This diffuses the collection of charge under both $\Phi1$ and $\Phi2$.
3) $\Phi1$ is set to 0v. This concentrates the collection of charge under $\Phi2$.
4) The same is repeated with $\Phi2, \Phi3$ and $\Phi3, \Phi1$.
5) The entire process repeats thereby shifting the charge three lines (or one pixel row) at a time.

At the bottom of the image section $\Phi3$ transfers all signal charges to the horizontal register (HR) which functions much like the VR except faster: the HR must transfer every line of pixels independently of all other lines to the read-out node. An obvious challenge faced by this system is how to prevent errant charge from accumulating out of sync with the shift process i.e. how to prevent new photoelectrons from being produced at intermediate lines while far lines are being shifted. The solution is to have interstitial dedicated shift channels in between columns of sensors, with the shift channels being masked off from exposure to light. This type of reading is called *interline transfer* because the accumulated charge is first moved one line over, into the shift channels. Naturally interline transfer shrinks photosensitive area by half and despite possible solutions (e.g micro-lenses being used to focus most of the light onto the unmasked sensors) this is one of the drawbacks of CCDs that CMOS imaging systems do not share.

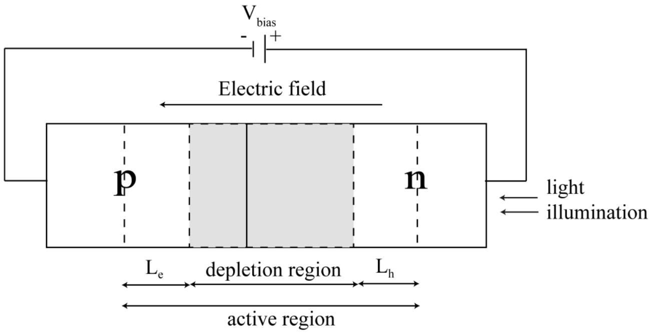

Fig. 5: Photodiode schematic. L_e, L_h are electron, hole diffusion lengths respectively[12]

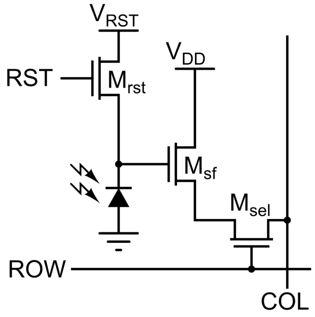

Fig. 6: Three transistor "pixel". M_{rst} is the reset transistor (enabling the photodiode to dump charge), M_{sf} buffers the charge on the photodiode (so that it can be read without loss), and M_{sel} enables a whole row of pixels to be read simultaneously (since all pixels in a physical row are tied to the same row line).

CMOS sensors consist of arrays of photodiodes (see figure 5). A photodiode is a p-n junction[8] operated in reverse bias mode[9]. When a photon of sufficient energy is absorbed by the diode, it creates an electron-hole pair. If the creation event happens within the active region then the hole moves out through the p-type material and the electron moves out through the n-type material. This establishes a *photocurrent* that can be read by a reading circuit (see figure 6). CMOS sensor arrays do not shift the charge from row to row like CCD arrays. In a CMOS sensor array, each pixel contains a transistor M_{sel} controlled by the voltage applied across a row (see figure 7). In order to read one row of pixels, a row line is raised high to turn on (close) all the M_{sel} transistors in the row. This brings the signals from all the pixels in that row to the shifter register below by way of the column lines. The shift register then outputs the values of the pixels. The high number of transistors needed per pixel in CMOS arrays has only recenty been manageable for semiconductor foundries. This, along with such artifacts as the "rolling shutter" effect produced by rowline reading, are some of the drawbacks of CMOS arrays relative to CCD arrays. Despite this CMOS arrays have become the most common imaging system in consumer goods such as cell phones and digital cameras due to their relatively simple mechanics.

Both CCD arrays and CMOS arrays only capture visible light. A microbolometer, on the other hand, measures the power in the infrared by exposing a thermistor[10] to the incident light. Since a thermistor's resistance changes as a function of its temperature, a key issue in the design of a microbolometer is the thermal isolation of the thermistor. With the maturation of micro-machining techniques (such as for MEMS[11]

[8] The interface between a p-type semiconductor (excess holes, i.e. positive charge carriers) and an n-type (excess electrons, i.e. negative charge carriers) semiconductor.

[9] With the p-type material at a lower voltage than the n-type. This causes both the holes and the electrons to flow away from the junction creating a depletion zone.

[10] An element with an electrical resistance that's a function of its temperature.

[11] Micro-electro-mechanical systems.

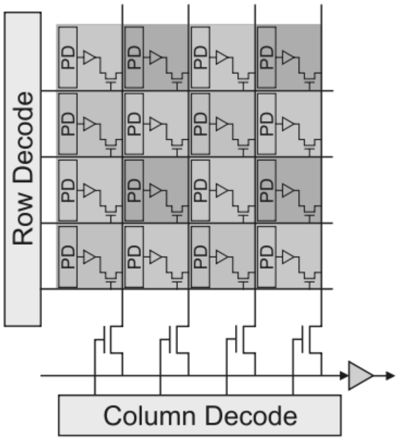

Fig. 7: CMOS array

devices) over the last some years, two level microbolometers consisting of a thermo-sensitive component suspended above (and insulated from) silicon have been built (see figure 8). These pixel packages are evacuated and therefore have good conduction, convection, radiation heat transfer properties. The actual thermo-sensitive component consists of a thermistor, an absorber (which aids in transfer of heat to the thermistor), and a reflector that creates a Fabry–Pérot optical cavity[12] (typically $\sim\lambda/4$[14]) that traps the infrared light. Typical materials for the thermistor are vanadium oxide and amorphous silicon owing to their high temperature coefficients of resistance[14], which in effect transform small changes in temperature into large changes in resistance. Measurements of the thermistor are performed by a read-out integrated circuit adjacent to the bridge in the silicon substrate. All told microbolometers are designed much differently from either CCD or CMOS arrays. It is as a result of this fact that high-resolution infrared cameras are not available.

Across all of these imaging systems there are ample avenues for the introductions of the kinds of errors that degrade image quality and across all of these imaging systems there are structures that impose limitations on resolving power. With that in mind we now proceed to formalizing the problem of super-resolution.

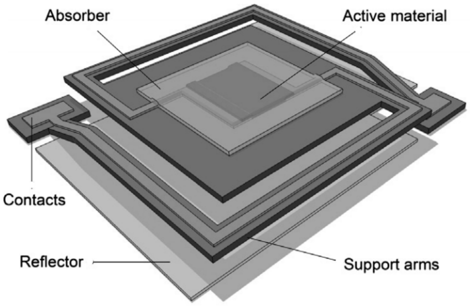

Fig. 8: Bridge structure of Honeywell microbolometer[13]

2.2 Mathematical notation

Upper case plain latin X, Y each denote channel × row × column *tensors*[13] representing LR and HR images respectively, with $(0, 0, 0)$ corresponding to the top left corner of the first channel of image. Often for the sake of simplicity we consider greyscale images, in which case we omit the channel dimension. Lower case plain latin x, y denote LR and HR *patches*[14] respectively. D, H, F, G variously refer to functions that operate on images. Bolded uppercase latin $\boldsymbol{X}, \boldsymbol{Y}$ are reserved for batches of images, i.e. batch size × channel × row × column tensors with $(0, 0, 0, 0)$ corresponding to the top left corner of the first channel of first image in the batch. Bolded lower case denote $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ etc. denote conventional column or row vectors. Unless otherwise specified $\|\cdot\|$ is the L_2 norm.

2.3 Imaging model

Figure 9 shows a conceptual model of the imaging process as carried out by an imaging system. The input to the system is a natural scene that is in effect sampled by the imaging system. In the idealized case the sampling is done at (or above) the Nyquist rate and no aliasing occurs. In practice there is noise and loss introduced at every step of the process: atmospheric turbulence plays a role at large distances, motion produces multiple views of the same scene but also induces blur, imperfections of the lenses further blur the image, and finally down-sampling by the sensor elements into pixels produces aliasing artifacts[15]. The noisy, blurry, down-sampled images are then further degraded by sensor noise. Each such image we call an LR sample. Let Y denote an idealized HR image of the scene from some fixed vantage point and assume the imaging system collects K LR samples X_k of Y. Formally the X_k are related to Y by

$$X_k = (D \circ H_k \circ A_k)(Y) + \varepsilon_k \tag{1}$$

where for the kth sample A_k (of K) is the operator representing motion (affine and perspective shift) induced by optical flow[16], H_k represents the blur operator, D represents the down-sampling operator (constant in time since it's typically a digital component of the imaging system), and ε_k represents the composite noise (environment and sensor noise).

