Distilling Information with Super-Resolution for Video Surveillance

Marco Cristani
Dipartimento di Informatica
Università di Verona
cristanm@sci.univr.it

Dong Seon Cheng
Dipartimento di Informatica
Università di Verona
cheng@sci.univr.it

Vittorio Murino
Dipartimento di Informatica
Università di Verona
vittorio.murino@univr.it

Donato Pannullo
Università di Verona
donato.pannullo@students.univr.it

ABSTRACT

A video surveillance sequence generally contains a lot of scattered information regarding several objects in cluttered scenes. Especially in case of use of digital hand-held cameras, the overall quality is very low due to the unstable motion and the low resolution, even if multiple shots of the desired target are available.

To overcome these limitations, we propose a novel Bayesian framework based on image super-resolution, that integrates all the informative bits of a target and condenses the redundancy. We call this process *distillation*.

In the traditional formulation of the image super-resolution problem, the observed target is (1) always the same, (2) acquired using a camera making small movements, and (3) the number of available images is sufficient for recovering high-frequency information. These hypotheses obviously do not hold in the concrete situations described above.

In this paper, we extend and generalize the image super-resolution task, embedding it in a structured framework that accurately distills the necessary information. In short, our approach is composed by two phases. First, a transformation-invariant video clustering coarsely groups and registers the frames, also defining a similarity concept among them. Second, a novel Bayesian super-resolution method uses this concept in order to combine selectively all the pixels of similar frames, whose result consists in a highly informative super-resolved image of the desired target.

Our approach is first tested on synthetic data, obtaining encouraging comparative results with respect to known super-resolution techniques and a definite robustness against noise. Second, real data coming from videos taken by a hand-held camera are considered, trying to solve the major details of a person in motion, a typical setting of video surveillance applications.

VSSN'04, October 15, 2004, New York, New York, USA.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*; I.4.3 [**Image Processing and Computer Vision**]: Enhancement

General Terms

Algorithms

Keywords

Super-resolution, generative model, video surveillance, machine learning.

1. INTRODUCTION

In the emerging field of video surveillance, the quality of video frames represents the essential source of information to identify, classify, or recognize targets of interest, like objects or people. The widespread use of low-cost hand-held cameras, web-cams, and mobile phones with cameras has multiplied the sources of videos' production, but at the cost of a lower quality. The reasons are several, starting from careless acquisitions from unexperienced operators to low resolution cameras with low gains. The analysis of this data is problematic at best, and useless in many cases, because the noise and the resolution may not allow any meaningful processing, even if the target turns out in a bunch of frames.

The problem of obtaining a highly informative image starting from noisy and coarsely resolved input images is known in literature as image super-resolution. When the input consists in only one low resolution image, we refer to the problem as "single-frame super-resolution" [10], when several frames are considered, we call it "multi-frame super-resolution" or simply super-resolution [12, 1]. There are other kinds of super-resolution currently on study, for example, the super-resolution enhancement of video [13, 3] whose goal consists in improving the quality of each single frame through the addition of high frequency information.

Recently, the attention devoted to the development of super-resolution algorithms is sensibly grown, in both the single image [9], and the multi-frame cases [5, 6, 14]. Clearly, in the latter case, the information encoded in the resulting image is considerably larger, giving a more accurate representation.

In video surveillance, several tasks of recognition and detection are in fact based on visual data [11]. For example, when the task consists in detecting the identity of a person captured with some camera device, highly detailed images are desirable. For this reason, the application of super-resolution techniques in this field is highly relevant.

In general, all super-resolution algorithms are based on three basic hypotheses:

1. all the images must portray the same scene, meaning that they can be compared without being deceived;

2. small movements of the scene should be present across images, such that each provides a slightly different "point of view" that can be integrated; in case of known large movements, this constraint may be relaxed by pre-registering the images;

3. the number of available images should be sufficient for recovering high frequency information.

These constraints are quite hard and penalizing, in particular for a video sequence, making the super-resolution image estimation possible only in supervised and controlled conditions, strongly reducing its applicability in wider contexts. For the sake of clarity, in the following we will use the term super-resolution as the process of combining several low resolution (LR) images in order to produce a higher resolution (HR) one, only when all the above constraints are satisfied. Under these circumstances, the super-resolution process inverts the generative process model in which the LR frames are generated from the HR one, when correctly warped, subsampled and blurred by the Point Spread Function (PSF) of the camera device [1].

The work present in the literature devoted to this problem are many and various, although it is possible to group the existing approaches into two subfields: in the first group, the alignment of the LR images (with known or estimated parameters) is separated from the fusion step, that estimates the remaining parameters (like the PSF width) [4]; in the second group, all the parameters are jointly estimated [8, 14]. In this second category the HR estimation is either performed by a maximum-likelihood (ML) approach, or in a Maximum A-Posteriori (MAP) fashion, regularizing the ill-conditioning of the ML framework using some Bayesian prior [1, 6]. Lastly, full Bayesian approaches are proposed, that explicitly take into account the uncertainty in the estimation of all the parameters driving the process of super-resolution [14]. In any case, the effectiveness of the above methods is based on the satisfaction of the above three basic constraints. When one of the three hypotheses is missing, the efficacy of any of the super-resolution methods fails seriously, as we see in the following.

In this paper, we show how it is possible to recover these constraints in arbitrary video sequences, by using a fully Bayesian framework, in order to acquire highly detailed information about the target of interest. Basically, with the term "target" we mean an object, person or, generally, an entity, present in most of the frames of the sequence. The idea is that the most frequent an object appears in the scene, the most presumably the object represents the target of interest whose information should be increased.

Our method is composed of two phases: in the first step, the clustering step, we exploit the transformed-invariant video clustering proposed in [7], using the transformed Gaussian model, in which each single frame of a video sequence is considered as generated by a representative ("mean") image, subject to a discrete transformation and sensor noise addition. After the clustering step the following information becomes available:

1. a low resolution image of each persistent target, over which it will be possible to obtain super-resolved information; such image is the mean of each Gaussian cluster, normalized with respect to invertible transformations;

2. the frames representing each target in the sequence, also normalized with respect to the invertible transformation;

3. the accuracy measure that exploits the goodness with which the targets are represented in the related frames, encoded in the covariance of the cluster;

4. a coarse alignment of the target for each frame, due to the normalization carried out.

