Exploiting Space-Time Statistics of Videos for "Hallucination"

Göksel Dedeoğlu

Thesis Proposal
February 2004

Committee Members
Takeo Kanade (Chair)
Simon Baker
Henry W. Schneiderman
Jonas August
William T. Freeman (External)

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

In this work, we address the task of enhancing the spatial resolution of video sequences, known as *super-resolution*. Specifically, we consider the problem of super-resolving a human face video by a very high ($\times 16$) zoom factor. Inspired by recent literature on hallucination and example-based learning, we formulate this task using a graphical model that encodes 1) spatio-temporal consistencies, and 2) image formation & degradation processes. A video database of facial expressions is used to learn a domain-specific prior for high-resolution videos. The problem is now one of probabilistic inference, in which we aim to find the high resolution video that best satisfies the constraints expressed through the graphical model. Traditional approaches to this problem using video data first estimate the relative motion between frames and then compensate for it, resulting effectively in multiple measurements of the scene. Our use of time is rather direct: We define data structures that span multiple consecutive frames, enriching our feature vectors with a temporal signature. We then exploit these signatures to find consistent solutions over time.

We present preliminary results in which a 8×6 pixel-wide face video, subject to translational jitter and additive noise, gets magnified to a 128×96 pixel video. Preliminary results show that, by exploiting both space and time, drastic improvements can be achieved in reducing both video flicker artifacts and mean-squared-error.

Proposed extensions include data-driven modifications of the graphical model and improvements on feature definitions as well as compatibility function designs. Finally, the task of human silhouette enhancement in videos is proposed to test our framework as a general, model-based regularization tool.

Contents

Figure 1: Given only a low-resolution video (top), can one estimate (or "hallucinate") the original high-resolution video (bottom)? Conventional methods, such as bicubic interpolation, are insuffi cient (middle). In this work we explore zooming using a database of videos with an inference procedure that enforces spatio-temporal consistency of the resulting hallucinated video. Our results more closely approximate the high-resolution video.

1 Introduction

Imagine we are given an extremely low resolution video (Fig. 1, top). *Assuming* that there is a human face in these images, can we guess the missing details, and estimate (or "hallucinate") a highly zoomed video that resembles the original (bottom)? In this work, we present a framework for this task, formulate it as an inference problem, and describe an algorithm for solving it.

The problem of estimating high-resolution image or video details has been known as *super-resolution* (SR). With regard to capturing fine details of a given scene, the most critical shortcomings of current imaging devices are the optical blur and the limited density of their sensing elements, which result in severely aliased samples. This situation is only aggravated when a camera is located far from the objects of interest, as is the case in most surveillance type of applications. Such physical limitations have motivated signal processing and pattern recognition approaches to SR, yielding a number of algorithms in the past two decades.

1.1 Zooming: An Inverse Problem

In the video zooming (or super-resolution) problem, one is given a low-resolution video, and asked to produce a high-resolution one. To understand the relationship between these two videos, let us first consider the simpler, static image case. Let us denote the high-resolution image by H ($MN \times 1$ vector), and the low-resolution image by L ($N \times 1$ vector), corresponding to a downscaling factor of $\sqrt{M}$ per dimension. From

the point of view of image formation, the relationship between H and L is relatively well understood [1]. A local linear averaging operator, denoted by A ($N \times MN$ matrix), maps high-resolution images onto the space of low-resolution ones. Furthermore, a pixel-wise independent Gaussian noise may be added to account for the imaging sensor noise. Hence, the observation model becomes

$$L = AH + \eta_L. \tag{1}$$

We observe that the problem of recovering the high-resolution image H amounts to inverting the operator A. However, because this is a many-to-one mapping, its inversion is mathematically ill-posed [2], necessitating some form of prior knowledge about images [3] to constrain the solution for H. For example, the smoothness assumption (Fig. 1, middle) would penalize strong edges, effectively constraining the solution space to that of smooth images.

The observation model in (1) generalizes to video sequences by simply redefining the variables H and L to be stacked vectors of video frames, and turning A into a block diagonal matrix. The ill-posed nature of the problem still persisting, a prior for high-resolution videos would still be needed: A commonly used prior assumption for videos is that of spatio-temporal smoothness.

1.2 Our approach

This proposal addresses the ill-posed nature of the zooming problem by adopting strong image and video priors. These are learned from training examples, and imposed selectively during zooming, as a function of observed data.

In order to keep the complexity of the overall system relatively low, we assume that we will be super-resolving human faces only: This allows us to focus our attention (and computational resources) from the larger space of all possible high-resolution videos to a much smaller subset thereof. To characterize this reduced space, we use a database of video examples taken from this domain. In order to address the variability of face features, we choose a patch-based representation, and allow recombinations of patches from different videos and time instants while synthesizing new ones. Global facial phenomena (over space and time) are further imposed through pair-wise couplings between these patches.

In our scheme, zooming is viewed as a prediction task: For this end, we downsample high-resolution videos to generate their low-resolution counterparts, and store these pairs for later referral. When presented with a low-resolution input, we essentially look-up the most similar low-resolution part of all low-high training pairs, and select its high-resolution part as the predicted prior. We use a graphical model to concisely express the relationship between low-resolution observations, predicted priors, and desired high-resolution outputs. The zooming problem is posed as one of probabilistic inference, in which we aim to find the high-resolution video that best satisfies the constraints expressed through the graphical model.

1.3 Outline

This thesis proposal is organized as follows: In section 2, we briefly review existing approaches to enhance the spatial resolution of images and videos, and highlight their underlying assumptions as well as limitations. In the light of this review, section 3 presents the particulars of our framework and its novel aspects. In section 4, we report preliminary experimental results and observations. Section 5 enlists validation experiments, extensions, as well as theoretical study directions as proposed work before the completion of this thesis. Finally, expected contributions are summarized, and a roadmap is proposed in sections 6 and 7, respectively.

2 Background

Super-Resolution (SR) aims to estimate high-resolution image or video details that the imaging sensor failed to capture due to optical blur and insufficient sampling rate [4, 5, 6]. In review of the existing literature on this problem, two classes of approaches can be identified: First, *reconstruction-based* methods aim to increase the effective sampling density. This type of resolution enhancement requires multiple low-resolution samples of the underlying scene to be fused into a coherent high-resolution estimate. The estimate, in turn, should be able to account for all undersampled low-resolution observations by simulating the image degradation process. The second and more recent class of approaches are *learning-based*, where low-resolution observations are used to predict lost high-resolution details using a training set. Their philosophical difference aside, both classes typically employ additional priors to encourage generic image properties such as local smoothness, edge preservation, and positivity.

2.1 Reconstruction-based Approaches

The fundamental problem with low-resolution images and videos is that they are severely undersampled, *i.e.*, aliased. Since the early 80s, much research in SR has been dedicated into alleviating this aspect by means of increasing the effective *sampling density*. The starting point is a set of sub-pixel shifted low-resolution observations of a scene (Fig. 2, left). Provided that one can accurately bring these images into a common coordinate frame, it becomes possible to estimate a "super-resolved" image, defined over a sampling grid finer than in any of the original observations (Fig. 2, right). The theoretical foundation of these approaches can be found in the Generalized Sampling Theory developed by Papoulis [7].

The intuition above immediately leads to the so-called *reconstruction constraint* that the high-resolution estimate would need to satisfy: After being warped to each low-resolution image's coordinate frame, followed by downsampling and decimation, the estimate should be able to regenerate all low-resolution observations up to a noise level. In the case of a video sequence, temporally neighboring frames form a natural set of candidates for providing such additional constraints. Provided that a video sequence is smooth enough for reliably and correctly recovering the relative motion between frames, such reconstruction constraints can be imposed.