We now consider the challenges and nuances of estimating motion and blur.

For a static 3-D scene and an imaging system with 6 degrees of freedom, the optical flow caused by motion is dependent on the geometry of the scene and potentially nonlinear (due to occlusion and motion parallax). This pertains to multiple image registration for MISR where we seek to relate X_{k+1} to X_k:

$$X_{k+1}(x, y) = X_k(x + v_x(x, y), y + v_y(x, y))$$

[12] An optical cavity made from two parallel reflecting surfaces that passes light only when it is in resonance with the cavity.

[13] A multidimensional array[15]. Not to be confused with the algebraic object.

[14] $m \times m$ pixel window, e.g. 3×3.

[15] CCD arrays, for example, employ 2×2 or 3×3 pixel binning, which is the practice of collapsing windows of pixels down to one pixel.

[16] "Optical flow can also be defined as the distribution of apparent velocities of movement of [patterns] in an image."[16]

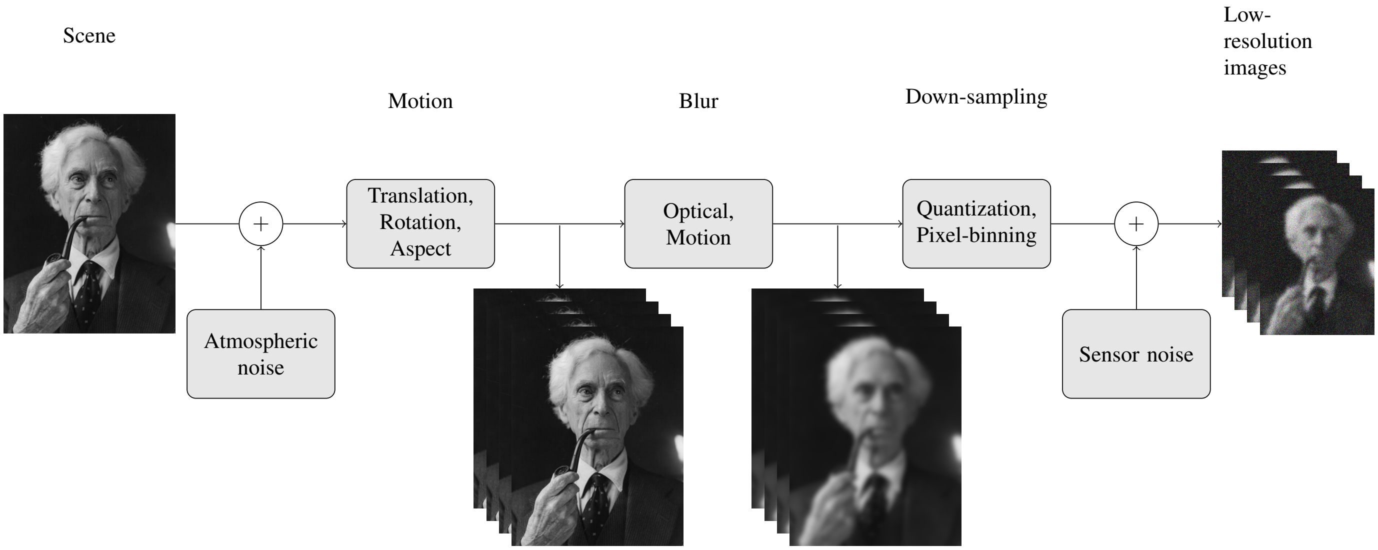

Fig. 9: The imaging model illustrating the relationship between a scene and final low-resolution images due to noise, motion, blur, and sampling.

For small motions we can approximate X_k by its first order Taylor series:

$$\begin{aligned} X_{k+1}(x,y) &= X_k(x+v_x(x,y), y+v_y(x,y)) && (2) \\ &\approx X_k(x,y) + v_x(x,y)\frac{\partial X_k}{\partial x} + v_y(x,y)\frac{\partial X_k}{\partial y} && (3) \end{aligned}$$

Evaluating equation 3 at every pixel gives a set of linear equations that enable us to fit one of the models in table 1. We focus on affine flow primarily because it is easy to estimate and secondarily because the composition of multiple affine transformations is an affine transformation (enabling us to register more than 2 images by building up the necessary transformations incrementally). Note that for nonstatic scenes the registration problem becomes "exponentially" more difficult as many more parameters need to be estimated. Furthermore registration and super resolution are not independent since the data being used to estimate the registration transforms is blurry and noisy; to wit perfectly resolved images could be much more effectively registered.

In general the optical transfer function (OTF) characterizes the blur of an imaging system[17]. We factor the OTF into three components:

$$H(u,v) = H_{\text{diff}}(u,v)H_{\text{abr}}(u,v)H_{\text{int}}(u,v) \qquad (4)$$

where u, v are horizontal and vertical spatial frequencies respectively (measured in cycles/mm), H_{diff} is blur due to diffraction, H_{abr} is blur due to lens aberrations, and H_{int} is blur due to imaging sensor shape (obtained by taking the Fourier transform of the shape an individual sensor in the imaging array). Blur due to diffraction in most imaging systems is due to diffraction through a circular aperture[1]:

$$H_{\text{diff}}(u,v) = \begin{cases} \frac{2}{\pi}\left(\frac{1}{\cos(\tau)} - \tau\sqrt{1-\tau^2}\right) & \text{if } \tau < 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\tau = \rho/\rho_c$, $\rho = \sqrt{u^2+v^2}$, $\rho_c = 1/\lambda N$ is the radial cutoff frequency of the aperture, N is the f-number[18] of the optics, and λ is the wavelength of light being diffracted. This is in fact the filter that produces the Airy pattern and therefore informs sensor array spacing in order to avoid aliasing. Wavelength independent blurring due to aberrations can be induced by various imperfections in the lenses such as spherical aberration, comatic aberration[19], or astigmatism. Furthermore, dispersion[20] blurs particular wavelengths of light. A good model for all of these effects is[18]:

$$H_{\text{abr}}(u,v) = \begin{cases} 1 - \left(\frac{25}{65}\right)^2\left(1 - 4\left(\tau - \frac{1}{2}\right)\right)^2 & \text{if } \tau < 0 \\ 0 & \text{otherwise} \end{cases}$$

Figure 10 shows an example OTF for an imaging system with a sensor spacing of 0.050 mm and therefore sampling frequency of 20 cycles/mm and $\rho_c = 83.3$ cycles/mm ($F = 3$ and $\lambda = 4\mu$m i.e. near infrared). Notice that ρ_c is much greater than the Nyquist rate ($\frac{1}{2} \times 20$ cycles/mm $= 10$ cycles/mm) and therefore many frequencies that are within the radial cutoff frequency will be aliased. This in particular can be mitigated by effectively increasing sampling rate using MISR. Notice also that like a typical transfer function the OTF is not flat and therefore attenuates high spatial frequencies. Simply applying gain to the image wouldn't solve the attenuation problems because of aliasing, but likewise this can be resolved after the effective sampling rate is increased using MISR.

The challenge of super-resolution is to solve the inverse problem of finding Y from one or several X_k. In general, since A_k, H_k, D_k are highly degenerate functions, the corresponding inverse problems are ill-posed without regularization and conditioning. The techniques that have been brought to bear on

[17] The optical transfer function is the spatial Fourier transform of the point spread function (the impulse response) of the optics. Spatial here means periodic in space rather than in time.

[18] The ratio of the system's focal length to the diameter of the aperture.

[19] Off-axis point sources appearing to have a tail (coma), due to variation in magnification in the image of the aperture stop.

[20] E.g. the cover of Pink Floyd's The Dark Side of the Moon.