This clustering step permits us to recover the three fundamental hypotheses required to place our data in a super-resolution framework, plus a similarity measure between frames and the mean representation, encoded by the Mahalanobis distance. Then, in the second step, the actual super-resolution task is carried out using the grouped frames of each cluster and actively embedding the accuracy measure, in order to build higher resolved images for each cluster, taking differently into account the value of each pixel of the frames involved in the process. The idea is that the more present a value in the normalized versions of the grouped LR frames, the more probably this pixel holds an important role in the HR image reconstruction step.

The most related super-resolution structure used as a basis for the proposed method is presented in [14]. In this work, a video of a static scene is considered subject to translation/rotation changes, even if arbitrary transformations can be dealt as well. The problem is managed in a full Bayesian approach, by marginalizing over the HR image to determine the LR images' registration parameters, and, possibly, the PSF parameters. Our method actively develops this structure, introducing a novel likelihood term in the Bayesian structure that weights differently each pixel in the estimation of the HR image, proportionally to its accuracy measure. Unlike the previous approaches and owing to the generative process applied to the input video sequence, the proposed probabilistic framework is able to recover super-resolution images of well defined targets in a fully automatic, efficient, and flexible manner. This is the main contribution of the present work.

The rest of the paper is organized as follows. In Sec. (2), the classical generative process is reported, which constitutes the fundamental basis of the proposed method. Its extension is described in Sec. (3), by introducing a novel likelihood term and its use in the context of the Bayesian framework customised to deal with the super-resolution problem. In Sec. (4), the method is tested and results are shown, also in comparative terms with respect to state-of-the-art algorithms. Finally, in Sec. (5), conclusions are drawn and future perspectives are envisaged.

2. THE BAYESIAN SUPER-RESOLUTION FRAMEWORK

When the three basic constraints are satisfied, we may conveniently represent a sequence of K LR images as the vectors $\mathbf{y}_k$, $k = 1, \ldots, K$, obtained by raster-scanning the images. The pixels intensities in the LR images are constrained to lie in the range $(-0.5, 0.5)$. This operation allows us to simplify some mathematical formulae in the following.

Given M image pixels, we want to produce a HR unrolled vector $\mathbf{x}$ made by N pixels, with $N \gg M$ depending on the linear (integer) magnification factor q.

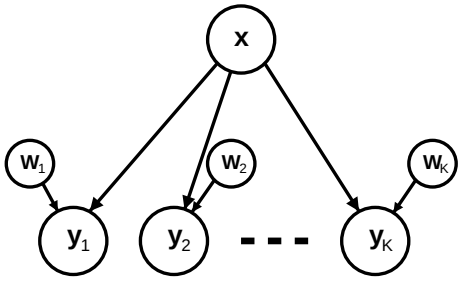

Figure 1: Super-resolution generative model.

The generative model is composed by a prior over the HR image together with an observation model (see Fig. 1) that explains each observed $\mathbf{y}_k$ as being generated independently from $\mathbf{x}$ by first applying a shift $\mathbf{s}_k$ and a rotation θ_k, then convolving with a Gaussian PSF with width γ, and finally downsampling to the lower resolution. All these steps are encoded in the downsampling transformation matrix $\mathbf{W}_k$, such that this model may be concisely described by the equation

$$\mathbf{y}_k = \mathbf{W}_k \mathbf{x} + \epsilon_k \tag{1}$$

where $\epsilon_k \sim \mathcal{N}(0, \mathbf{\Lambda})$ represents an independent noise term capturing both camera noise and the uncertainty in the generative process. $\mathbf{\Lambda}$ is a diagonal covariance matrix with variance elements equal to λ^2, describing the (per pixel) *uncertainty*. This parameter is usually heuristically set a priori.

The likelihood of observing an image $\mathbf{y}_k$ in the given model is described by the density

$$p(\mathbf{y}_k|\mathbf{x}, \mathbf{s}_k, \theta_k, \gamma) \sim \mathcal{N}(\mathbf{y}_k; \mathbf{W}_k\mathbf{x}, \mathbf{\Lambda})$$

where

$$\mathcal{N}(\mathbf{y}_k; \mathbf{W}_k\mathbf{x}, \mathbf{\Lambda}) = \left(\frac{1}{2\pi\lambda}\right)^{\frac{M}{2}} \exp\left\{-\frac{\| \mathbf{y}_k - \mathbf{W}_k\mathbf{x} \|^2}{2\lambda^2}\right\} \tag{2}$$

In the $M \times N$ downsampling transformation matrix $\mathbf{W}_k$ the m-th row, $m = 1, \ldots, M$, produces the m-th LR pixel value, convolving the N HR pixels with a gaussian mask centered according to the transformation parameters $\mathbf{s}_k, \theta_k$ and whose span is determined by the PSF width parameter γ. If we denote the matrix element located at the m-th row and n-th column with the superscript $^{<mn>}$, we have:

$$\mathbf{W}_k^{<mn>} = \frac{\widetilde{\mathbf{W}}_k^{<mn>}}{\sum_{n'=1}^{N} \widetilde{\mathbf{W}}_k^{<mn'>}} \tag{3}$$

where each m-th row has been normalized from

$$\widetilde{\mathbf{W}}_k^{<mn>} = \exp\left\{\frac{||\mathbf{v}^{<n>} - \mathbf{u}_k^{<m>}||^2}{\gamma^2}\right\} \tag{4}$$

The above equation clearly explains why γ is being referred to as the 'width' of the PSF: the larger its value, the larger the number of neighboring pixels to be weighted, thus, the resulting image results more blurred.

In Eq. (4) the vector $\mathbf{v}^{<n>}$ is the position of the n-th HR pixel, while the vector $\mathbf{u}_k^{<m>}$ is the center of the PSF and is located according to the following transformation:

$$\mathbf{u}_k^{<m>} = \mathbf{R}_k(\mathbf{u}^{<m>} - \overline{\mathbf{u}}) + \overline{\mathbf{u}} + \mathbf{s}_k \tag{5}$$

where $\mathbf{u}^{<m>}$ is the position of the m-th LR pixel, rotated around the center $\overline{\mathbf{u}}$ (coincident for simplicity with the centre of the image) by the matrix

$$\mathbf{R}_k = \begin{bmatrix} cos\theta_k & sin\theta_k \\ -sin\theta_k & cos\theta_k \end{bmatrix} \tag{6}$$

and then shifted by the amount $\mathbf{s}_k$. In both Eqs. (4) and (5) the vectors $\mathbf{v}^{<n>}$ and $\mathbf{u}^{<m>}$ are chosen once and for all images on a regular grid based only on the resolution.