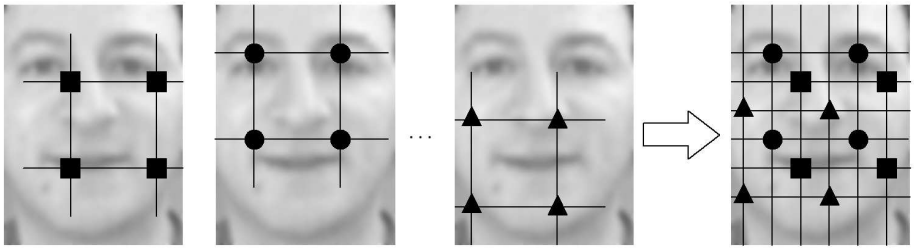

Figure 2: Reconstruction-based approaches aim to increase the spatial density of samples. Provided that multiple low-resolution images (left, 2×2 pixels) can be aligned correctly into a common coordinate frame, interpolation over a fi ner sampling grid can yield a "super-resolved" image (right, 4×4 pixels).

Reconstruction-based line of research dates back to Tsai and Huang's [8] frequency-domain technique. Their work did not model optical blur or observation noise, and restricted the relative motion between low resolution images to global translations. Since then a variety of models and computational tools have been successfully applied to SR, gradually relaxing the assumptions about allowed motion types while incorporating more realistic observation noise models (comprehensive reviews can be found in [5] and [6]). In the following, we briefly touch upon established computational tools that follow the reconstruction-based paradigm.

Least-squares (LS) methods were used in both frequency [9] and spatial domains [10, 11]. Using spatio-temporal regularization, [12] sought LS solutions to enhance both spatial and temporal resolution of videos. In [13], non-uniform interpolation was followed by deblurring for enhancing the spatial resolution of images. Inspired by Computed Tomography algorithms, [14] developed the Iterated Back Projection (IBP) algorithm. Projection-Onto-Convex-Sets (POCS) was first used in [15] as a means of incorporating image priors. Later, [16] took into account optical blur, nonzero aperture time, and sampling on arbitrary lattices using the POCS formulation.

Probabilistic inference tools have also been employed to solve SR problems. Spatial-domain Bayesian formulations such as [17, 18, 19] provided a rigorous estimation framework, especially with regard to spatial priors. Using probabilistic graphical models such as Markov Random Fields (MRF) [20], one can conveniently express desired local image properties such as smoothness and edge-preservation. Extending MRFs into the temporal domain, [21] used spatio-temporal smoothness priors for pixels which were successfully tracked between adjacent frames in time. Assuming an additive Gaussian noise model, Bayesian inference leads to well-defined Maximum A Posteriori (MAP) estimates of high-resolution images that can be computed efficently. This is in contrast to POCS and IBP methods that may not have unique solutions. Variations on the Bayesian approach include Gaussian process priors [22] and dynamic tree inference algorithms [23].

Accurate image alignment plays a critical role in all reconstruction-based methods. Issues related to motion estimate ambiguities and partial occlusions were addressed explicitly in [24]. In a similar vein, [25] defined and used what they called validity and

segmentations maps for eliminating inaccurate motion estimates and enabling object-based tracking. Confirming the importance of the quality of motion estimates, [26] cast the registration and SR problems jointly, and solved them iteratively. In [27], a set of feasible motions were maintained to avoid early commitment to potentially erroneous initial estimates. More recently, [28] provided an analysis of the influence of image alignment and warping errors on the quality of super-resolved images.

The type of allowed motions was relaxed in [29] and [30], which computed dense optical flow fields between multiple views. A robust optical flow method was developed in [28], emphasizing the consistency of recovered flow fields to ensure higher registration accuracy. The difficulty in estimating such high-dimensional motions models is that each flow vector has to be estimated from a relatively small number of pixels. This is in contrast to more global, parametric motion models [31] which remain more reliable as long as the underlying scene motion can be well approximated. For example, inter-frame motions were modeled as homographies in [32], which estimated their parameters using feature-based random sampling procedures, known to be robust against outliers.

What makes reconstruction-based SR possible is the registration of independently acquired samples of a signal into a common coordinate frame. When this crucial step is not well-defined, or simply too difficult due to noisy data, SR cannot be performed.

2.2 Learning-based Approaches

All SR methods presented in the previous section are based on the principle of increasing the effective sampling density. In contrast, learning-based approaches rely on the premise that a low resolution observation alone may contain enough information to make reasonable predictions about its high-resolution counterpart, or features thereof (such as edges). The essence of these techniques is to use a training set of high-resolution images and their artificially downsampled, low-resolution versions to learn a joint occurence model. This model can take a variety of forms, including a set of learnt interpolation kernels, a look-up table of pairs of low-high resolution image patches, or their coefficients in alternative representations. At the time of applying the learnt model, the task is to predict high-resolution data from the observed low-resolution data.

Learning-based algorithms for SR are relatively recent, and have mostly been restricted to static images. The approach of [33] hypothesized that similar image neighborhoods remained similar across scales, and proposed to learn this structure locally from training samples. A set of interpolation kernels was extracted in an unsupervised fashion, and resolution enhancements by a factor of 2 per dimension were reported. In [34, 35], an example-based learning scheme was applied to generic images, and zooming results up to a factor of 4 were reported. A direct application of this method to video sequences was attempted in [36], but severe video artifacts were observed. To achieve more coherent videos, the heuristic of *re-using* high-resolution solutions of preceeding frames was proposed [1].

An interesting aspect of learning approaches is that they can be very powerful

[1] To the best of our knowledge, [36] is the only learning-based SR method that was applied to video.

when images are limited to a particular domain: For instance, [37] considered super-resolving human faces only, and furthermore, employed inhomogeneous (*i.e.*, location-specific) priors. Their recognition algorithm referred to a database of registered face images, and collected best matching image patches given the input, enabling convincing results with zoom factors up to 8. More recently, [38] addressed the same application domain using a two-step procedure, by first estimating a global face via Principal Component Analysis (PCA), and then fitting nonparametric local models. Another example in the face domain is [39], which constrained high-resolution solutions to lie in patch-wise "face subspaces" found by PCA.

In principle, the performance of learning-based techniques is limited by the amount of discriminative information that "sneaks" from high-resolution training samples to their low-resolution counterparts during the downsampling process. The challenge for these algorithms is to retain as much of this information as possible while being able to generalize to other samples drawn from the domain of interest.

2.3 Main Issues

In the light of the literature review presented in previous sections, we now take a closer look at the problem of zooming human face videos, and test whether the assumptions underlying SR algorithms still hold.

As a starting point, we will consider the image registration task. It is a fundamental tool for all reconstruction-based methods. A quick examination of high-resolution video frames such as in Fig. 1 immediately reveals that facial expressions cause a variety of face motions and appearance changes. For instance, a smile or a surprised look causes eyebrows to rise and slightly bend. Cheeks and the lower contour of the chin move and deform as one speaks. Dimples, wrinkles and laugh lines appear and disappear as a result of facial muscle activity. Among more drastic appearance changes are the eyeblinks, occurring within a fraction of a second: Eyelids quickly occlude and then reveal the eyes. Similarly, lips change shape throughout speech, and parts of the speaker's teeth and tongue become intermittently visible for short periods. Unfortunately, from an image registration point of view, such visual phenomena represent the very cases for which the motion estimation problem is not well-defined. Turning our attention to the low-resolution version of the same face video, we notice that the downsampling process has largely destroyed the complex visual phenomena described above. One may wonder whether nonparametric motion models such as optical flow would be able to recover smooth deformations in certain areas of the face, and thus make SR feasible, at least locally. Unfortunately, since the effects of occluded or newly appeared visual structures get irreversibly mixed in with their neighboring pixels, we realize that a good portion of face pixels is unavoidably contaminated with such unmodeled variations. Based on these considerations, it is clear that recovering the *correct* sub-pixel displacements on such small scales is not practical.

The correctness argument about registering video frames may seem too conservative. After all, there is a large body of literature and working systems that estimate complex motion fields for classification and detection purposes [40, 41, 42]. The crucial factor that distinguishes the reconstruction-based SR problem is the *accuracy* requirement on the motion estimation, whereas in other domains, *repeatibility* takes priority.

Unless the low-resolution images are registered accurately, SR estimates will be simply incorrect. However, in any motion-based classification task, as long as the same estimate errors are consistently reproduced, the performance of classifiers or detectors will not suffer.