Flow Type	Model	When Applicable
Affine	$v_x(x, y) = p_1x + p_2y + p_3$ $v_y(x, y) = p_4x + p_5y + p_6$	Planar scene with orthographic projection
Planar Projective	$v_x(x, y) = \dfrac{p_1 + p_2x + p_3y}{p_7 + p_8x + p_9y} - x$ $v_y(x, y) = \dfrac{p_4 + p_5x + p_6y}{p_7 + p_8x + p_9y} - y$	Planar scene with full prospective projection
Quadratic	$v_x(x, y) = \omega_Z y + \dfrac{\omega_X xy}{l} - \dfrac{\omega_Y x^2}{l} - \omega_Y l$ $\approx p_1y + p_2xy + p_3x^2 + p_4$ $v_y(x, y) = -\omega_Z x - \dfrac{\omega_Y xy}{l} + \dfrac{\omega_X y^2}{l} + \omega_X l$ $\approx p_5y + p_6xy + p_7x^2 + p_8$	Approximate for prospective projection with only $\omega_X, \omega_Y, \omega_Z$ euler angle rotations (l is focal length)
Quadratic	$v_x(x, y) \approx p_1x + p_2y + p_3x^2 + p_4xy + p_5$ $v_y(x, y) \approx p_6x + p_7y + p_8y^2 + p_9xy + p_{10}$	Approximate for planar scene with full prospective projection

Table 1: Optical flow models[17]. Note x, y are pixel coordinates and p_i are parameters that need to be estimated.

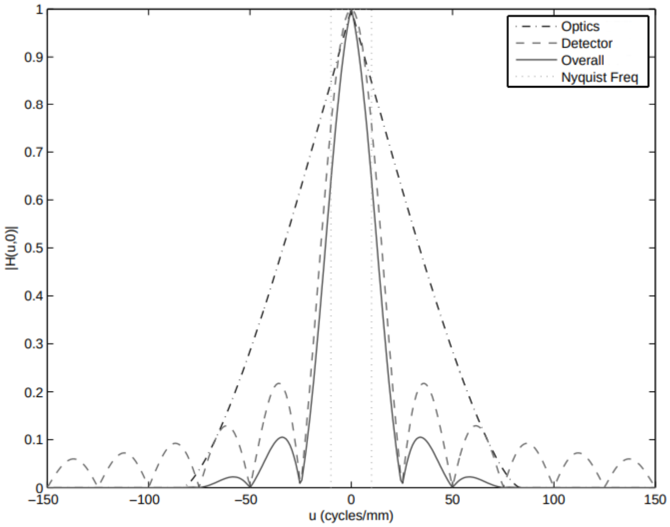

Fig. 10: OTF magnitude cross-section for[19]

the problem range from interpolation to statistical estimation to example based learning.

3 Classical Algorithms

3.1 Registration

3.2 Interpolation

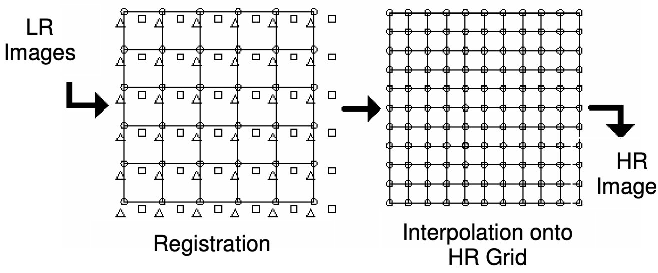

Fig. 11: LR image registration on an HR grid[21]

Suppose that H_k is linear spatial[21] and time invariant. Suppose further that A_k is affine. Then $H := H_k$ commutes with A_k[20] and eqn. 5 becomes

$$\begin{aligned} X_k &= (D \circ A_k \circ H)(Y) + \varepsilon \\ &= (D \circ A_k)(H(Y)) + \varepsilon \\ &= (D \circ A_k)(V) + \varepsilon \end{aligned} \tag{5}$$

where $V := H(Y)$. This naturally suggests interpolation in order to recover V (since X_k, in this framing, is simply shifted samples of V). Note in this context we use interpolation very broadly, i.e. to connote filling in missing values using neighboring (in some sense — not necessarily geometrically) values. This class of techniques proceed by first registering images on a high resolution grid (see figure 11) then interpolating at the "missing" pixels in the HR grid to recover V, and finally denoising and deconvolution (of H) to recover Y. Since in general consecutive X_k have non-uniform shifts (relative to X_1) the interpolation is non-uniform and improvisations on this theme use various weighting schemes for adjacent LR pixels[22].

For example Alam et al.[22] uses weighted nearest neighbors: for every pixel to be interpolated the three nearest pixels

[21] In analogy with Linear Time Invariant (i.e. linear and constant in space).

[22] An LR pixel is one sampled from an LR image and embedded in an HR grid. An HR pixel is a pixel in an HR grid.

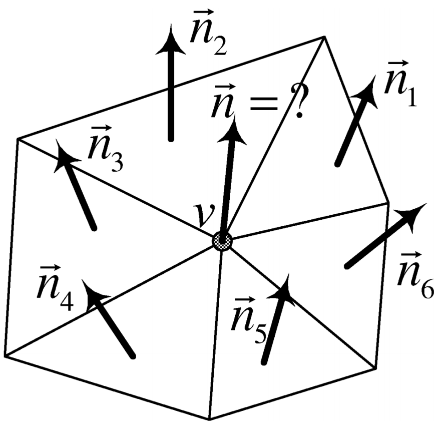

Fig. 12: Delaunay triangulation for fitting splines at LR pixels[23]. v is an LR pixel. Note that v is at z equal to the pixel value

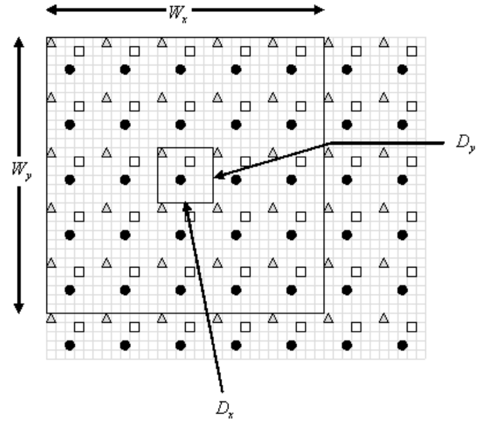

Fig. 13: Wiener filter super resolution estimation window of dimension $D_x \times D_y$ and observation window of dimension $W_x \times W_y$[24]

are weighted inversely by their distance (according to HR grid distance) and then their weighted sum is assigned to that pixel. This non-uniform interpolation is then followed by application of a Wiener filter whose design is informed by the OTF of the particular imaging system they study (which they do not estimate i.e. they assume they can model accurately). Lertrattanapanich et al.[23] base their algorithm on interpolants which require knowledge of gradients (e.g. splines) and mediate the non-uniform sampling by using a weighted average (by area) of those gradients in adjacent Delaunay cells; to be precise they produce a Delaunay triangulation of all LR pixels and compute the gradients (see figure 12) according to

$$\vec{n} = \sum_{i=1}^{m} \frac{B_j \vec{n_j}}{B} \text{ where } B = \sum_{i=1}^{m} B_i$$
$$\frac{\partial z}{\partial x} = -\frac{n_x}{n_z} \text{ and } \frac{\partial z}{\partial y} = -\frac{n_y}{n_z}$$

where B_i is the area of the ith Delaunay cell. Unfortunately this intricate solution is not robust to noise in real images.

A more sophisticated method for non-uniform interpolation uses parametric models for the auto-correlation between LR pixels and the cross-correlation between LR pixels and interpolated pixels to estimate wiener filter weights[24]. These weights are then used to average nearby pixel values. The algorithm operates on a sliding *estimation window* whose dimensions D_x, D_y are chosen such that the effective sampling rate exceeds the Nyquist rate for a given ρ_c. The pixel values for the estimation window are a function of the wiener filter weights of nearby LR pixels within an *observation window* whose dimensions W_x, W_y are an integer multiple of D_x, D_y (see figure 13). The weights $\boldsymbol{w}$ are defined as the solution to the minimum mean squared error (MMSE) filter problem, i.e. the finite impulse response (FIR) wiener filter:

$$\boldsymbol{w} = R^{-1}\boldsymbol{p} \tag{6}$$

where R is the auto-correlation of the LR pixels in the observation window and $\boldsymbol{p}$ is the cross-correlation between the pixels to be estimated and the LR pixels. Then R and $\boldsymbol{p}$ are both constructed by sampling a parametric model that weights pixels in the observation window according to distance: R is constructed by sampling from

$$C_1(r) := \sigma_d^2 \rho^r * G(r) \tag{7}$$

and $\boldsymbol{p}$ is constructed by sampling from

$$C_2(r) := \sigma_d^2 \rho^r * G(r) * G(-r) \tag{8}$$

In the case of R, r is distance on the HR grid, σ_d is related to the empirical variance of all LR pixels in a given observation window and $G(r)$ is a smoothing kernel (e.g. Gaussian). Thus by evaluating C_1 for all $r = r(n_1, n_2)$ distances between LR pixels n_1, n_2 we can construct R. Similarly for $\boldsymbol{p}$, $r = r(m, n)$ is the distance between pixel-to-be-estimated m and LR pixel n. Note that R is an $N \times N$ matrix where $N = KW_xW_y/D_xD_y$, i.e. how many LR pixels there are in the observation window, and $\boldsymbol{p}$ is an $N \times 1$ column vector uniquely computed for each pixel in the estimation window. The scheme is effective but suffers from issues with the spatial isotropy of the auto-correlation and cross-correlation models.