The Bayesian approach to the super-resolution estimates the high resolution image $\mathbf{x}$ inverting the ill-posed system (1), considering all the K LR images:

$$p(\mathbf{x}|\{\mathbf{y}_k, \mathbf{s}_k, \theta_k\}, \gamma) = \frac{p(\mathbf{x}) \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}, \mathbf{s}_k, \theta_k, \gamma)}{p(\{\mathbf{y}_k\}|\{\mathbf{s}_k, \theta_k\}, \gamma)} \tag{7}$$

$$= \frac{p(\mathbf{x}) \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}, \mathbf{s}_k, \theta_k, \gamma)}{\int_{\mathbf{x}} p(\mathbf{x}) \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}, \mathbf{s}_k, \theta_k, \gamma)} \tag{8}$$

where the posterior over the HR image $\mathbf{x}$, given the properly shifted, rotated LR images and the width of the PSF, is calculated as the product of the regularizer prior and the K likelihood terms, as usual in the MAP approach, normalized by a marginal factor that represents the uncertainty over the parameters.

The prior over the HR image $p(\mathbf{x})$ provides a more constrained solution space fixing the ill-conditioned system of the likelihood terms. Much effort has been spent in literature to study case-oriented priors [1, 4], but, to deal with arbitrary sequences, a Gaussian prior is adopted [14], that constrains the correlation of nearby pixels:

$$p(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{Z}_\mathbf{x}) \tag{9}$$

with the covariance matrix $\mathbf{Z}_\mathbf{x}$ having the form

$$Z_{i,j} = A \exp\left\{-\frac{\| \mathbf{v}^{<i>} - \mathbf{v}^{<j>} \|^2}{r^2}\right\} \tag{10}$$

Here $\mathbf{v}^{<i>}$, $\mathbf{v}^{<j>}$ denote as before the spatial positions of HR pixels i and j in the two-dimensional image space; A and r represent the weakness of the prior and the correlation range, respectively.

Considering the whole term in Eq. (7) as the product between the prior and the joint likelihood Gaussian terms, normalized by a convolution of Gaussian densities, after some matrix manipulations (see [15]), it is possible to rewrite the posterior in the following Gaussian form:

$$p(\mathbf{x}|\{\mathbf{y}_k, \mathbf{s}_k, \theta_k\}, \gamma) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) \tag{11}$$

where

$$\mathbf{\Sigma} = \left(\mathbf{Z}_x^{-1} + \lambda^{-2} \sum_{k=1}^{K} \mathbf{W}_k^T \mathbf{W}_k\right)^{-1} \tag{12}$$

$$\boldsymbol{\mu} = \lambda^{-2} \mathbf{\Sigma} \left(\sum_{k=1}^{K} \mathbf{W}_k^T \mathbf{y}_k\right) \tag{13}$$

This posterior is basically a prior-compensated pseudo inverse matrix. The covariance in Eq. (12) encodes the uncertainty over each HR pixel: this uncertainty is mainly driven by the smallest value between the prior covariance and the covariance of the likelihood term, weighted with noise variance λ^2.

Our approach generalizes this Bayesian framework, modifying the likelihood terms with additional knowledge, gained by a previous clustering step. In this way, it becomes possible to estimate the HR resolution image, taking differently into account each pixel of the LR images.

3. THE PROPOSED APPROACH

In the Bayesian framework of Sec. (2), the role of the Gaussian noise additive term is mainly to capture the *global* discrepancy between the model and the data. This term acts equally and indipendently on all image pixels, thanks to the safe conditions guaranteed by the three strong assumptions. However, we cannot maintain these hypotheses in a more general case, in which the targets to be super-resolved may be more than one (depending on their permanence in the video sequence), freely moving in cluttered scenes, and also having entirely distinct appearances. This variability translates into different confidences for each area of the LR images and hence different 'weights' for the estimation of the HR representation.

To take into account these uncertainties, the observation model of Eq. (1) must be changed. Our approach makes two major modifications by considering several targets to super-resolve in the scene, and exploiting more suitable noise processes. In short, (Step 1) we initially perform a clustering of the video sequence, each cluster containing most shots of a particular target. In the same process, every frame is coarsely aligned, inverting its spatial transformation to maximise the cluster coherence. Subsequently, (Step 2) we derive a different SR generative model for each cluster. This integrates the information brought by the LR frames into a HR image for each target.

In the following subsections, we will see how these steps can be performed in a statistical fashion, accurately weighing the large existing uncertainty. We will show how persistent objects in the video sequences scatter precious information that can and will be distilled with super-resolution.

3.1 Low-resolution frame clustering and registration

In order to detect salient targets in the frames we basically perform clustering among LR images, considering the appearance of persistent targets captured in the sequence, invariantly with respect to camera movements. The model over which the clustering step operates is the one proposed in [7], in which each single frame of a video sequence is considered as generated by a representative ("mean") image, subject to a discrete transformation and sensor noise addition. Performing clustering into this space corresponds to individuate in the M-dimensional pixel space a set of clusters c, $c = 1, \ldots, C$, of LR frames. The c-th cluster is representative of a particular set of LR images, that we denote as $\{\mathbf{y}_k\}^c$. This clustering process is possible by modeling the LR frames of the sequence via a transformed mixture of Gaussians. In formulae, the generative process captured by this model is the following (omitting for clarity the index k, present in each term):

$$\mathbf{y} = \mathbf{T}\boldsymbol{\mu}_c + \mathbf{T}\boldsymbol{\Phi}_c\mathbf{T}^T + \boldsymbol{\Psi}, \tag{14}$$

in which all the possible invertible discretized transformations applied at the image $\mathbf{y}$, namely $\mathbf{T}$, are taken into account, considering the wrap-around effect. Each of the C cluster is modelled with a Gaussian function $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Phi}_c)$. The M-th dimensional $\boldsymbol{\mu}_c$ represents a possible image that produces $\mathbf{y}$ via the probable $\mathbf{T}$ transformation, extracted from all the possible M-dimensional transformation with probability l. This process is carried out with pixel level uncertainty proportional to the diagonal covariance matrix $\boldsymbol{\Phi}_c$, plus the noise captured with a zero mean Gaussian process $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ diagonal matrix, with elements equal to ψ^2.