The discussion above on correctness raises the question of whether frame-to-frame registration is the only possible mode of use for video data. As degraded as they may be, low-resolution videos (Fig. 1, top), still carry *some* information about the scene. In this proposal, we develop a framework that exploits low-resolution video sequences and regularities found therein from a recognition perspective [2]. We follow the learning-based SR paradigm, and base the low-to-high resolution prediction task upon *spatio-temporal signatures* that low-resolution video pixels exhibit.

The model we propose for super-resolving videos is inspired by the following key aspects of earlier work: By limiting our learning task to faces only, and using a spatially varying prior as in [37], we keep the computational requirements relatively low. Inspired by the use of spatial couplings in [34], we model both spatial *and* temporal consistencies in the super-resolved videos. In contrast to [36], we do not resort to re-seeding our high resolution hypothesis space with earlier solutions, but instead model and deal with *temporal visual phenomena* directly. As such, complex re-appearance and occlusion events, which haunt alignment algorithms, can actually be taken advantage of as rich temporal signatures that distinguish facial expressions from each other.

3 Framework

In this section, we develop a probabilistic inference framework for estimating high-resolution video details from their noisy low-resolution versions. Following [37], we use the term *video hallucination* for stressing the recognition-based nature of our approach.

3.1 Graphical Model for Spatio-Temporal Constraints

In modeling the high-zoom problem, we aim to integrate our domain knowledge about the videos of interest with the physical principles of image formation. For this end, we first introduce our graphical model for the formation of low-resolution observations. For clarity, section 3.1.1 first describes this generative model for the static image case, and section 3.1.2 extends the latter to the temporal dimension.

3.1.1 Generative Image Model

A graphical model is a concise tool for expressing causal and statistical dependence relationships between random variables of interest. Specifically, two nodes which are not connected by a link are independent when conditioned upon their neighbors. Such conditional independencies will play a crucial role in Section 3.2, where we will articulate our theory of high-zoom inference.

[2] This work has been submitted to the IEEE 2004 Conference on Computer Vision and Pattern Recognition, and is currently under review.

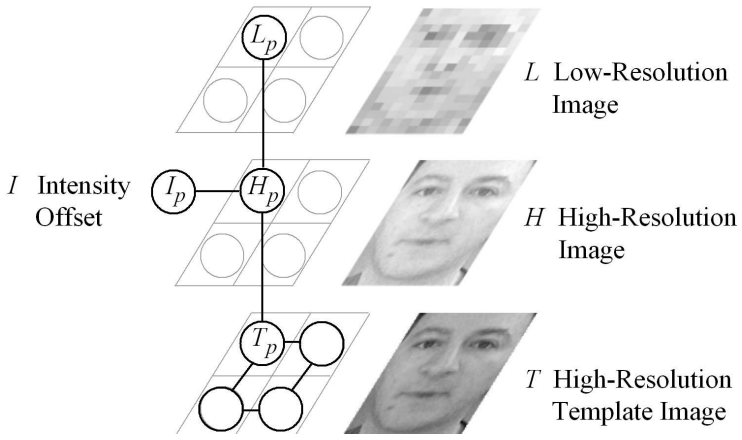

Figure 3: Model of blur and degradation (See section 3.1.1)

Our model for low-resolution observations comprises three steps: organized upwards in Fig. 3, 1) Generation of template image T, 2) addition of illumination offset I to generate a noisy high-resolution image H, and 3) down-sampling and corruption for forming the low-resolution image L. We now discuss each of these steps in detail.

The starting point is a high-resolution template image T, generated following a prior model about possible images in the domain. Building a generative statistical model of T that can account for all possible face images represents a formidable challenge. In order to circumvent this modeling problem, we will take a non-parametric approach, and draw samples from a large database of examples. Since capturing all possible variations of facial expressions and features requires a very large number of examples to be stored, one can adopt local models, defined over image patches, and treat them independently, as in [37]. Such a choice, however, fails to capture those events which span multiple patches, resulting in unrealistic face compositions. As a computational trade-off between treating these patches all independently and building a full statistical co-occurence model, we will impose compatibility constraints only between neighboring patches. In particular, we will use a Markov Random Field (MRF) (Fig. 4, left) to model spatial interactions, allowing us to compose face template images without noticeable artifacts.

After the template image T is formed, we consider a deviation from the illumination conditions in which the prior model was built: An intensity offset I is added to T for producing the high resolution image H. Finally, we model the severe blur and downsampling operations for obtaining the low-resolution observation L by a linear, local-averaging operator followed by additive noise [1].

3.1.2 Exploiting Time

Just as neighboring pixels in natural images tend to be highly correlated, so too are consecutive frames in video sequences. In our work, we exploit these temporal dependencies in further constraining the space of high resolution solutions. By extending the

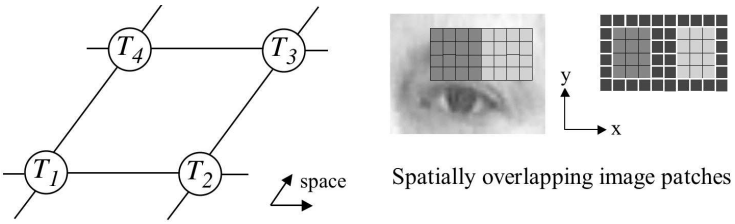

Figure 4: Spatial coupling between neighboring template patches is shown in the Markov Random Field graphs for image super-resolution. The black pixels (right) indicate the locations where neighboring patches must have similar intensity to be considered compatible.

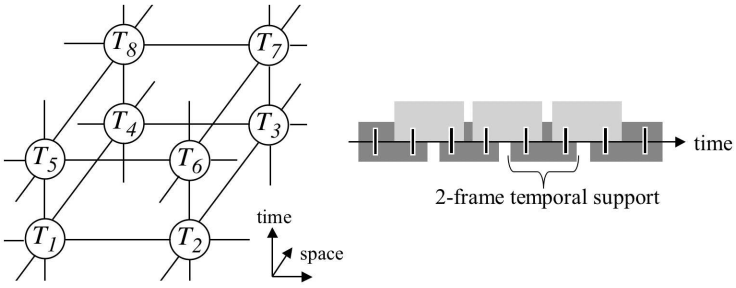

Figure 5: Spatio-temporal coupling between neighboring template patches is shown in the Markov Random Field graphs for video superresolution. Pixel differences over whole patches of overlapping video frames are evaluated for compatibility.

MRF framework into the time dimension (Fig. 5, left), we model couplings between consecutive frames. This results in a three-dimensional network of *video patches*, defined as data structures spanning multiple consecutive frames: For instance, as shown in Fig. 5 (right), we can choose a temporal support of 2 frames for the nodes in T, and make consecutive nodes overlap by one frame. This is equivalent to stating that the underlying video sequence is first-order Markov in time.

Our scheme gives the temporal dimension an unconventional role compared to earlier approaches to super-resolution: In the literature, the relative motion between frames is estimated, then eliminated through warping or optical flow. These approches are essentially two dimensional, treating time, in effect, as a nuisance parameter to be compensated for.

By contrast, we take advantage of the richer local signature that the combination of space *and* time provides. In fact, the very small size of inputs (8×6 pixels) considered in this work would make the recovery of facial motions (e.g., opening and closing of the eyelids and mouth, and the appearance of pupils and teeth) particularly difficult. Avoiding this motion estimation problem, our representation deals with complicated visual phenomena such as occlusions, appearance of new structures, and non-diffeomorphic deformations naturally, in terms of interacting chunks of high-resolution video that constitute the nodes in T.

3.2 Inferring the High-Zoom Video

In this section, we formulate the problem of super-resolving videos by combining the conditional independencies in our graphical model with a basic observation that we call the *unique template* assumption.