One of the most sophisticated of these non-uniform interpolation schemes employs the kernel regression framework and *steering kernels* (see 15). In this context we start with all X_k registered to a common HR grid and consider pixel values $Y(\boldsymbol{x}_i)$ at pixel coordinates $\boldsymbol{x}_i := (x_{i1}, x_{i2})$ as the measured data pairs $(\boldsymbol{x}_i, Y(\boldsymbol{x}_i))$. Recall that kernel regression frames the estimation problem as

$$Y(\boldsymbol{x}_i) = Z(\boldsymbol{x}_i) + \varepsilon \tag{9}$$

where Z is the to-be-estimated *regression function* that "predicts" Y as a function of $\boldsymbol{x}$. Then the Nadaraya–Watson estimator (NWE)[25] $\hat{Z}$ for Z is

$$\hat{Z}(\boldsymbol{x}) = \frac{\sum_{i=1}^{P} K(\boldsymbol{x} - \boldsymbol{x}_i) Y(\boldsymbol{x}_i)}{\sum_{i=1}^{P} K(\boldsymbol{x} - \boldsymbol{x}_i)} \tag{10}$$

where P indexes over all pixels in the HR grid and K is a *kernel function* whose purpose is to decay the contribution of $\boldsymbol{x}_i$ if it's in some sense far from $\boldsymbol{x}$. Note that $\hat{Z}(\boldsymbol{x})$ can also be seen as a weighted filtering of Y. In conventional

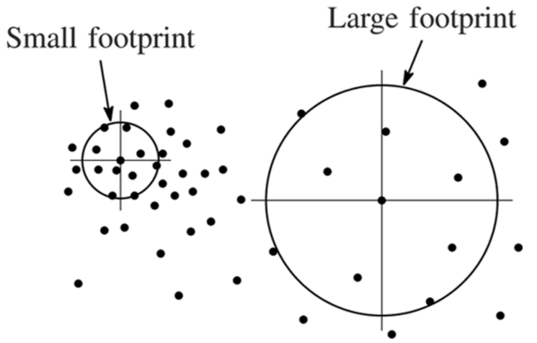

Fig. 14: Kernel footprint as a function of sample density[26]

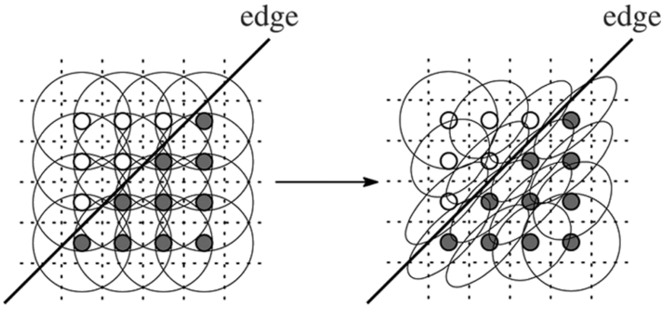

Fig. 15: Adapting kernel shape as a function of local directed structure[26]

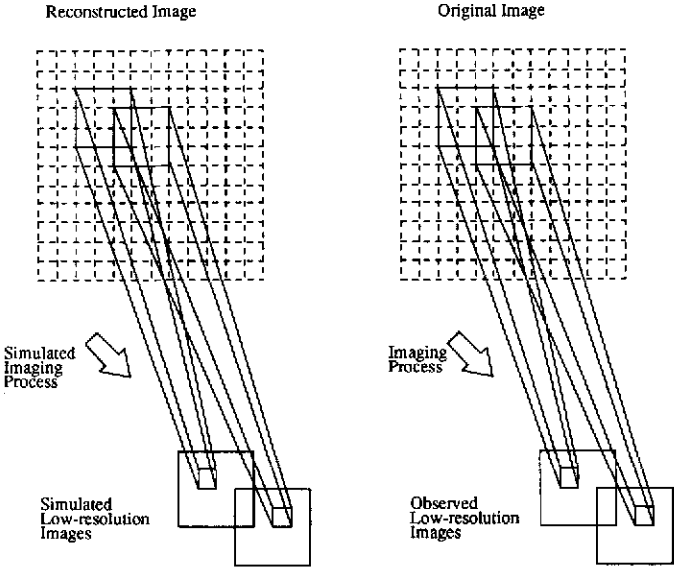
Reconstructed Image
Original Image
Simulated Imaging Process
Imaging Process
Simulated Low-resolution Images
Observed Low-resolution Images

Fig. 16: Iterative back projection[29]

kernel regression K might be any non-negative, symmetric, unimodal[27] function with augmented with an additional h parameter that controls the "bandwidth" or "footprint" of the kernel, i.e.

$$K_h(\boldsymbol{x} - \boldsymbol{x}_i) \coloneqq \frac{1}{h} K\left(h^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)\right) \tag{11}$$

This bandwidth parameter h can be generalized to a *smoothing kernel* H in order to make $K = K_H$ adaptive to the local structure of the pixels, e.g. to have larger footprints in sparsely sampled regions and have smaller footprints in densely sampled regions (see figure 14). Ultimately though it is desirable to have kernels that can adapt to directed structure in the image, i.e. "steerable" kernels that filter strongly along an edge and weakly across an edge. This is accomplished by, for example, using a Gaussian as the kernel:

$$K_{H_i}(\boldsymbol{x} - \boldsymbol{x}_i) \propto \frac{\exp\left\{-(\boldsymbol{x} - \boldsymbol{x}_i)^T H_i^{-1} (\boldsymbol{x} - \boldsymbol{x}_i)\right\}}{\sqrt{\det H_i}} \tag{12}$$

and identifying H_i with $\nabla^2 Z(\boldsymbol{x}_i)$ (since gradients capture edge structure). An estimate $\hat{H}_i$ of $\nabla^2 Z(\boldsymbol{x}_i)$ can be obtained by looking at covariances of empirical gradients (i.e. the HR grid registered image convolved with a difference filter). Unfortunately this is a naive estimate that is often rank deficient or unstable (both leading to instances where $\hat{H}_i$ isn't invertible). One solution is to parameterize H_i:

$$H_i = \gamma_i U_{\theta_i} \Lambda_{\sigma_i} U_{\theta_i}^T$$

where U_{θ_i} is a rotation matrix, $\Lambda_{\sigma_i} = \operatorname{diag}\left(\sigma_i, \sigma_i^{-1}\right)$ is an "elongation" matrix, and γ_i is a scaling parameter, with each of $\gamma_i, \theta_i, \sigma_i$ estimated from the data in a more robust way. An alternative kernel is the bilateral kernel[28] that defines closeness according to geometric and radiometric distance:

$$\begin{aligned} K_S(\boldsymbol{x} - \boldsymbol{x}_i) &\coloneqq \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma_S^2}\right\} \\ K_R(\boldsymbol{x}, \boldsymbol{x}_i) &\coloneqq \exp\left\{-\frac{\|Y(\boldsymbol{x}) - Y(\boldsymbol{x}_i)\|^2}{2\sigma_R^2}\right\} \\ K_B(\boldsymbol{x}, \boldsymbol{x}_i) &\coloneqq K_S(\boldsymbol{x} - \boldsymbol{x}_i) K_R(\boldsymbol{x}, \boldsymbol{x}_i) \end{aligned} \tag{13}$$

where σ_S parameterizes spatial distance weight and σ_R parameterizes "radiometric" distance weight.

In general non-uniform interpolation techniques are intuitive and typically (relatively) computationally efficient but they assume an unrealistic observation model (namely that of affine flow).