The covariance term $\boldsymbol{\Phi}_c$ in this framework is highly meaningful, because it models the uncertainty with which the $\{\mathbf{y}_k\}^c$ frames represent the mean image $\boldsymbol{\mu}_c$ For simplicity, this covariance matrix is modelled as a diagonal variance matrix, assuming all the LR pixel as independent each other. This assumption implies that the lower the variance correspondent to the pixel $\boldsymbol{\mu}^{<m>}$, the more uncertain is that pixel value, the lower useful information is brought by the aligned component images $\{\mathbf{y}_k\}^c$.

The likelihood of the image $\mathbf{y}_k$ with respect to this framework is modelled using the density

$$p(\mathbf{y}_k|\mathbf{T}_k, c) \sim \mathcal{N}(\mathbf{y}_k; \mathbf{T}_k\boldsymbol{\mu}_c, \mathbf{T}_k\boldsymbol{\Phi}_c\mathbf{T}_k^T + \boldsymbol{\Psi}) \tag{15}$$

All the necessary parameters of this model are learned in a Maximum Likelihood fashion using an exact Expectation Maximization algorithm (for additional details, about both the model and the learning step, please refer to [2])

Once the model is learned, several useful inferences are possible. For example, given an image $\mathbf{y}_k$, it is possible to choose the most probable generative cluster $\widetilde{c}$:

$$\widetilde{c} = \underset{c}{\operatorname{argmax}} \; p(c|\mathbf{y}_k)$$

This operation effectively groups all the K images in C sets of $\{\mathbf{y}_k\}^c$ frames, all depicting the same target. In the same way it is possible to infer, for each LR image, the most probable transformation $\widetilde{\mathbf{T}}_k$ that is applied to the cluster mean $\boldsymbol{\mu}_{\widetilde{c}}$ that most probably has generated it:

$$\widetilde{\mathbf{T}}_k = \underset{\mathbf{T}_k}{\operatorname{argmax}} \; p(\mathbf{T}_k|\mathbf{y}_k, \widetilde{c})$$

Therefore, a LR registration is available considering $\mathbf{T}_k^{-1}\mathbf{y}_k$ for $\widetilde{\mathbf{y}}_k = \{\mathbf{y}_k\} \in \widetilde{c}$, for each cluster c. So, after the clustering, we register all the LR images belonging to a particular cluster with a LR pixel precision, obtaining, for each cluster, a set of $\{\widetilde{\mathbf{y}}_k\}^c$.

Therefore, the clustering step allows us to recover *nearly* all the necessary constraints of the classical super-resolution framework. Actually we obtain C groups of aligned LR images (in the M-dimensional image space) $\{\widetilde{\mathbf{y}}_k\}^c$, showing each one in the center of the image the c-th persistent target. The mean LR representation of this target is encoded in the mean of each Gaussian cluster, normalized with respect to invertible transformations. The difference with an ordinary super-resolution algorithm is that in the classical formulation all the LR images globally represent the target, while here all the frame consist in a target part, relevant in the process of super-resolution, fused with a clutter part, useless.

In the following section, we will show how to embed the clustering results in the super-resolution framework, enriching the Bayesian structure, and obtaining for each cluster a precise super-resolved image, via an estimation process that considers differently the LR images pixels.

3.2 Distilling Information with Super-Resolution

After the clustering step, we obtain C groups of LR images, each one formed by $K_c = |\{\tilde{\mathbf{y}}_k\}^c|$ LR aligned frames. Each one of these group of LR frames will produce an high resolved version of the target represented, and may be considered independently from the other clusters. Therefore, we consider the cluster c together with the K_c corresponding LR images omitting for the sake of clarity the cluster index c. Disregarding the effect of the normalized transformation $\mathbf{T}_k$, the likelihood term of of Eq. (15) becomes

$$p(\tilde{\mathbf{y}}_k|c) \sim \mathcal{N}(\tilde{\mathbf{y}}_k; \boldsymbol{\mu}, \boldsymbol{\Phi} + \boldsymbol{\Psi}) \tag{16}$$

with respect to the generative process:

$$\tilde{\mathbf{y}}_k = \boldsymbol{\mu} + \boldsymbol{\Phi} + \boldsymbol{\Psi}, \tag{17}$$

The last two factors of the right hand part of Eq. (17) can be condensed in a covariance term $\widetilde{\boldsymbol{\Phi}}$, that we call *contextual covariance*. This covariance term takes into account the sensor noise, and the uncertainty in the generative model, and the per pixel uncertainty over the mean image. This covariance depends of how well a target is represented in the M-dimensional pixel space by the frames belonging to the c-th cluster, from which comes the term "contextual". At the same time, the contextual covariance explains how likely the pixels of each image contains useful information for the mean image estimation process. We claim that this uncertainty holds proportionally in the estimation of a super-resolved version of the mean image. Therefore, we inject this structured uncertainty directly in the generative process of the HR image, obtaining the novel LR image generative model:

$$\tilde{\mathbf{y}}_k = \mathbf{W}_k\mathbf{x} + \widetilde{\boldsymbol{\Phi}} \tag{18}$$

taking into account of the carried out LR sub-pixel transformations $\mathbf{s}_k, \theta_k$ plus the PSF width γ, in the matrix $\mathbf{W}_k$.

The corresponding likelihood term becomes:

$$p(\tilde{\mathbf{y}}_k|\mathbf{x}, \mathbf{s}_k, \theta_k, \gamma) \sim \mathcal{N}(\tilde{\mathbf{y}}_k; \mathbf{W}_k\mathbf{x}, \widetilde{\boldsymbol{\Phi}}) \tag{19}$$

Comparing this likelihood with the old LR image appearance model of Eq. (2), we observe that in this case each pixel of the LR image has a different weight with respect to the estimation of the HR image. Leaving the other terms unchanged as for the formulation of [14], we embed this novel likelihood term in the Bayesian posterior density over the HR image $\mathbf{x}$. Expanding opportunely the term, we can obtain the new Gaussian formulation of the posterior in the form

$$\mathcal{N}(\mathbf{x}; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \tag{20}$$

where

$$\bar{\boldsymbol{\Sigma}} = \left(\mathbf{Z}_x^{-1} + \sum_{k=1}^{K} \mathbf{W}_k^T \widetilde{\boldsymbol{\Phi}}^{-1} \mathbf{W}_k\right)^{-1} \tag{21}$$

$$\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{\Sigma}} \left(\sum_{k=1}^{K} \mathbf{W}_k^T \widetilde{\boldsymbol{\Phi}}^{-1} \mathbf{y}_k\right) \tag{22}$$

The process underlying the reconstruction of the HR image is once again the EM procedure. In this case, the LR images $\{\tilde{\mathbf{y}}_k\}$ are the visible variables, the HR image $\mathbf{x}$ is considered as a hidden variable, forming together the so called complete data. In the standard formulation of the EM framework, the task is to maximize the complete data likelihood with respect to the hidden parameters, in our case $\boldsymbol{\Omega} = \{\mathbf{s}_k, \theta_k, \gamma\}$. The EM algorithm essentially operates in two step: in the E-step it computes the expectation over the hidden variable of the log of the complete data likelihood, given an arbitrary set of values for the hidden parameters $\boldsymbol{\Omega}^{(i)}$, and the visible variables. In the M-step, the EM algorithm maximizes the computed expectation with respect to the hidden parameters $\boldsymbol{\Omega}^{(i+1)}$, obtaining the next set of hidden parameters values. These two steps are iteratively computed, until a convergence criteria is reached.