Using our graphical model, we pose the problem of super-resolution as one of finding the Maximum A Posteriori (MAP) high-resolution image H_{MAP} and the illumination offset I_{MAP} given the low-resolution image L:

$$(H_{MAP},\ I_{MAP}) \triangleq \arg\max_{H,I}\ logP(H,\ I \mid L).$$

To express the MAP estimate in terms of known quantities, we first marginalize over the unknown template image T:

$$P(H,\ I \mid L) = \sum_{T} P(H, I, T \mid L).$$

By applying the chain rule[3] twice, the posterior becomes

$$\begin{aligned} & \sum_{T} \Big\{ P(H \mid I, T, L)\ P(I, T \mid L) \Big\} \\ = & \sum_{T} \Big\{ P(H \mid I, T, L)\ P(I \mid T, L)\ P(T \mid L) \Big\}. \end{aligned}$$

Using Bayes rule in the first term, the posterior becomes

$$\sum_{T} \Big\{ \frac{P(L \mid H, I, T)\ P(H \mid I, T)}{P(L \mid I, T)}\ P(I \mid T, L)\ P(T \mid L) \Big\}.$$

Observing the conditional independence[4] $P(L \mid H, I, T) = P(L \mid H)$ entailed by our graphical model, and capturing the denominator by a constant C, we rewrite the posterior as

$$C \sum_{T} \Big\{ P(L \mid H)\ P(H \mid I, T)\ P(I \mid T, L)\ P(T \mid L) \Big\}. \tag{2}$$

At this point, we would like to tease out a premise that underlies the entire enterprise of super-resolution. The very assumption that we can perfectly succeed at the task of super-resolution (i.e., obtain H uniquely and to arbitrary resolution) implies that the underlying distribution $P(T \mid L)$ is peaked around the true high-resolution solution. As an approximation, we assume that this posterior is a delta-function at the true configuration, which we estimate using the input.

Unique Template Assumption. Assume that the probability $P(T \mid L)$ over all possible configurations of T is highly concentrated around $T^* = T^*(L)$, i.e.,

$$P(T \mid L) \approx \delta(T - T^*). \tag{3}$$

[3]The chain rule asserts that $P(X, Y) = P(X \mid Y)P(Y)$.

[4]Two nodes which are not connected by a link are independent when conditioned upon their neighbors.

Deferring the computation of T^* until section 3.2.4, we substitute (3) into (2) so that $P(H, I \mid L)$ is approximately

$$C\ P(L \mid H)P(H \mid I, T^*)P(I \mid T^*, L). \tag{4}$$

Using (4), H_{MAP} and I_{MAP} approximately maximize

$$\log P(L \mid H) + \log P(H \mid I, T^*) + \log P(I \mid T^*, L). \tag{5}$$

The individual terms of this objective function have natural interpretations: The first states that H_{MAP} should increase likelihood of the *reconstructed* observation L. The second encourages those H_{MAP} that differ from T^* up to an intensity offset, effectively imposing only a *gradient match* to the template T^*. Finally, the last enforces the *illumination* I_{MAP} to be consistent with the assumed template T^* and observation L.

3.2.1 Finding the Peak Template T^*

Now we describe our method for computing the peak template T^* in (3) by estimating the maximum of $P(T \mid L)$. Using Bayes rule, we first rewrite this posterior in terms of likelihood and prior terms. Observing that nodes in L are conditionally independent given the high-resolution template T, we obtain a factorized likelihood term

$$P(T \mid L) \propto P(L \mid T)\ P(T) = \prod_{p=1}^{N} P(L_p \mid T_p)\ P(T). \tag{6}$$

Unfortunately, in the case of extremely blurred images, the likelihood term $P(L_p \mid T_p)$ is too weak; that is, many templates match with a given L_p. One remedy to this problem is based on the observation that there are spatial dependencies in the observed data. Thus, by pooling contextual information about L_p into a local feature vector, one can make the likelihood term more descriptive. The downside of such an extension is that the factorized form of (6) will no longer be valid. In section 3.2.3, we will present the details of such a feature vector, and expose our assumptions for achieving a computationally tractable algorithm.

3.2.2 The Template Prior

We restrict the space of possible T's to a domain-specific collection of example templates. For this end, a database is generated from training data by artificially downsampling high-resolution images and computing their low-resolution feature images. As shown in Fig. 6, we store these examples patch-wise, in that each record is a quadruple, (t_k, n_k, f_k, s_k), containing high-resolution template patch pixels t_k, a thin strip of surrounding pixels n_k, the feature vector f_k computed at the corresponding low-resolution pixel, and the location of the template s_k.

The MRF model assigns a probability to each template patch configuration T, and according to Hammersley-Clifford theorem, $P(T)$ is a product $\prod_{T_p,T_q} \phi(T_p, T_q)$ of compatibility functions $\phi(T_p, T_q)$ over all pairs of neighboring nodes. We define ϕ

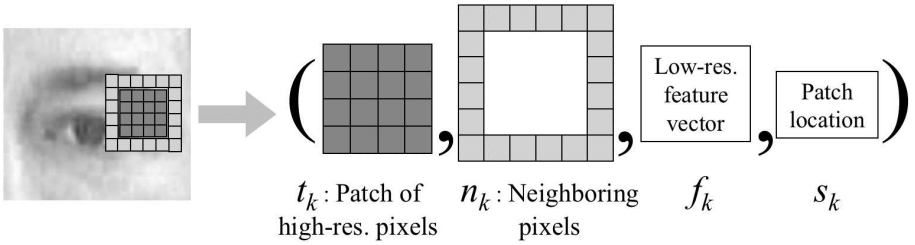

Figure 6: Each database entry contains an image patch, the neighboring pixels (for enforcing consistency), a feature vector (for matching to the low-resolution image), and its location (for supporting non-homogeneous spatial statistics). This structure is repeated for all frames within the temporal support considered.

using similarity similarity between pixel values in the overlapping areas of example patches.

$$\phi(T_p = t_k, T_q = t_l) \propto exp\Big(- \sum_{overlap} \big(t_k(u) - n_l(v)\big)^2 - \sum_{overlap} \big(n_k(u) - t_l(v)\big)^2 \Big).$$

3.2.3 The Feature Vector

To render the likelihood term more descriptive, we use a multi-scale feature vector derived from the low resolution observation L. Following [37], we adopted the *parent vector* [43] as our feature F_p, which stacks together local intensity, gradient and Laplacian image values at multiple scales. Fig. 7 (left) shows a 1-dimensional version of Fig. 3, with the feature vector nodes added.

Factorization Assumption. Observe that we have two random fields, F and T, that are coupled through the image degradation model. For computational tractability, we invoke the pseudo-likelihood approximation [20] to assume that $P(F \mid T)$ factorizes across feature image pixels:

$$P(F \mid T) \approx \prod_{p=1}^{N} P(F_p \mid T). \tag{7}$$

Correspondingly, the graphical model of Fig. 7 (left) is simplified to Fig. 7 (right).

The likelihood $P(F_p = f_p \mid T_p = t_k)$ will be defined using the similarity between the feature vectors f_p and f_k, where k is an index to database entries. For a spatially-varying (i.e., inhomogeneous) prior for T_p, we consider a similarity of the form

$$P(F_p = f_p \mid T_p = t_k) \propto \begin{cases} exp\big(- \|f_p - f_k\|^2 \big) & \text{if } s_k = p, \\ 0 & \text{otherwise.} \end{cases}$$

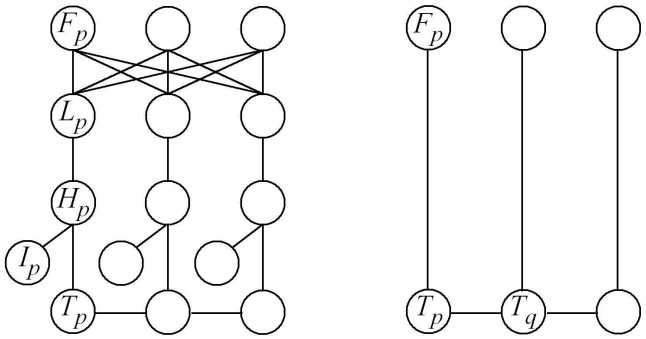

Figure 7: Interactions involved in determining optimal template T^*. For illustration purposes, a 1-dimensional version of the model in Fig. 3 is shown on left. After applying the factorization assumption, the resulting graph structure (right) is tractable enough to apply inference methods such as ICM.