3.3 Estimation

Statistical estimation methods cast SR as an inference problem. One of the earliest successful SR algorithms[29] proposed an iterative scheme inspired by the back-projection method commonly used to reconstruct 2-D objects from 1-D projections in computer-aided tomography. Recall eqn. 1. Then the idea is to take the current estimate of the HR image $\hat{Y}^i$, see if after motion and down-sampling $(D \circ A_k)(\hat{Y}^i)$ it is near the LR samples X_k, and add a correction when it is not (see figure 16):

$$\hat{Y}^{i+1} = \hat{Y}^i + \sum_{k=1}^{K} (D \circ A_k)^{-1} \Big((D \circ A_k)(\hat{Y}^i) - X_k\Big) \tag{14}$$

where $\hat{Y}^i$ is the current estimate of the blurred HR image, $(D \circ A_k)(\hat{Y}^i)$ is the projection of the current estimate to low resolution, and $(D \circ A_k)^{-1}\Big((D \circ A_k)(\hat{Y}^i) - X_k\Big)$ is the *back-projection*. This process iterates until convergence i.e. $|(D \circ A_k)(\hat{Y}^i) - X_k| < \delta$ for some δ. Irani *et al.* [29]

also convolve the back-projection with a smoothing kernel as a form of regularization since the estimation problem is in general ill-posed (there are many $\hat{Y}^i$ that will project down to a pixel-distance neighbor of X_k). It can be shown[30] that for ε_k distributed $(0, R_k)$-Normal, $\hat{Y}$ is none other than the maximum likelihood estimate (MLE) for Y. We can see this by recognizing that eqn. 14 is just the Richardson iterative[31] solution to

$$L(Y) = \frac{1}{2}\left\| \boldsymbol{X} - \begin{bmatrix} D_1 \circ A_1 \\ D_2 \circ A_2 \\ \vdots \\ D_K \circ A_K \end{bmatrix} Y \right\|^2 \tag{15}$$

since

$$\nabla_Y L = 0 \iff \sum_{k=1}^{K} (D \circ A_k)^{-1}\left((D \circ A_k)(Y) - X_k\right) = 0$$

and therefore $\hat{Y}_i \to \hat{Y}$ is the MLE (since MLE is the solution to least squares[32]). Though this is one of the oldest SR algorithms it's recently been revisited and re-imagined as a deep neural network architecture (DNN)[33].

Another estimation technique employs a Kalman filter[34] to estimate Y. Let $\boldsymbol{x}_k = \text{vec}(X_k)$ be the vectorization[23] of X_k and $\boldsymbol{y} = \text{vec}(Y)$ likewise. If we assume linear models for each of D, A_k, H_k (i.e. all representable as matrices) and a well-behaved optical flow model (most pixels in image $\boldsymbol{x}_k$ appear in image $\boldsymbol{x}_{k-1}$) then we can image $\boldsymbol{y}$ as a sequence of images $\boldsymbol{y}_k$ related in time by

$$\boldsymbol{y}_k = A'_k \boldsymbol{y}_{k-1} + \eta_k \tag{16}$$

where A'_k is the *relative* motion operator, $A'_k \boldsymbol{y}_{k-1}$ is conventional matrix-vector multiplication, and η_k is the only source of new pixels (noise[24] distributed $(0, Q_k)$-Normal). Consequently eqn. 1 becomes

$$\boldsymbol{x}_k = D H_k \boldsymbol{y}_k + \varepsilon_k \tag{17}$$

and the pair of eqns. 16, 17 can be seen to constitute a linear dynamical system with $\boldsymbol{y}_k$ the state of the system, A'_k the state transition, η_k the state noise, $\boldsymbol{x}_k$ the measurement, DH_k the measurement model, and ε_k the measurement noise. Note that the HR image conditioned on all previous LR images (measurements)

$$\boldsymbol{y}_{k|s} := \boldsymbol{y}_k | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_s \tag{18}$$

with $s \le k$, is a Gaussian process (GP), with mean $\bar{\boldsymbol{y}}_{k|s} = E\left[\boldsymbol{y}_{k|s}\right]$ and covariance

$$C_{k|s} := E\left[(\boldsymbol{y}_k - \bar{\boldsymbol{y}}_{k|s})(\boldsymbol{y}_k - \bar{\boldsymbol{y}}_{k|s})^T\right] \tag{19}$$

By definition the MMSE $\hat{\boldsymbol{y}}_{k|s}$ of $\boldsymbol{y}_{k|s}$ is $\bar{\boldsymbol{y}}_{k|s}$ and therefore $C_{k|s}$ is the covariance of the error of the estimate $\hat{\boldsymbol{y}}_{k|s}$. The Kalman filter proceeds in two steps: an *a priori* (before measurement) prediction step and an *a posteriori* (after measurement) update step. The prediction step iterates on the estimate (and its covariance) given all previous measurements:

$$\hat{\boldsymbol{y}}_{k|k-1} = A'_k \hat{\boldsymbol{y}}_{k-1|k-1} \tag{20}$$

$$C_{k|k-1} = A'_k C_{k-1|k-1} (A'_k)^T + Q_k \tag{21}$$

This can be seen as a one-step propagation of the estimate "in the direction" of the previous measurement. Then the update step incorporates new information from a measurement:

$$\hat{\boldsymbol{y}}_{k|k} = \hat{\boldsymbol{y}}_{k|k-1} + K_k(\boldsymbol{x}_k - DH_k \hat{\boldsymbol{y}}_{k|k-1}) \tag{22}$$

$$C_{k|k} = (I - K_k D H_k) C_{k|k-1} \tag{23}$$

where the Kalman gain

$$K_k := \frac{C_{k|k-1}(DH_k)^T}{DH_k C_{k|k-1}(DH_k)^T + R_k} \tag{24}$$

weights the contribution of the prediction and the measurement[25]. Note that in the Kalman framework D, H_k, A'_k, R_k, Q_k are all assumed to be known. In Elad *et al.* [34] the assumption is D, H_k, R_k are known functions of camera parameters, A'_k can be estimated by an image registration algorithm and Q_k can be approximated:

$$Q_k \approx \alpha_k A'_k C_{k|k} (A'_k)^T \tag{25}$$

where α_k is chosen such that the approximation upper-bounds the true Q_k. They comment that this stronger auto-correlation for $\boldsymbol{y}_k$ adds "pseudo-noise" to the system and biases the Kalman filter to "rely" more on the measurements than the state transition model.

SR can also be posed as a Bayesian maximum a posterior (MAP) estimation problem. Let $\boldsymbol{H} := (H_1, \ldots, H_k)$ be the vector of blur operators applied to Y in order to produce $\boldsymbol{X}$ and similarly $\boldsymbol{A}$. Then

$$\begin{aligned} \hat{Y} &= \underset{Y}{\text{argmax}}\ P\left(Y|\boldsymbol{X}\right) \\ &= \underset{Y}{\text{argmax}} \int_{D,\boldsymbol{H},\boldsymbol{A}} P\left(Y, D, \boldsymbol{H}, \boldsymbol{A}|\boldsymbol{X}\right) \\ &= \underset{Y}{\text{argmax}} \int_{D,\boldsymbol{H},\boldsymbol{A}} \frac{P\left(\boldsymbol{X}|Y, D, \boldsymbol{H}, \boldsymbol{A}\right) P(Y) P(D, \boldsymbol{H}, \boldsymbol{A})}{P(\boldsymbol{X})} \end{aligned} \tag{26}$$

$$= \underset{Y}{\text{argmax}} \int_{D,\boldsymbol{H},\boldsymbol{A}} P\left(\boldsymbol{X}|Y, D, \boldsymbol{H}, \boldsymbol{A}\right) P(Y) P(D, \boldsymbol{H}, \boldsymbol{A}) \tag{27}$$

where in eqn. 26 we've used the independence of Y and $D, \boldsymbol{H}, \boldsymbol{A}$[35] and in eqn. 27 we've used that $\boldsymbol{X}$ is a constant

[23]
$$\text{vec}\left(\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}\right) = \\ [a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, \ldots, a_{1n} \ldots a_{mn}]$$

[24] Noise here doesn't necessarily mean unwanted high frequency variation but simply a source of randomness. This is in close affinity with how generative machines such as variational auto-encoders and generative adversarial networks are understood.