In our case, the E-step is equal to

$$\mathrm{E}\left[\log p\left(\{\tilde{\mathbf{y}}_k\}, \mathbf{x}|\boldsymbol{\Omega}^{(i+1)}\right) \middle| \{\tilde{\mathbf{y}}_k\}, \boldsymbol{\Omega}^{(i)}\right] \tag{23}$$

that expanded gives

$$\int_{\mathbf{x}} \log p\left(\{\tilde{\mathbf{y}}_k\}, \mathbf{x}\middle|\boldsymbol{\Omega}^{(i+1)}\right) f\left(\mathbf{x}\middle|\{\tilde{\mathbf{y}}_k\}, \boldsymbol{\Omega}^{(i)}\right) \tag{24}$$

Here, the f function is the marginal distribution of the unobserved data, dependent on both the LR images $\{\tilde{\mathbf{y}}_k\}$ and on the current parameters $\boldsymbol{\Omega}^{(i)}$, that in our case represent the posterior over the high resolution image. So, in the E-step we calculate this posterior distribution, that holds in Gaussian form as shown in Eq. (20).

In the M-step, we maximize the expectation of Eq. (23), taking the delta function approximation

$$f\left(\mathbf{x}\middle|\{\tilde{\mathbf{y}}_k\}, \boldsymbol{\Omega}^{(i)}\right) = p(\mathbf{x}|\{\tilde{\mathbf{y}}_k, \mathbf{s}_k, \theta_k\}, \gamma) \tag{25}$$

$$= \mathcal{N}(\mathbf{x}; \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}) \tag{26}$$

$$\sim \delta(\mathbf{x}_{MAP}) \tag{27}$$

$$= \delta(\boldsymbol{\mu}) \tag{28}$$

as performed in [15], because the mean of the posterior Gaussian density corresponds to its Maximum a Posteriori. Therefore, this step calculates the maximum of the log of the complete data likelihood, in order to obtain the new parameters $\boldsymbol{\Omega}^{(i+1)}$, calculated in the maximum value assumed by the hidden variable $\mathbf{x}_{MAP}$ i.e. at $\boldsymbol{\mu}$. The maximization can be obtained using some gradient descent technique, as performed in [14], until a convergence criteria is reached, at iteration i_{max}. At that final iteration, we have reached the locally optima hidden parameters values $\boldsymbol{\Omega}^{(i_{max})}$, that allow us to build the super-resolved version of the target, simply injecting them in the Eq. (20), and taking the mean $\bar{\boldsymbol{\mu}}$.

In this formulation, we choose to embed only the contextual covariance obtained from the clustering step, ignoring the mean image information, that would be integrated in the prior term in some fashion. In this way it is possible to evaluate if a more expressive uncertainty information can drive expressively the SR resolution estimation process, leaving the task of the building of the HR image completely to the LR images. The potential advantages are several, and in the experimental section we justify them with some results. Actually, rather than trying to estimate from a video sequence of frames only one high resolution image, our method permits to estimate several HR representations, built from

a reasonable clustering over the entire set of frames, opportunely pre-aligned at LR level of pixels. Moreover, the estimated HR image takes as informative values the ones present in the LR images associated to a low variance in the corresponding mean image; in this way, the shift, rotation, and PSF parameters are estimated referring over "safe" values, in a statistical sense. Finally, as shown in Eq. (20), the computation of the huge $N \times N$ covariance $\bar{\mathbf{\Sigma}}$ is necessary; in the original formulation of [14], a reduced area is considered, chosen arbitrarily in the image; in our method instead, we choose that area in a completely automatic fashion, which center corresponds to the lowest contextual variance obtained after the clustering step.

4. EXPERIMENTAL SESSION

The proposed approach can be considered from one side as an instance of super-resolution, but starting with different initial hypotheses. Actually, as we saw in Sec. (3.2), after the clustering step the three basic constraints are only nearly satisfied, because the initial aligned images do not show exactly the same scene, rather, they depict a local prominent object immersed in different background scenes. In this sense, our method compensates this uncertainty, introducing a structured covariance term in the generative process of the LR images.

Even if our method outperforms all the image super-resolution algorithms (assumed standard initial conditions), we still try to compare our algorithm with other techniques considering the following situation: an arbitrary LR frame sequence $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\}$ is considered, and the clustering step is applied to get C sets of aligned LR frames. The cluster number C is easily defined heuristically, in accord to [7]. A model selection procedure may be embedded in the process, but this goes beyond the aim of this paper. Therefore, we obtain C instances of the same problem to which our algorithm can be applied, and compared with a standard bi-cubic image interpolation method, and with the Bayesian image super-resolution algorithm proposed in [14]. In this way we measure the effective robustness and efficacy of our method, estimating super-resolved targets from different frames, originally unordered in time, and evaluating the effectiveness of the structured covariance term. In the following experiments, for convenience sake we consider only translations, i.e. rotations are not taken into account.

4.1 Synthetic data

The synthetic example we used to test our method is shown in Fig. 2a: the main target is an image of a lowercase black letter 'r' on a white background of size 32×32 pixels. The first sequence of low resolution images were generated by randomly shifting this image, convolving with a Gaussian PSF with $\gamma = 2$ and finally downsampling to a resolution of 8×8 pixels. The goal of this experiment was then to reproduce as best as possible the high resolution image, comparing the results of our method and the other algorithms. Fig. 2a shows one of the 30 images thus generated. Fig. 2b (left and right column) show the mean and covariance diagonal of the cluster resulting after the application of the clustering step. Note how the least variance is expressed by the leg of the letter and the background.