Using the factorized form (7), T^* is approximately

$$\arg\max_{T} \prod_{p=1}^{N} P(F_p \mid T_p) \prod_{(p,q)} \phi(T_p, T_q). \tag{8}$$

3.2.4 ICM Algorithm for T^*

Maximizing the joint probability of T in (8) to obtain T^* is a nontrivial task. We adopt a greedy approach commonly taken in the field of Bayesian image estimation: The Iterated Conditional Modes (ICM) algorithm [44] takes advantage of the Markov structure, and maximizes local conditional probabilities sequentially.

input : observed feature vectors F
output : template image T^*

/ initialize T^* with local Maximum Likelihood estimates */*
for *all video patches p* **do**

$$T_p^* \leftarrow \arg\max_{t_k} P(F_p = f_p \mid T_p = t_k)$$

end

/ choose a video patch, and update it using its neighbors */*
repeat
pick a random location p

$$T_p^* \leftarrow \arg\max_{t_k} P(F_p | T_p^* = t_k) \prod_{q \in N(p)} \phi(T_p^* = t_k, T_q^*)$$

until T^* *converges*;

Algorithm 1: Finding T^* with ICM

3.3 Hallucinating the High-Zoom Video

Now we introduce the details of the likelihood models of image formation and observation. We show that, once we have computed T^*, video hallucination (computing H_{MAP} and I_{MAP}) only requires a quadratic minimization. We then note that a particular case of our framework provides a probabilistic interpretation for the objective function used in a previous approach for static images [37].

3.3.1 Likelihood Models

After the high-resolution template image T^* is composed, an intensity offset field I is applied, producing the high-resolution image H, where

$$H = T^* + I + \eta_H.$$

To express the uncertainties due to both template and illumination models, we include pixel-wise independent additive Gaussian noise $\eta_H \sim N(0,\ diag(\sigma_H))$:

$$P(H \mid T, I) = \prod_{h=1}^{MN} \frac{1}{\sigma_H \sqrt{2\pi}} exp\Big(- \frac{\big(H(h) - T(h) - I(h)\big)^2}{2\ {\sigma_H}^2} \Big).$$

After the high-resolution image H is blurred and downsampled, sensor noise is added, resulting in our model for the low-resolution observation L:

$$L = AH + \eta_L,$$

where matrix A is a local averaging operator with N rows and MN columns. We assume a pixel-wise independent noise model for L:

$$P(L \mid H) = \prod_{l=1}^{N} \frac{1}{\sigma_L \sqrt{2\pi}}\ exp\Big(- \frac{\big(L(l) - (AH)(l)\big)^2}{2\ {\sigma_L}^2} \Big).$$

3.3.2 Computing H_{MAP} and I_{MAP}

Assuming that I and the kernel for blur operator A do not vary within each patch, one can show that $-\log P(I \mid T^*, L)$ is a quadratic form. Combined with the likelihood models above, we can evaluate (5). Thus, H_{MAP} and I_{MAP} minimize

$$\|L - AH\|^2 + \frac{\sigma_L^2}{\sigma_H^2}\|T^* + I - H\|^2 + \|L - AT^* - I\|^2. \tag{9}$$

Individual terms above have intuitive interpretations: From left to right, first, we require the high-resolution image H to be able to *reconstruct* the observation L as closely as possible. Second, we would like H to *match* T^* up to an illumination shift I. Third, I may not take arbitrary values, and should be a *consistent illumination offset* with respect to T^* and L. Finally, we observe that (9) is quadratic in the unknowns H and I, and employ a gradient descent scheme for this minimization.

3.3.3 Comparison to Baker-Kanade

The algorithm obtained in the previous subsection can be compared to the approach of Baker and Kanade for static images [37]. Their objective function

$$H_{BK} = \arg\min_{H} \ \|L - AH\|^2 + \lambda\|\nabla H - \nabla T\|^2$$

consists of reconstruction (first) and *gradient match* (second) terms, where T is a recognition-based prior image. In our method, we obtained

$$H_{MAP}, I_{MAP} = \arg\min_{H,I} \|L - AH\|^2 + \frac{\sigma_L^2}{\sigma_H^2}\|T^* + I - H\|^2 + \|L - AT^* - I\|^2. \tag{10}$$

For simplicity, consider specializing the illumination offset in our model to be constant across the image (i.e., $I(h) = I_c$ for $h = 1, 2, .., MN$). Then, Euler-Lagrange equations for the mimimization of (10) w.r.t. a high-resolution pixel h become

$$\big(A^T(L - AH)\big)(h) \ + H(h) - T^*(h) - I_c = 0. \tag{11}$$

Now consider two neighboring high-resolution pixels, $h = i$ and $h = j$, for which the first terms of (11) can be taken as approximately equal: When we backproject an error in the low-resolution reconstruction onto high-resolution pixels, we cannot distinguish the contribution of high-resolution pixel i from that of its neighbor j. Therefore, taking the difference of their constraints yields

$$H(i) - H(j) \approx T^*(j) - T^*(j). \tag{12}$$

(12), in turn, suggests an approximate match of the gradientsof H and T^* at pixel i

$$\nabla H(i) \approx \nabla T^*(i).$$

In other words, our scheme encourages the *gradients* of the hallucination H to match those in T^*. Hence, our graphical model setup and its subsequent specialization motivates the objective of *matching gradients*, just as in [37]. Again, note that [37] dealt with static images only, without modeling spatial interactions.

4 Preliminary Results

4.1 Training Data and Testing

We generated our database of face template patches from a 1200 frame-long (40 sec) video of a speaking person, where the face covered an area of 128×96 pixels. The global motion in this video was removed using a translation-only motion model.

In our learning, we used individual low resolution pixels as patches, corresponding to 16×16 pixel-wide high resolution patches in both T and H. The neighboring pixels come from the 2-pixel wide frame that surrounds each patch (Fig. 6). The feature vector stacks 12-dimensional (composed of intensity, horizontal and vertical derivatives, and

Laplacian, each computed over 3 scales) vectors for each frame within the temporal support considered.

In order to generate the test data, we used a separate, 30 frame-long video sequence of the same person, whose translational motion is removed as above. After adding translational jitter noise (zero-mean Gaussian with $\sigma = 1$ high-resolution pixel), we blurred and downsampled this test video at a resolution of 8×6 pixels (examples of such images can be seen in the top row of Fig. 10). We also added Gaussian noise (zero-mean, $\sigma = 1$) to its intensity values to account for uncertainties in sensing. Finally, since our data sets exhibited minimal change in the illumination conditions, we considered a constant illumination offset value for the entire image.

To better contrast the roles of spatial and temporal couplings, we ran multiple hallucination experiments, in which we turned these couplings on and off, and varied the range of temporal interaction from one to five frames.

4.2 Spatial Interaction

Fig. 10, displays a selected subset of frames, corresponding to time instants t=2, 4, 14, and 19, for three such settings (the entire set of frames is in video 1). In the first row, 8×6 input images are displayed, whereas the last row shows the underlying 128×96 pixel-wide ground truth images.

The second row shows hallucination results with no interaction among patches T_p (i.e., each patch in each frame is hallucinated independently, using the local Maximum Likelihood estimate computed in step 1 of Alg. 1). We observe that the results look very patchy due to blocking artifacts, and extraneous edges. For the third row, we ignore temporal interactions but enforce spatial interactions, so that hallucination is performed independently for each frame, or *frame-wise*. We note that many of the blocking artifacts have disappeared, but unfortunately, hallucinations now contain some incorrect estimates of the underlying face motions (e.g., closed vs. open eyelid and mouth).

4.3 Spatio-Temporal Interaction

In the fourth row of Fig. 10, we included representative results for temporal hallucination, where we used three frames of temporal support. First, we note hallucinations become more correct when temporal interactions are allowed (compare the opening of eyelid and mouth with spatial-only hallucinations).

Inspected as *static images*, the results in Fig. 10 already exhibit considerable improvements due to both spatial-only, and spatio-temporal modeling of the problem at hand. Moreover, as can be verified from the attached video files, we find our results as *video sequences* to be even more compelling: Frame-to-frame transitions that are not directly observable in static images can have perceptually detrimental effects when seen as a time sequence. We observe that such flicker artifacts, amply present in frame-wise hallucinations, vanish by a large extent when temporal couplings are taken into account (i.e., when two or more frames of temporal support are used). These observations show that time plays a crucial role as a regularizer in our inference.