[25] This is better understood in the more general linear dynamic systems case where $\boldsymbol{y}_k = A_k \boldsymbol{y}_k + B_k \boldsymbol{u}_k + \varepsilon_k$ and $B_k \boldsymbol{u}_k$ is known "controlled" input. Then the prediction includes a $B_k \boldsymbol{u}_{k-1}$ term and the Kalman gain effectively mediates between controlled and uncontrolled inputs.

with respect to the maximization. While there exist reasonable priors for Y, marginalizing over $D, \boldsymbol{H}, \boldsymbol{A}$ is still difficult due to the high-dimensionality of each. Therefore assuming $D, \boldsymbol{H}, \boldsymbol{A}$ can be estimated independently as $\hat{D}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{A}}$, eqn. 27 becomes

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\ P\left(\boldsymbol{X}\middle|Y;\hat{D},\hat{\boldsymbol{H}},\hat{\boldsymbol{A}}\right)P(Y) \tag{28}$$

where the semicolon indicates $\hat{D}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{A}}$ are known parameters of the conditional distribution. This casts $\hat{Y}$ the standard MAP estimate of Y. Note that an equivalent formulation of MAP maximizes the log-likelihood instead of maximizing the likelihood:

$$\begin{aligned}\hat{Y} &= \underset{Y}{\operatorname{argmax}}\left[\log\left(P\left(\boldsymbol{X}\middle|Y;\hat{D},\hat{\boldsymbol{H}},\hat{\boldsymbol{A}}\right)P(Y)\right)\right]\\ &= \underset{Y}{\operatorname{argmax}}\left[\log P\left(\boldsymbol{X}\middle|Y;\hat{D},\hat{\boldsymbol{H}},\hat{\boldsymbol{A}}\right)+\log P(Y)\right]\end{aligned} \tag{29}$$

Various choices for $P\left(\boldsymbol{X}\middle|Y;\hat{D},\hat{\boldsymbol{H}},\hat{\boldsymbol{A}}\right)$ and the prior $P(Y)$ (and consequent choice of optimization strategies) characterize this class of SR techniques. Again assume D, H_k, A_k linear and given ε_k in eqn. 1 distributed $(0, rI)$-Normal

$$P\left(\boldsymbol{X}\middle|Y;\hat{D},\hat{\boldsymbol{H}},\hat{\boldsymbol{A}}\right) \propto \exp\left\{-\frac{\left\|\boldsymbol{X}-\hat{D}\hat{\boldsymbol{H}}\hat{\boldsymbol{A}}\boldsymbol{y}\right\|^2}{2r^2}\right\} \tag{30}$$

where here $\boldsymbol{X} = (\operatorname{vec}(X_1), \dots, \operatorname{vec}(X_k))$ and

$$\hat{D}\hat{\boldsymbol{H}}\hat{\boldsymbol{A}} := \begin{bmatrix}\hat{D}\hat{\boldsymbol{H}}_1\hat{\boldsymbol{A}}_1\\ \hat{D}\hat{\boldsymbol{H}}_2\hat{\boldsymbol{A}}_2\\ \vdots\\ \hat{D}\hat{\boldsymbol{H}}_K\hat{\boldsymbol{A}}_K\end{bmatrix} \tag{31}$$

After a suitable choice for the prior it can be seen that eqn. 29 is just regularized regression. For example when choosing a Gibbs[35] distribution as the prior, i.e.

$$P(Y) \propto e^{-\alpha B(Y)} \tag{32}$$

where $B(Y)$ is called the *potential*, eqn. 29 becomes

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmin}}\left[\left\|\boldsymbol{X}-\hat{D}\hat{\boldsymbol{H}}\hat{\boldsymbol{A}}\boldsymbol{y}\right\|^2+\lambda B(Y)\right] \tag{33}$$

where the aggregate regularization parameter λ absorbs α from $P(Y)$ and r from ε_k. There are many other choices for the prior in eqn. 29, whose effect is to bias the estimator $\hat{\boldsymbol{y}}$ towards "natural" images. Alternatively we can (and will) take eqn. 33 as our starting point and explicitly choose the regularizer $B(Y)$.

One of the simplest priors is a $(0, Q)$-Normal, where Q is symmetric positive definite[26] (PD) and captures the covariance. This corresponds to the regularizer taking the form

$$B(Y) = \boldsymbol{y}^T Q \boldsymbol{y} \tag{34}$$

where $\boldsymbol{y} = \operatorname{vec}(Y)$. Since Q is symmetric PD eqn. 33 becomes

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmin}}\left[\left\|\boldsymbol{X}-\hat{D}\hat{\boldsymbol{H}}\hat{\boldsymbol{A}}\boldsymbol{y}\right\|^2+\lambda\left\|\sqrt{Q}\boldsymbol{y}\right\|^2\right] \tag{35}$$

where $\sqrt{Q}$ is U of the Cholesky decomposition $Q = U^T U$. Equation 35 is Tikhonov regularized regression. Letting $G = \hat{D}\hat{\boldsymbol{H}}\hat{\boldsymbol{A}}$ the closed-form solution to eqn. 35 is

$$\hat{\boldsymbol{y}} = \left(G^T G + \lambda Q\right)^{-1} G^T \boldsymbol{X} \tag{36}$$

Nguyen *et al.* [36] use cross-validation to determine the regularization parameter λ (by partitioning the pixels into a "fit" and "validate" set). In general, rather than explicitly computing inverses in eqn. 36, in practice $\hat{\boldsymbol{y}}$ is found by solving the regression problem via optimization (for high-dimensional matrices optimization is faster than inversion). Nguyen *et al.* use conjugate gradient descent[27] to optimize eqn. 35. They argue that $\hat{D}\hat{\boldsymbol{H}}\hat{\boldsymbol{A}}$ is ill-conditioned[28] and to that end, since the convergence rate of conjugate gradients is dependent on the condition number[37], propose pre-conditioners[29] to improve the converge rate.

An issue with the multivariate Normal prior is that it strongly enforces global smoothness, penalizing sharp edges. One solution is to use Huber loss[38]

$$L_\delta(a) := \begin{cases} a^2 & \text{for } |a| \le \delta \\ 2\delta|a| - \delta^2 & \text{otherwise} \end{cases} \tag{37}$$

to explicitly parameterize the penalty for gradients (i.e. high-frequency features). Huber loss enforces local smoothness (since it's quadratic for $|a| \le \delta$) but permits edges (since it's linear for $|a| > \delta$). Capel *et al.* [39] implement this by composing $L_\delta(a)$ with a first-order gradient operator[30] as the potential in eqn. 32. Another gradient penalty that encourages sparse gradients (i.e. local smoothness and steep edges) is Total Variation (TV) norm[40]:

$$\|u\|_{\text{TV}} := \int_\Omega \|\nabla u\|\, d\Omega \tag{38}$$

where u is a smooth image of bounded variation (i.e. such that the integral converges) over domain Ω. In the context of SR this amounts to setting

$$B(Y) = TV(Y) := \|\nabla Y\|_1 \tag{39}$$

Farsiu *et al.* [41] introduce a *bilateral* TV norm

$$BTV(Y) := \sum_{k=0}^{N}\sum_{l=0}^{N}\alpha^{k+l}\|Y - S_x^k S_y^l Y\|_1 \tag{40}$$

where S_x^k, S_y^l are shift operators (i.e. S_x^k shifts Y by k pixels in the horizontal) and N is arbitrary. The bilateral TV norm factors in gradients at several scales ($Y - S_x^2 Y$ is an approximation of the horizontal gradient at twice the scale of $Y - S_x^1 Y$) and decays their contribution according to their

[26] A symmetric real matrix Q is positive definite if $\boldsymbol{y}^T Q \boldsymbol{y} > 0$ for all non-zero $\boldsymbol{y}$.

[27] Two vectors $\boldsymbol{u}, \boldsymbol{v}$ are conjugate with respect to G if $\boldsymbol{u}^T G \boldsymbol{v} = 0$. Conjugate gradient descent is gradient descent but with conjugate gradients (it has better convergence properties).

[28] The condition number κ of a function is a measure of how sensitive it is to small perturbations; for a matrix G it is defined $\kappa(G) := \sigma_{\max}(G)/\sigma_{\min}(G)$ where $\sigma_{\max}(G), \sigma_{\min}(G)$ are the maximum and minimum singular values of G respectively.

[29] A pre-conditioner of a matrix G is an approximation of G that has a better condition number. Nguyen *et al.* propose a pre-conditioner with singular values clustered around 1 in order that $\kappa(G) \approx 1$.

[30] Let $u = (1, 2, 1)$ and $v = (1, 0, -1)$ then $h = uv^T$ is the first order Sobel filter and $\nabla Y = \left(h * Y, h^T * Y\right)$.