In a second series of experiments we replaced the white background with a sequence of images coming from a movie of moving terrains. The goal was to mimick the presence of

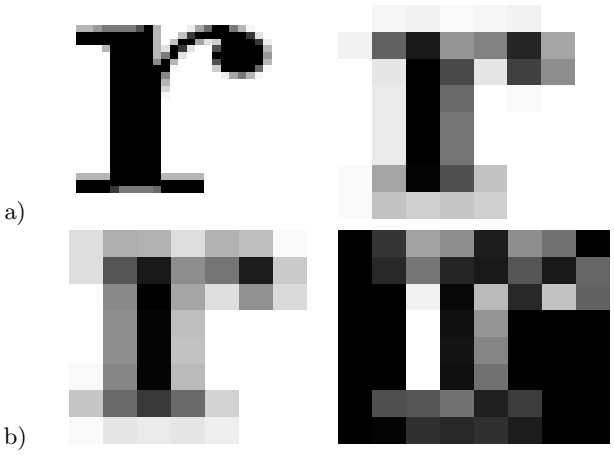

Figure 2: Clustering results over synthetical data : a) example of an original 32×32 frame and (right) subsampled 8×8 image; b) mean image(left) and diagonal structured covariance (right) of the cluster; the darker the pixel value in the covariance, the lowest the corresponding variance.

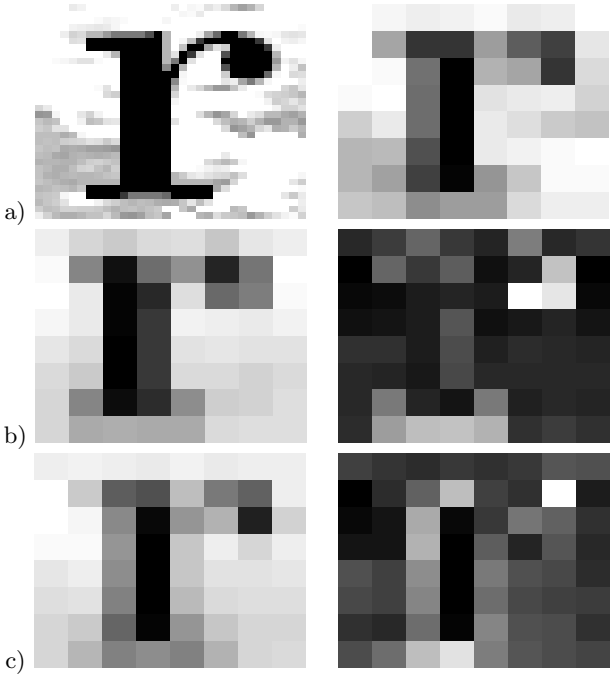

Figure 3: Clustering results over synthetical data with background added: a) example of an original 32×32 frame and (right) subsampled 8×8 image; b) mean image(left) and diagonal structured covariance (right) of the first cluster; c) mean image(left) and diagonal structured covariance (right) of the second cluster; the darker the pixel value in the covariance, the lowest the corresponding variance.

structured background behind our shaking object. Fig. 3a shows one HR frame and its corresponding LR image. Since we wanted to extract meaningful information, the clustering

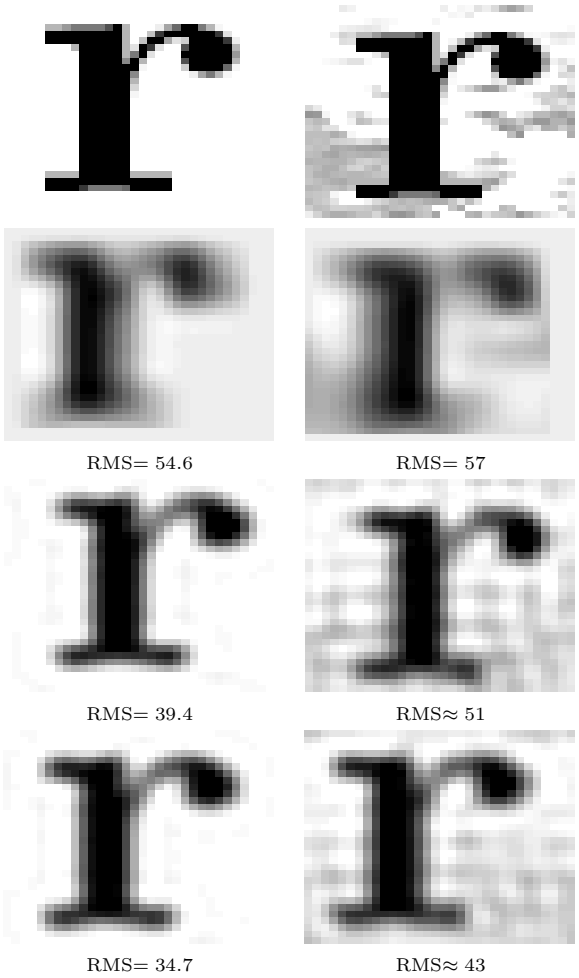

Figure 4: Super-resolved images and RMS values obtained; at the right column the clean sequence and respective results at the left, the sequence with added background and the corresponding obtained results. From the top row: one of the original frame 32×32; single-image interpolations; Bishop's method; our approach.

was trained to recognize two clusters, shown in Fig. 3b-c. Note how the first cluster seems to be uncertain if expressing the foreground or the background, since both the leg of the letter and the background on the right have low variances. On the other side, the second cluster is confident on the letter and cares little of the background. We thus decided to super-resolve the images of this cluster.

Fig. 4 compares the results obtained by interpolating a single LR image, applying Bishop's algorithm, and the present approach. The left column proves that our method performs slightly better than Bishop's provided with perfect data. The results on the noise-corrupted images are more interesting (right column). Even if Bishop's algorithm does a very good job, sometimes it is deceived by the structured noise and provides a less-than-perfect result. The most evident failures are the shrinking of the leg width and the smaller serifs, whereas our approach consistently reproduces all the details, showing that it's not misleaded.