In order to quantify the role of time, we provide an empirical analysis of the effect of various levels of temporal couplings. While varying the amount of temporal

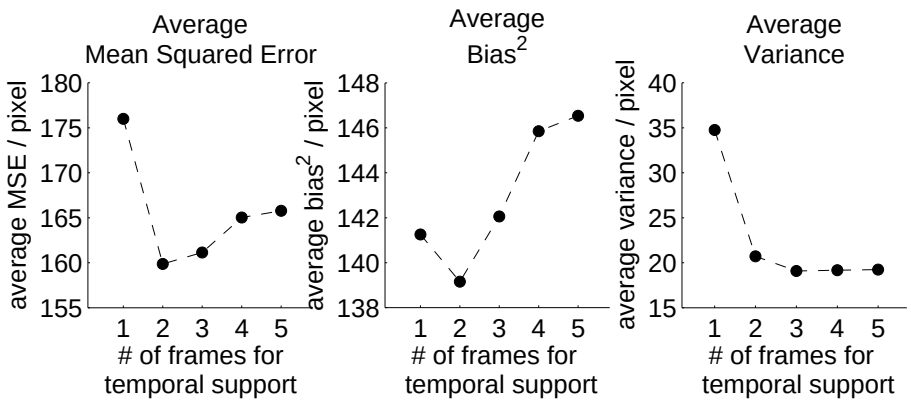

Figure 8: Bias-Variance Trade-Off: We generated and applied independent jitter and additive noise to the same input data for a total of 36 hallucination experiments, and compared all output videos against the ground truth. To summarize the measured bias and variance videos, we plot their value averaged spatially and temporally over the entire video sequence. Enforcing spatio-temporal couplings reduces the Mean Squared Error (left), primarily by reducing the variance and enhancing the stability of hallucinated videos (right). However, stronger temporal couplings induce a larger bias (middle).

support in the nodes of T from a single frame (i.e., frame-wise hallucination, using spatial coupling only) to five frames, we compared the resulting hallucination videos to the ground truth video using the L_2-norm. Fig. 8 (left) shows a noticeable drop in the Mean-Squared-Error (MSE) metric as soon as temporal couplings are considered. In fact, the Bias-Variance decomposition of MSE [45] reveals a more interesting phenomenon: Temporal models dramatically reduce the variance of our video hallucinator (Fig. 8, right), resulting in more stable videos. However, as temporal couplings become stronger, the bias also increases.

To further analyze the reduction in the amount of video flicker artifacts, we have measured frame-to-frame differences between consecutive time instants (i.e., temporal derivatives) in videos, and investigated how well these matched. Fig. 9 plots the L_2-norm of the errors (relative to the ground truth video) in estimated temporal derivatives as a function of time. We notice that errors observed in frame-wise hallucinations are consistently higher compared to those of temporal hallucinations. In addition, the variability in error is lower when temporal couplings are used (bottom curve).

5 Proposed Work

In this section, we present a roadmap of activities and study topics for the extension and validation of the hallucination framework. A few of the design choices that lead to the overall model of section 3 will be revisited, and alternatives or potential improvements will be discussed.

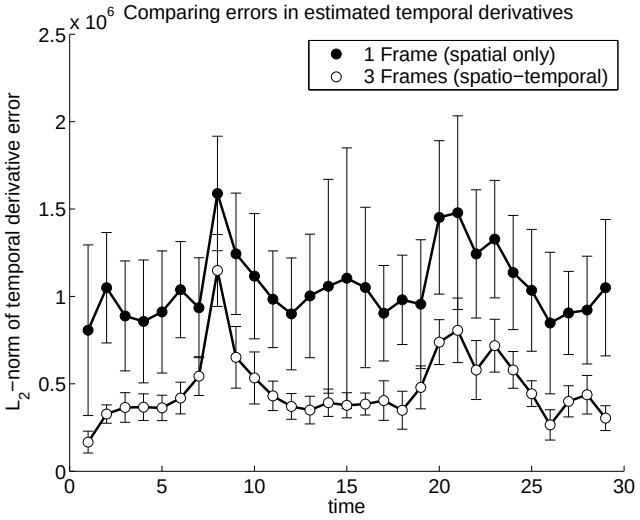

Figure 9: Incorporation of temporal couplings reduces the errors in the estimates of temporal derivatives. The two peaks observed around frames 8 and 21 are due to blinking eyes, indicating that both algorithms are challenged. Error bars indicate one standard error for a sample set of size 36.

5.1 Validation Experiments

The experimental results presented in section 4 already expose the benefits of using spatial and temporal interactions in hallucinating high-zoom videos. However, our training and testing data sets have dealt with only one subject's videos. A crucial requirement of any learning-based method is its generalization capability, and this naturally raises the question of how well our face video hallucination would generalize to people who do not appear in the training set. For the case of *static* face hallucination, earlier works of [37, 38, 39] already confirmed the feasibility of the task, for zooming factors of up to 8 (or 64 per number of pixels). For validation purposes, we will populate the training set with videos of more subjects.

5.2 Robustness Experiments

In our experiments, we considered and simulated two types of noise: The low-resolution observation noise was modeled as additive Gaussian, and the face alignment (into the rectangular area which was zoomed in) noise was simulated using translational Gaussian jitter.

We propose to extend the realism level of these experiments by incorporating partial occlusions of the face, motion blur, and intermittently missing video frames. Such experiments would simulate the kinds of video data one would expect in a surveillance or low-bandwidth communication scenario.