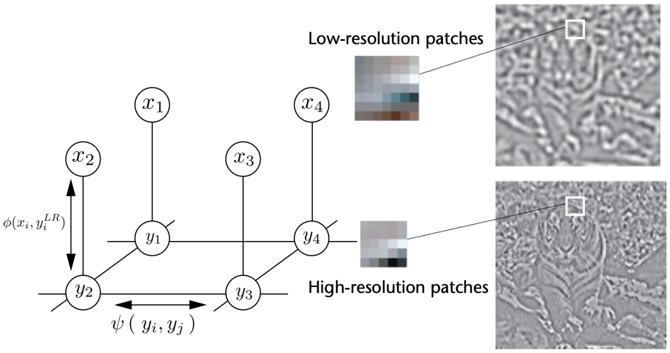

Fig. 17: Hidden Markov random field modeling spatial relationships between LR patches x_i and example LR patches y_i^{LR} and HR patches y_i[43]

spatial distance (α^{k+l} decays as a function of shift distance). It can also be shown to be equivalent[42] to filtering $\boldsymbol{X}$ by the bilateral kernel (see eqn. 13).

Though there are many priors (or regularizers) that have been studied none really capture natural images in all of their variation. For that we need to actually learn from real data i.e. examples images.

3.4 Example based

Example based techniques *learn* a mapping from LR patches to HR patches based on a training set and then use that mapping to predict details in new (unseen) images. Freeman *et al.* [43] argues that the most important features to reconstruct are the high-frequency features. Therefore they store only the correspondence between high-pass filtered, contrast-normalized versions of example LR patches and HR patches. Note that they pre-process the LR images by cubic-spline interpolating to the higher pixel sampling resolution and that HR patches are sampled to overlap by one or more pixel widths at their edges. Naive mosaicing of these matching HR patches produces poor results due to noise and the ill-posed nature of super-resolution. Their solution to this issue is using a hidden Markov random field[31] (HMRF) to model spatial consistency between adjacent HR patches y_i and between LR-HR patch pairs x_i, y_i (see figure 17). They then compute the MAP estimate of the HMRF to obtain the smoothest assignment of HR patches: the HMRF model postulates that the conditional probability of any HR patch assignment $\boldsymbol{y}$ given observed LR patches $\boldsymbol{x}$ is

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{P(\boldsymbol{x})} \prod_{i \bullet\!\!-\!\!\bullet j} \psi(y_i, y_j) \prod_i \phi(x_i, y_i^{LR}) \tag{41}$$

where x_i are observed LR patches and y_i are unknown (to-be-inferred) HR patches (along with their learned-mapping example LR patches y_i^{LR}). Note $P(\boldsymbol{x})$ is a normalization constant, $i \bullet\!\!-\!\!\bullet j$ indicates the product is only over adjacent HR patches, $\psi(y_i, y_j)$ encodes the compatibility of adjacent HR patches according to

$$\psi(y_i, y_j) = \exp\left\{-\frac{\|y_i - y_j\|}{2\sigma^2}\right\} \tag{42}$$

where $\|y_i - y_j\|$ is only computed on the overlapping pixels, and $\phi(x_i, y_i^{LR})$ similarly encodes the compatibility between the observed LR patch x_i and the example LR patch y_i^{LR} corresponding to the estimated HR patch y_i. To make the MAP computation tractable they choose only 16 candidate example LR patches y_i^{LR}. They compute the MAP estimate using belief propagation[32].

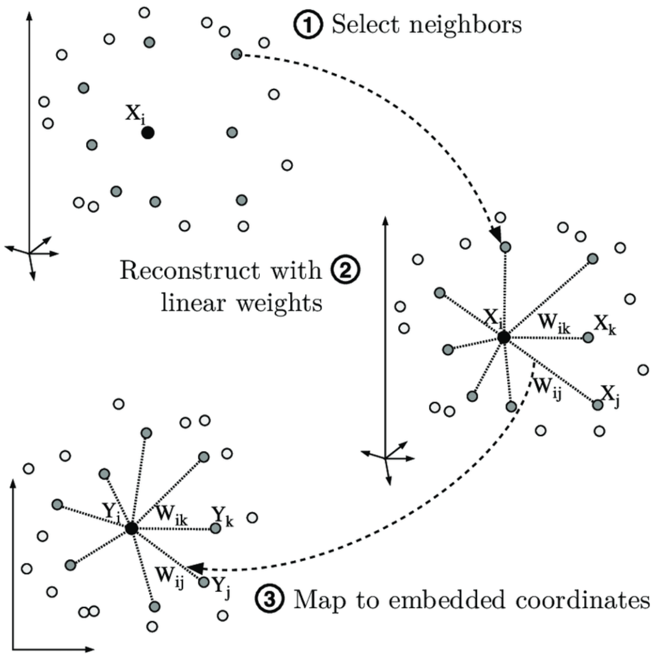

Fig. 18: NNE algorithm from LR patches x_i to HR patches y_i[45]

The main problem with LR-HR patch pair technique is that it necessitates an enormous database of patches in order for it to generalize. Chang *et al.* remedy this problem by using ideas from a manifold[33] learning[34] technique called locally linear embedding[46] (LLE) that they call nearest neighbor embedding (NNE). First they characterize both the space of LR patches and HR patches as manifolds. NNE (and LLE) is based on the intuition that a well-sampled smooth manifold is locally linear: a LR patch x_i on the LR manifold and its neighbors $x_j, x_k, \ldots$ lie in a locally linear subspace of the manifold. Given this assumption, one can characterize the local geometry of that subspace by finding convex reconstruction weights W_{ij} of any x_i from only its neighbors: by minimizing

[31] A Markov random field (MRF) is a collection of random variables $x_i, y_j, \ldots$ with conditional dependence represented by pairings and satisfying the *pairwise Markov* property: any two random variables that aren't paired are conditionally independent of each other given (conditioned on) all other variables. A *hidden* Markov random field is simply a MRF where some of the random variables aren't observed.

[32] An efficient way to compute marginals of joint probabilities that have structure by reusing partial sums (i.e. passing messages)[44].

[33] A collection of points that locally resembles Euclidean space.

[34] Dimensionality reduction.

$$\hat{W} = \underset{W}{\operatorname{argmin}} \sum_i \left\| x_i - \sum_{x_j \in N(x_i)} W_{ij} x_j \right\|^2 \text{ s.t.} \tag{43}$$

$$\sum_{x_j \in N(x_i)} W_{ij} = 1 \text{ for all } i$$

where $N(x_i)$ is the neighborhood of x_i. These weights W_{ij} are invariant with respect to rotation, rescaling, and translation of the LR patch and its neighborhood[46] and therefore should remain valid in an embedding space (i.e. the HR manifold) coordinate system (see figure 18). The patches y_i in the HR manifold are then linear functions of the same reconstructions weights, i.e.

$$y_i = \sum_{x_j \in N(x_i)} W_{ij} y_j \tag{44}$$

Note that neighboring patches in HR space are constrained to overlap and in fact the neighborhoods are computed using gradient features of the LR and HR patches rather than the raw patches (they argue this allows for better generalization i.e. smaller example sets).

4 Deep Learning Algorithms

5 Future Research

6 Conclusion

7 Appendix

TODO: work out diffraction circular aperture TODO: workout poisson noise

Acknowledgments

References

[1] J. W. Goodman, *Introduction to Fourier Optics*, 3rd. Roberts & Co. Publishers, 2005, pp. 76–78.

[2] P. Scholz, “Focused ion beam created refractive and diffractive lens techniques for the improvement of optical imaging through silicon,” PhD thesis, Jul. 2012. DOI: 10.14279/depositonce-3270.

[3] J. Kennedy, O. Israel, A. Frenkel, R. bar-shalom, and H. Azhari, “Improved image fusion in pet/ct using hybrid image reconstruction and super-resolution,” *International journal of biomedical imaging*, vol. 2007, p. 46 846, Jan. 2007. DOI: 10.1155/2007/46846.

[4] L. R. F.R.S., “Xxxi. investigations in optics, with special reference to the spectroscope,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 8, no. 49, pp. 261–274, 1879. DOI: 10.1080/14786447908639684. eprint: https://doi.org/10.1080/14786447908639684. [Online]. Available: https://doi.org/10.1080/14786447908639684.