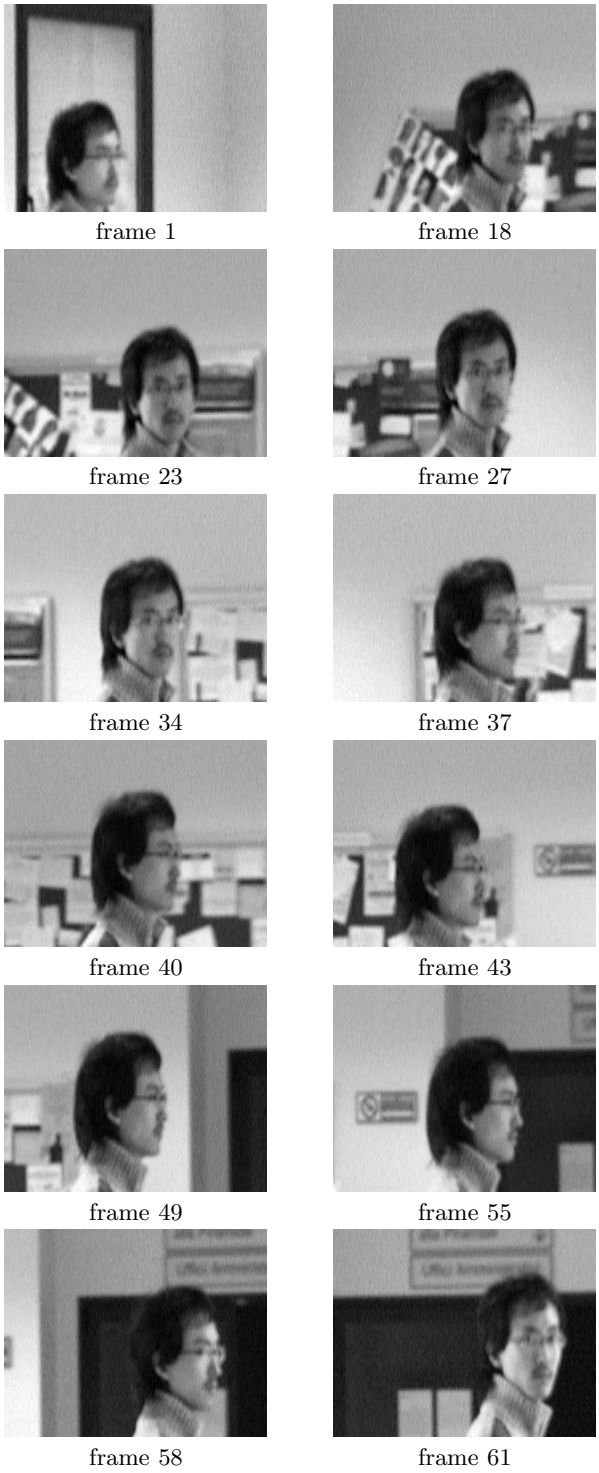

Figure 5: Original high resolution "Indoor sequence".

4.2 Real data

To evaluate the practical effectiveness of our method, we capture several video sequences using a Sony TRV120E handheld camera, at 20 Hz acquisition rate. We report the most significant sequence, named the "indoor sequence". The

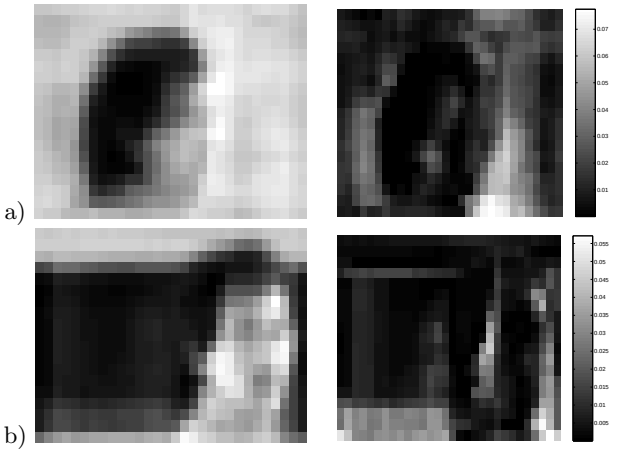

Figure 6: Clustering results for the subsampled indoor sequence: the mean and diagonal covariance images of the $C = 1$ (a) and of the $C = 3$ (b) cluster; in the covariances, the darker the pixel intensity the lowest the value.

video sequence consists in a person walking quickly through an indoor environment, alternating randomly different poses and expressions, as shown in Fig. 5.

Video data are acquired at 320×240 pixels, and the subject captured is about 10 meters away from the camera. Then, 100 frames of the sequence are initially subsampled by a linear magnification factor $q = 4$, with a PSF width $\gamma = 2$. This artificial subsampling provides a sort of ground truth useful in comparison with the obtained SR results. We perform the clustering step with $C = 4$ clusters, obtaining 4 groups of images aligned at LR pixel level, from which we choose the ones with high certainty information, in order to perform the HR image estimation. Therefore, we consider only two clusters for $C = 1$ and $C = 3$, with respectively 16 and 14 LR images, which mean and diagonal covariance images are depicted in Fig. 6.

From that figure, it is possible to evaluate the uncertainty exploited by the two groups of images, decoded by the correspondent diagonal covariance images; moreover, such images are rectangular patches of dimensions 19×30 LR pixels, centered in the lowest covariance pixel, extracted automatically from the full M-dimensional LR images. All the LR aligned frames are considered in the following using these dimensions (Fig. 7).

As explained in Sec. (3), we distill the information contained in the several LR images, estimating only one HR representation per cluster. To this end, we run the EM procedure as explained in Sec. (3.2) until a reasonable convergence rate is reached. Then, we estimate from the same images the classical Bayesian image super-resolution. Finally, we perform bi-cubic interpolation on a random LR image per cluster.

The results are shown in Fig. 8: in correspondence of low cluster covariance, our method actively uses the pixel information contained in the LR images, correctly estimating the shift parameters and the gamma value and estimating correctly the face, in both poses. Conversely, the classical Bayesian framework erroneously takes equally into account all the LR pixel values and produces bad results.