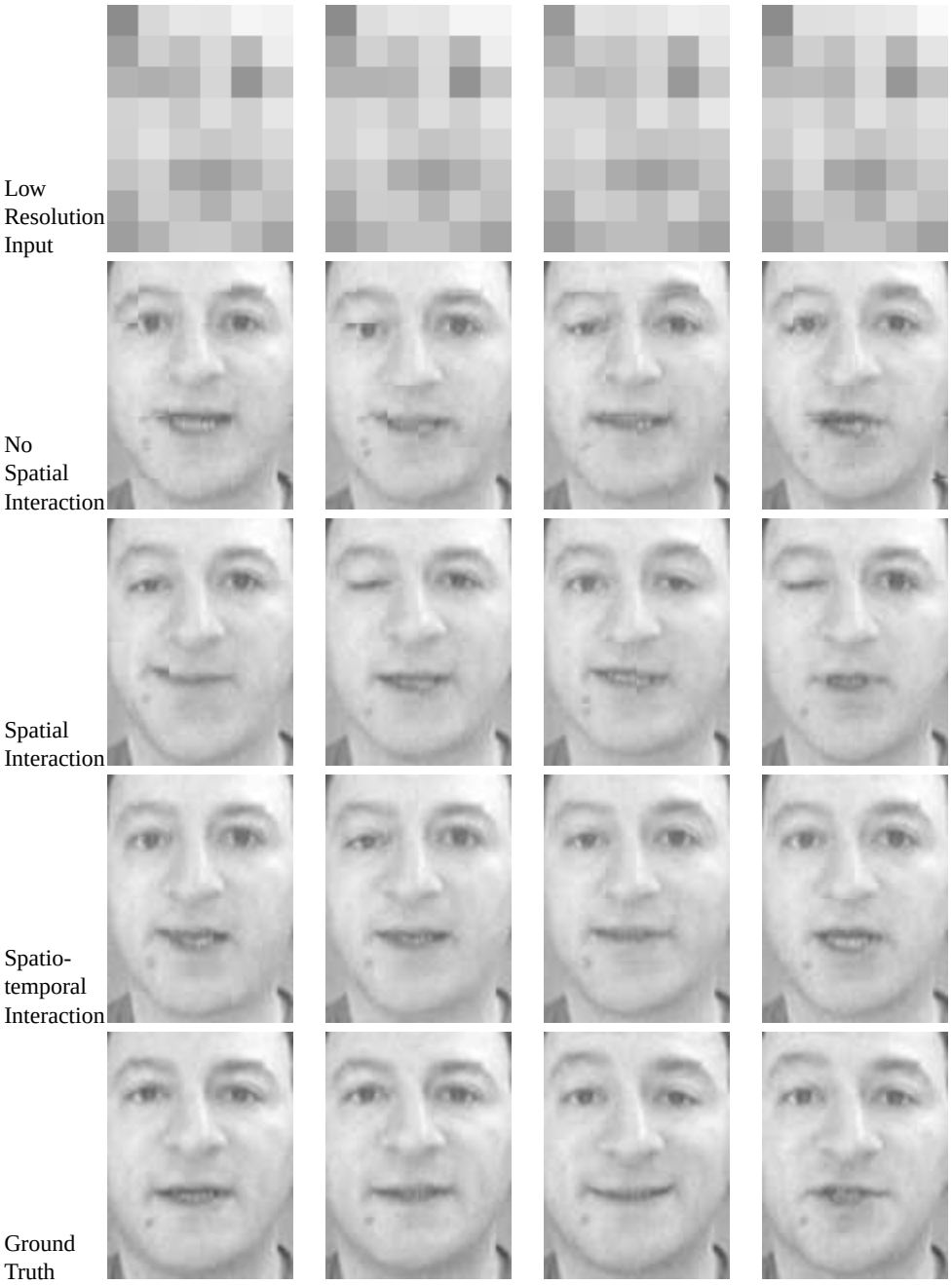

Figure 10: The regularizing role of time for video hallucination (see section 4).

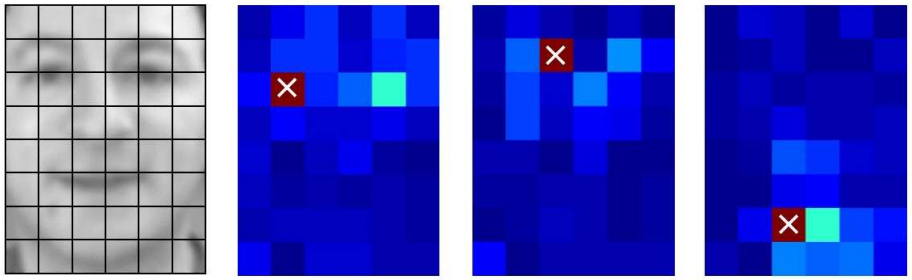

Figure 11: The leftmost image shows a high-resolution face, overlaid with the 8×6 grid of its low-resolution version. The other images show intensity-mapped estimates of the mutual information between a selected low-resolution pixel (marked with a $\times$, corresponding to the eye, eyebrow, and mouth) and all other pixels.

5.3 Domain-specific Design

As touched upon in section 2, our initial design of the graphical model and its specialization to human faces were inspired by the earlier works of [34] and [37]. In this section, we reconsider some of these design choices, and propose alternatives that may further exploit the particular domain of human face videos.

5.3.1 Graphical Model Topology

Currently, our MRF is defined on a regular 3-dimensional lattice, and clique potentials of order three and higher and assumed to be zero. The connectivity of each node in the model is limited to its 6 neighbors: 4 spatial and 2 temporal. Although theoretically this structure can already express global statistical properties, one may expect a different connectivity to be better suited to this particular domain. For example, the underlying symmetry of faces could be used explicitly, by connecting left-right side pixels and imposing additional interaction constraints. We propose to experiment with such variations on the topology of the graphical model.

5.3.2 Feature Selection for Contextual Information

In section 3.2.3, we adopted the multi-scale feature vector of [43] in an attempt to pool contextual information about a low-resolution pixel. However, the face domain exhibits a lot more structure and regularity than generic textures, for which this feature vector was originally designed. We expect features derived from face videos to be more powerful in pooling relevant information. One commonly used measure of the relevancy between variables is mutual information [46]. To give an intuition on this, as done by [47], we measured and displayed mutual information between low-resolution pixels in Fig. 11.

We observe that, while there is usually some degree of mutual dependency between immediately neighboring pixels, much stronger dependencies usually exist between the

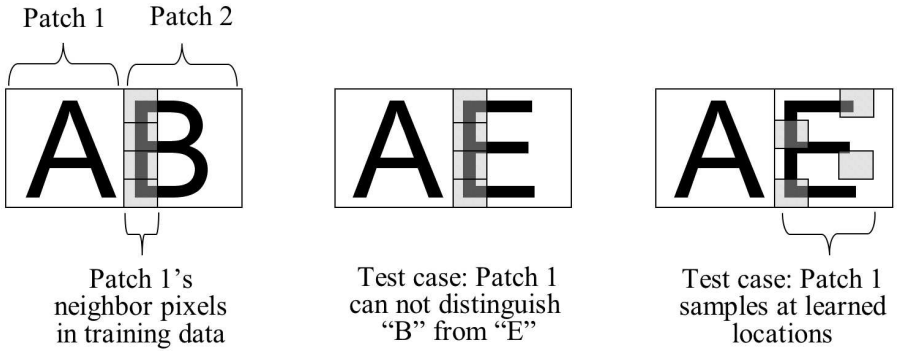

Figure 12: In this synthetic example, we show how the spatial compatibility function can be specialized to sample more informative pixels of surrounding patches. Imagine that the training set has a mixture of "AB" and "AE" patterns. Using a surrounding-pixel-strip approach, Patch 1 would be unable to distinguish between its "B" and "E" neighbors (left and middle). With a data-driven selection of sampling positions, the co-occurence of these letters can be captured more accurately (right).

left and right half pixels of faces. We propose to develop a feature selection framework for designing more effective contextual pooling mechanisms.

5.3.3 Spatial Compatibility

In section 3.2.2, we decribed our face template image prior which used pairwise compatibility functions between spatially neighboring image patches (*i.e.*, nodes of the graph). These functions captured the extend to which the co-occurence of a given pair of template patch configurations was plausible. More specifically, the *neighboring pixels* variable was used to compare a given set of neighbors with all stored in the database.

We observe that the surrounding-pixel-strip design is very much geared towards penalizing visual discontinuities along patch seams, and may even be misleading in capturing the *co-occurence* of two template patches. Consider the synthetic example shown in Fig. 12: Most informative pixels are not along the seams of patches, but elsewhere, determined by data. We propose to generalize the surrounding-pixel-strip concept to that of most-informative-pixel-set by following information theoretic measures.

5.4 Other Extensions

5.4.1 Spatially Homogeneous Priors

This work used a spatially inhomogeneous prior for the template T. While such priors require input images to be registered, they also render database referencing and feature comparison steps more efficient. Although we challenged the registration assumption with translational jitter noise, we propose to study space-invariant priors in more depth.

5.4.2 Dimensionality Reduction

As we increase the number of subjects in the video database, we expect the computational requirements to also increase. We propose to study and experiment with dimensionality reduction techniques for alleviating the computational requirements.

5.4.3 Illumination Models

Since our test data set did not include illumination variations, the additional power of our intensity offset model remains to be tested. In this respect, we propose to extend the current constant illumination offset model to linear and quadratic functions. Of potential utility would also be scene specific, but reduced-dimensional models that could be estimated using PCA.

5.5 A Framework for Spatio-Temporal Regularization ?

The spatio-temporal framework developed in this document can also be seen as a general, model-based tool for restoring corrupted measurements. In the case of video super-resolution, the corruption took the form of spatial resolution degradation. By referring to a database of low-high resolution video pairs, we were able to make reasonable predictions about the underlying, unobserved high-resolution video.

A potential application domain for our model is the silhouette enhancement task, which would, for example, benefit human identification efforts from surveillance videos [48]. It has been reported that the recognition performance of such systems is highly sensitive to the quality of extracted silhouettes: These are usually computed using background subtraction techniques, which assume that the foreground object's appearance will be different from that of the background. In situations where this assumption fails, silhouettes are not computed correctly, and have "holes". For instance, if a person's clothing color matches with anything in the background, parts of the silhouette would be missing as the person occludes those parts of the background. In noisy video sequences, such mishaps occur frequently and unpredictably, resulting in choppy silhouettes as seen in Fig. 13 (bottom).