[5] D. L. Fried, “Optical resolution through a randomly inhomogeneous medium for very long and very short exposures,” *J. Opt. Soc. Am.*, vol. 56, no. 10, pp. 1372–1379, Oct. 1966. DOI: 10.1364/JOSA.56.001372. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=josa-56-10-1372.

[6] E. Van Reeth, I. W. K. Tham, C. H. Tan, and C. L. Poh, “Super-resolution in magnetic resonance imaging: A review,” *Concepts in Magnetic Resonance Part A*, vol. 40A, no. 6, pp. 306–325, 2012. DOI: 10.1002/cmr.a.21249. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cmr.a.21249. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cmr.a.21249.

[7] J. Shermeyer and A. V. Etten, “The effects of super-resolution on object detection performance in satellite imagery,” *CoRR*, vol. abs/1812.04098, 2018. arXiv: 1812.04098. [Online]. Available: http://arxiv.org/abs/1812.04098.

[8] M. Robbins, “Final test guideline,” May 2014.

[9] M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. Macdonald, V. Mahajan, and E. Van Stryland, *Handbook of Optics, Third Edition Volume I: Geometrical and Physical Optics, Polarized Light, Components and Instruments(Set)*, 3rd ed. New York, NY, USA: McGraw-Hill, Inc., 2010, ISBN: 0071498893, 9780071498890.

[10] J. R. Janesick, T. Elliott, S. Collins, M. M. Blouke, and J. Freeman, “ Scientific Charge-Coupled Devices,” *Optical Engineering*, vol. 26, no. 8, pp. 692–714, 1987. DOI: 10.1117/12.7974139. [Online]. Available: https://doi.org/10.1117/12.7974139.

[11] J. Pawley, *Handbook of Biological Confocal Microscopy*, ser. Cognition and Language. Springer, 1995, pp. 918–919, ISBN: 9780306448263. [Online]. Available: https://books.google.com/books?id=16Ft5k8RC-AC.

[12] Y. Xu, “Fundamental characteristics of a pinned photodiode cmos pixels,” 2015.

[13] Y. E. Kesim, E. Battal, M. Y. Tanrikulu, and A. K. Okyay, "An all-zno microbolometer for infrared imaging," *Infrared Physics & Technology*, vol. 67, pp. 245–249, 2014, ISSN: 1350-4495. DOI: https://doi.org/10.1016/j.infrared.2014.07.023. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350449514001479.

[14] R. Ambrosio, M. Moreno, J. Mireles Jr., A. Torres, A. Kosarev, and A. Heredia, "An overview of uncooled infrared sensors technology based on amorphous silicon and silicon germanium alloys," *physica status solidi c*, vol. 7, no. 3-4, pp. 1180–1183, 2010. DOI: 10.1002/pssc.200982781. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pssc.200982781. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pssc.200982781.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016, p. 33, ISBN: 0262035618, 9780262035613.

[16] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981, ISSN: 0004-3702. DOI: https://doi.org/10.1016/0004-3702(81)90024-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0004370281900242.

[17] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998, ISBN: 0132611082.

[18] R. R. Shannon, " Aberrations And Their Effects On Images," in *Geometrical Optics*, R. E. Fischer and W. J. Smith, Eds., International Society for Optics and Photonics, vol. 0531, SPIE, 1985, pp. 27–48. DOI: 10.1117/12.946501. [Online]. Available: https://doi.org/10.1117/12.946501.

[19] P. Milanfar, *Super-Resolution Imaging*, ser. Digital Imaging and Computer Vision. CRC Press, 2017, pp. 44–45, ISBN: 9781439819319. [Online]. Available: https://books.google.com/books?id=fjTUbMnvOkgC.

[20] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1187–1193, Aug. 2001. DOI: 10.1109/83.935034.

[21] S.-C. Lin and C.-T. Chen, " Reconstructing Vehicle License Plate Image from Low Resolution Images using Nonuniform Interpolation Method," Tech. Rep. 1, p. 21.

[22] 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," Tech. Rep. 5, 2000.

[23] S. Lertrattanapanich and N. K. Bose, "High resolution image formation from low resolution frames using delaunay triangulation," *IEEE Transactions on Image Processing*, vol. 11, no. 12, pp. 1427–1441, Dec. 2002. DOI: 10.1109/TIP.2002.806234.

[24] R. Hardie, "A fast image super-resolution algorithm using an adaptive wiener filter," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2953–2964, Dec. 2007. DOI: 10.1109/TIP.2007.909416.

[25] E. Nadaraya, " On Estimating Regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964. DOI: 10.1137/1109020. [Online]. Available: https://doi.org/10.1137/1109020.

[26] H. Takeda, S. Member, S. Farsiu, P. Milanfar, and S. Member, " Kernel Regression for Image Processing and Reconstruction," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 16, no. 2, p. 349, 2007. DOI: 10.1109/TIP.2006.888330. [Online]. Available: http://www.soe.ucsc.edu/%20%7B%20~%20%7D%20htakeda..

[27] M. Wand and M. Jones, *Kernel Smoothing*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994, ISBN: 9780412552700. [Online]. Available: https://books.google.com/books?id=GTOOi5yE008C.

[28] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98, Washington, DC, USA: IEEE Computer Society, 1998, pp. 839–, ISBN: 81-7319-221-9. [Online]. Available: http://dl.acm.org/citation.cfm?id=938978.939190.

[29] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Model and Image Processing*, vol. 53, pp. 231–239, 1991.

[30] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646–1658, Dec. 1997. DOI: 10.1109/83.650118.

[31] R. S. Anderssen and G. H. Golub, "Richardson"s non-stationary matrix iterative procedure.," Stanford, CA, USA, Tech. Rep., 1972.

[32] G. Casella and R. Berger, *Statistical Inference*. Duxbury Resource Center, Jun. 2001, ISBN: 0534243126.

[33] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," *CoRR*, vol. abs/1803.02735, 2018. arXiv: 1803.02735. [Online]. Available: http://arxiv.org/abs/1803.02735.

[34] M. Elad and A. Feuer, "Super-resolution reconstruction of image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 817–834, Sep. 1999. DOI: 10.1109/34.790425.

[35] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, " Joint MAP Registration and High-Resolution Image Estimation Using a Sequence of Undersampled Images," Tech. Rep. 12, 1997.

[36] Nhat Nguyen, P. Milanfar, and G. Golub, "A computationally efficient superresolution image reconstruction algorithm," *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 573–583, Apr. 2001. DOI: 10.1109/83.913592.

[37] A. van der Sluis and H. A. van der Vorst, " The rate of convergence of Conjugate Gradients," *Numerische Mathematik*, vol. 48, no. 5, pp. 543–560, Sep. 1986,

ISSN: 0945-3245. DOI: 10.1007/BF01389450. [Online]. Available: https://doi.org/10.1007/BF01389450.

[38] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964. DOI: 10 . 1214 / aoms / 1177703732. [Online]. Available: https://doi.org/10.1214/aoms/1177703732.

[39] D. Capel and A. Zisserman, "Super-resolution enhancement of text image sequences," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 1, Sep. 2000, 600–605 vol.1. DOI: 10.1109/ICPR.2000.905409.

[40] L. I. Rudin, S. Osher, and E. Fatemi, " Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992, ISSN: 0167-2789. DOI: https : / / doi . org / 10 . 1016 / 0167 - 2789(92) 90242 - F. [Online]. Available: http : / / www . sciencedirect . com / science / article / pii / 016727899290242F.

[41] 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, Oct. 2004. DOI: 10 . 1109 / TIP. 2004 . 834669.

[42] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Transactions on Image Processing*, vol. 11, no. 10, pp. 1141–1151, Oct. 2002. DOI: 10.1109/TIP.2002.801126.

[43] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, no. 2, pp. 56–65, 2002.

[44] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring Artificial Intelligence in the New Millennium*, G. Lakemeyer and B. Nebel, Eds., Morgan Kaufmann Publishers, Jan. 2003, ch. 8, pp. 239–236, ISBN: 1-55860-811-7. [Online]. Available: https://www.merl.com/publications/TR2001-22.

[45] G. Donatti, "Memory organization for invariant object recognition and categorization," PhD thesis, Jun. 2016. DOI: 10.13140/RG.2.1.2880.7921.

[46] L. K. Saul and S. T. Roweis, "An introduction to locally linear embedding," *unpublished*, 2000. [Online]. Available: https://cs.nyu.edu/~roweis/lle/papers/lleintro.pdf.