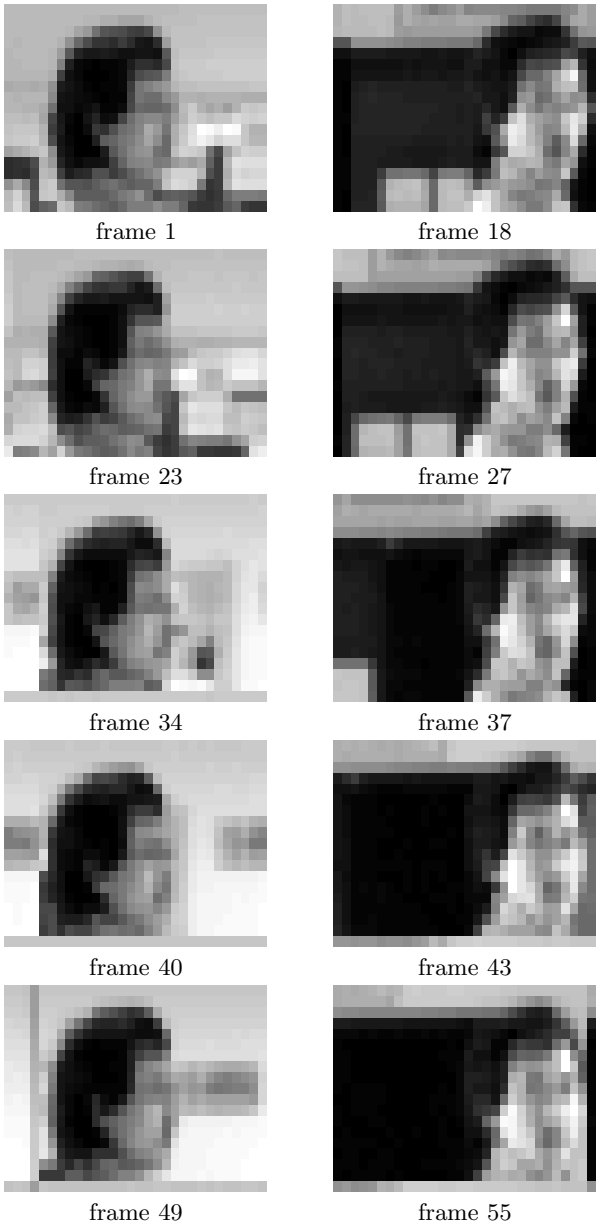

Figure 7: Clustered and aligned LR patches: on the left column, the patches of the LR images belonging to the cluster $C = 1$; on the right the ones relative to the cluster $C = 3$.

Therefore, in order to discover the effectiveness and the robustness of our method, the native images acquired with the camera are considered as LR frames, composing them to build an HR image. In this case, the linear magnification factor applied is $q = 2$. Here, the clustering process considers as prominent one cluster, depicted in Fig. 9.

The obtained results are reported in Fig. 10. In this case, the high certainty produced by the clustering step provides good input data to both the super-resolution algorithms, that estimate similar HR images.

Figure 8: High resolution image estimation results: on the right are the results for the $C = 1$ cluster, on the left for the $C = 3$ cluster; from the top: a low resolved image, bi-cubic interpolation, classical Bayesian super-resolution, our approach.

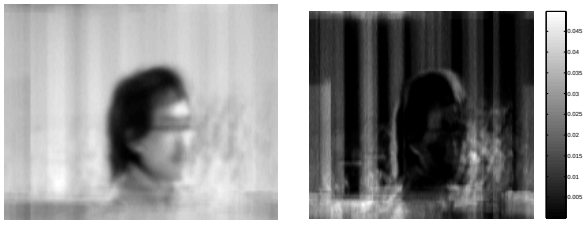

Figure 9: Clustering results for the indoor sequence, taken at the original 320×240 resolution: the mean and diagonal covariance images of the $C = 2$ cluster; the darker the pixel value in the covariance, the lowest the corresponding variance.

5. CONCLUSIONS

In the present paper we propose a novel information-extraction scheme, able to provide informative high resolved patterns from a sequence captured with an hand-held camera. In this sense our approach performs a sort of super-resolution, although all the super-resolution algorithms start from a well constrained hypothesis, here completely missed.

Figure 10: High resolution image estimation results: on the left are the results for the estimated cluster, on the right some enlargements; from the top: a LR image, bi-cubic interpolation, classical Bayesian super-resolution, our approach.

Our framework is formed from a translation invariant video clustering step, in which similar low resolution frames are clustered together. The clustering rule normalizes and groups those frames that exhibit recurrent visual patterns, invariantly respect the transformation. After the grouping, we obtain a measure of the cluster's compactness, encoded in the covariance matrix of the clusters. We actively use this information in the extraction of high resolution patterns, using a Bayesian iterative approach. This approach builds for each cluster an high resolved image, considering differently each pixel of each low resolution frame, depending on the covariance of the cluster: the higher the variance of a low resolution pixel, the lower the contribute of that pixel in the construction phase of the high resolution one.

The results show that the obtained results are good and informative for a detection task, even if not completely photo-realistic: this effect is caused primarily from the clustering step. In facts, the higher the compactness of the cluster, the higher the contribute of each pixel of each low resolution image in the high resolved one. Anyway, the growing diffusion of home-made video data useful for surveillance encourages the development of technics capable to increase the

resolution of the video pattern, not merely using one frame interpolation.

The future perspectives are to embed the process of clustering in the super-resolution step, using a structured generative model that clusters and improves the resolution of the low resolution step in one shot. In this method we can more compactly consider the uncertainty in the video data, producing more highly defined information.

6. REFERENCES

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167 – 1183, September 2002.

[2] J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report ICSI-TR-97-021, ICSI, 1997.

[3] C. Bishop, A. Blake, and B. Marthi. Super-resolution enhancement of video. In C. Bishop and B. Frey, editors, *Proceedings Artificial Intelligence and Statistics*, 2003.

[4] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson. Super-resolved surface reconstruction from multiple images. In G. R. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 293–308. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1996.

[5] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Comput. Graph. Appl.*, 22(2):56–65, 2002.

[6] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *Int. J. Comput. Vision*, 40(1):25–47, 2000.

[7] B. Frey and N. Jojic. Transformation-invariant clustering using the EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):1 – 17, January 2003.

[8] R. Hardie, K. Barnard, and E. Armstrong. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6:1621–1633, 1997.

[9] K. Kim, M. Franz, and B. Scholkopf. Kernel Hebbian algorithm for single-frame super-resolution. In A. Leonardis and H. Bischof, editors, *Statistical Learning in Computer Vision (SLCV 2004)*, pages 135–149, 2004.

[10] O. Kursun and O. Favorov. Single-frame super-resolution by a cortex based mechanism using high level visual features in natural images. In *WACV02*, pages 112–117, 2002.

[11] PAMI. Special issue on video surveillance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[12] R. Schultz and R. Stevenson. A Bayesian approach to image expansion for improved definition. *IEEE Transactions on Image Processing*, 3(3):233–242, 1994.

[13] E. Shechtman, Y. Caspi, and M. Irani. Increasing space-time resolution in video. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 753–768. Springer-Verlag, 2002.

[14] M. Tipping and C. Bishop. Bayesian image super-resolution. In *Neural Information Processing Systems - NIPS'2002*, Vancouver, 2002.

[15] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.