Our spatio-temporal model can be readily used to restore silhouette videos with intermittently missing parts. The key observation is that when images are blurred and downsampled, missing parts and holes tend to be "filled in" by surrounding image structures. Therefore, by first reducing the spatial resolution of these silhouettes, and then going back to their original resolution by hallucination, we can expect to fill in the missing parts. Note that this would require a database of valid, fully observed silhouettes. Such models can be acquired under controlled background and lighting conditions, where the silhouettes can be very clean. We propose to test whether our framework can result in higher recognition rates.

6 Expected Contributions

To summarize, in this thesis proposal we formulated a framework for the task of hallucinating high-zoomed face videos. We cast the problem as one of probabilistic infer-

Figure 13: From [48], examples of human silhouette images (bottom) extracted using background subtraction.

ence, and dealt with the temporal nature of the problem directly. Through experiments, we visually displayed and quantified the benefit of incorcorating spatial *and* temporal couplings among units of estimated high-resolution videos.

This thesis is expected to make the following contributions:

- It presented a novel learning-based approach to super-resolving videos. It is original in the way it exploits both spatial and temporal aspects of the video domain, without requiring the estimation of any intermediate variables such as optical flow fields.
- It extended existing 2-dimensional spatial graphical models to a 3-dimensional space-time model to deal with video sequences. Consistencies in the time dimension are imposed via temporal signatures, which represent a *novel mode of use* of time in the field of super-resolution.
- It prescribed a versatile framework that can be used to address other computer vision tasks of spatio-temporal nature. It can be seen as a general, model-based regularization tool that can restore corrupted and missing observations.

7 Schedule

Below is a tentative schedule of development and testing activities for the completion of this thesis:

- Spring and Summer 2004: Develop theory for and implement domain-specific design of the graphical model connections, and derive new feature vector and neighborhood definitions.
- Summer 2004: Extend the video database to include multiple people for validation experiments. Study and experiment with dimensionality-reduction techniques.

- Fall 2004: Test the framework in the silhouette enhancement domain.
- Spring 2005: Work on extending the illumination model.
- Summer 2005: Write thesis and defend.

Acknowledgements

Insightful questions, comments, and suggestions from Sanjiv Kumar, Al Rizzi, Yanxi Liu, and the miscelaneous reading group are gratefully acknowledged.

References

[1] H. C. Andrews and B. R. Hunt, *Digital Image Restoration*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1977.

[2] C. A. Vogel, *Computational Methods for Inverse Problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2002.

[3] B. Chalmond, *Modeling and Inverse Problems in Image Analysis*. New York: Springer-Verlag, 2003.

[4] S. Chaudhuri, Ed., *Super-Resolution Imaging*. Boston: Kluwer Academic Publisher, 2001.

[5] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," in *IEEE Signal Processing Magazine*, May 2003, pp. 21–36.

[6] S. Borman and R. Stevenson, "Spatial resolution enhancement of low-resolution image sequences a comprehensive review with directions for future research," University of Notre Dame, Notre Dame, IN, Tech. Rep., 1998.

[7] A. Papoulis, "Generalized sampling theorem," *IEEE Transactions on Circuits and Systems*, vol. 24, pp. 652–654, November 1977.

[8] R. Y. Tsai and T. S. Huang, "Multi-frame image restoration and registration," in *Advances in Computer Vision and Image Processing*, Greenwich, CT, 1984, pp. 317–339.

[9] S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 38, pp. 1013–1027, June 1990.

[10] M. Elad and A. Feuer, "Restoration nof a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646–1658, December 1997.

[11] ——, "Super-resolution reconstruction of image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 817 – 834, 1999.

[12] E. Shechtman, Y. Caspi, and M. Irani, "Increasing space-time resolution in video," in *Proc. of the 7th European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 2350. Springer-Verlag Heidelberg, 2002, pp. 753–768.

[13] H. Ur and D. Gross, "Improved resolution from sub-pixel shifted pictures," *CVGIP: Graphical Models and Image Processing*, vol. 54, pp. 181–186, March 1992.

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

[15] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Amer. A*, vol. 6, no. 11, pp. 1715–1726, November 1989.

[16] A. J. Patti, M. I. Sezan, and A. M. Tekalp, "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Transactions on Image Processing*, vol. 6, no. 10, pp. 1064–1076, 1997.

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

[18] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson, "Super-resolved surface reconstruction from multiple images," in *Maximum Entropy and Bayesian Methods*, G. R. Heidbreder, Ed. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1996, pp. 293–308.

[19] B. Bascle, A. Blake, and A. Zisserman, "Motion deblurring and super-resolution from an image sequence," in *Proc. of the European Conference on Computer Vision (ECCV)*, 1996, pp. 573–582.

[20] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Tokyo: Springer, 2001.

[21] S. Borman and R. L. Stevenson, "Simultaneous multi-frame map super-resolution video enhancement using spatio-temporal priors," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, vol. 3, 1999, pp. 469–473.

[22] M. E. Tipping and C. M. Bishop, "Bayesian image super-resolution," in *Advances in Neural Information Processing Systems (NIPS 2002)*, vol. 15. The MIT Press, 2003.

[23] A. Storkey, "Dynamic structure super-resolution," in *Advances in Neural Information Processing Systems (NIPS 2002)*, vol. 15. The MIT Press, 2003.

[24] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *Journal on Visual Communications and Image Representation*, vol. 4, no. 4, pp. 324–335, 1993.

[25] P. E. Eren, M. I. Sezan, and A. M. Tekalp, "Robust, object-based high-resolution image reconstruction from low-resolution video," *IEEE Transactions on Image Processing*, vol. 6, no. 10, pp. 1446–1451, October 1997.

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

[27] N. R. Shah and A. Zakhor, "Resolution enhancement of color video sequences," *IEEE Transactions on Image Processing*, vol. 8, no. 6, pp. 879–885, June 1999.

[28] W. Zhao and H. Sawhney, "Is super-resolution with optical flow feasible?" in *Proc. of the 7th European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 2350. Springer-Verlag Heidelberg, 2002, pp. 599–613.

[29] S. Baker and T. Kanade, "Super resolution optical flow," Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-99-36, 1999.

[30] Z. Jiang, T.-T. Wong, and H. Bao, "Practical super-resolution from dynamic video sequenes," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, June 2003, pp. II–549 – II–554.

[31] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. of the European Conference on Computer Vision (ECCV)*, 1992, pp. 237–252.

[32] D. Capel and A. Zisserman, "Computer vision applied to super resolution," in *IEEE Signal Processing Magazine*, May 2003, pp. 75–86.

[33] F. M. Candocia and J. C. Principe, "Super-resolution of images based on local correlations," *IEEE Transactions on Neural Networks*, vol. 10, no. 2, pp. 372–380, March 1999.

[34] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25 – 47, 2000.

[35] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, March/April 2002.

[36] C. M. Bishop, A. Blake, and B. Marthi, "Super-resolution enhancement of video," in *Proceedings Artificial Intelligence and Statistics*, C. M. Bishop and B. Frey, Eds. Society for Artifi cial Intelligence and Statistics, 2003.

[37] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167 – 1183, 2002.

[38] C. Liu, H.-Y. Shum, and C.-S. Zhang, "A two-step approach to hallucinating faces: Global parametric model and local nonparametric model," in *Proc. of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. I–192 – I–198.

[39] D. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, December 2001, pp. II–627 – II–634.

[40] M. Shah and R. Jain, *Motion-Based Recognition*. Dordrecht: Kluwer Academic Publishers, 1997.

[41] C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, March 1995.

[42] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the 2003 IEEE International Conference on Computer Vision*, 2003, pp. 726–733.

[43] J. S. DeBonet and P. A. Viola, "A non-parametric multi-scale statistical model for natural images," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 10. The MIT Press, 1998.

[44] J. E. Besag, "On the statistical analysis of dirty pictures (with discussion)," *Journal of the Royal Statistical Society B*, vol. 48, no. 3, pp. 259 – 302, 1986.

[45] G. Casella and R. L. Berger, *Statistical Inference*. Belmont, California: Duxbury Press, 1990.

[46] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: John Wiley & Sons, 1991.

[47] H. Schneiderman and T. Kanade, "Object detection using the statistics of parts," *International Journal of Computer Vision*, 2002.

[48] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identifi cation from body shape and gait," pp. 351–356, October 2002.