

# High-Zoom Video Hallucination by Exploiting Spatio-Temporal Regularities

Göksel Dedeoğlu, Takeo Kanade and Jonas August  
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213  
{dedeoglu, tk, jonas}@cs.cmu.edu

## Abstract

In this paper, 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 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. Traditional approaches to this problem using video data first estimate the relative motion between frames and then compensate for it, effectively resulting 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. In our experiments, a  $8 \times 6$  pixel-wide face video, subject to translational jitter and additive noise, gets magnified to a  $128 \times 96$  pixel video. Our results show that by exploiting both space and time, drastic improvements can be achieved in both video flicker artifacts and mean-squared-error.

## 1 Learning-based Super-Resolution

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, super-resolved video that resembles the original (bottom)? In this paper, we present a model for this task, formulate it as an inference problem, and describe an algorithm for solving it.

The problem of estimating high resolution image details is commonly referred to as Super-Resolution (SR) [9], although in practice approaches may differ in their use of a single static image, a sequence of thereof, or a video of a dynamic scene [17, 13, 11, 16]. Mathematically, such problems are highly ill-posed [18], motivating the use of Bayesian techniques and generic smoothness assumptions about high resolution images [8] (Fig. 1, middle).



Figure 1: Given only a low-resolution video (top), how can one estimate (or “hallucinate”) the original high-resolution video (bottom)? Unfortunately, simple methods such as bicubic interpolation are insufficient (middle). In this paper we explore zooming using a database of videos with an inference procedure that enforces spatio-temporal consistency of the resulting hallucinated video.

Recently, learning-based approaches to SR have produced compelling results [12, 2, 5, 15, 6]. The essence of these techniques is to use a training set of high resolution images and their low resolution counterparts to build a co-occurrence model (stored either directly as image patches, or as coefficients of alternative representations). At the time of applying the learnt model, the task is to predict high resolution data from the observed low resolution data. In [12], 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 to video sequences was attempted in [4], but severe video artifacts were found. As a remedy, an *ad-hoc* solution was proposed, which consisted of *re-using* high-resolution solutions for achieving more coherent videos.

An interesting aspect of learning approaches is that they can be made much more powerful when images are limited to a particular domain: For instance, [2] considered super-resolving human faces only. 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.

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 spa-

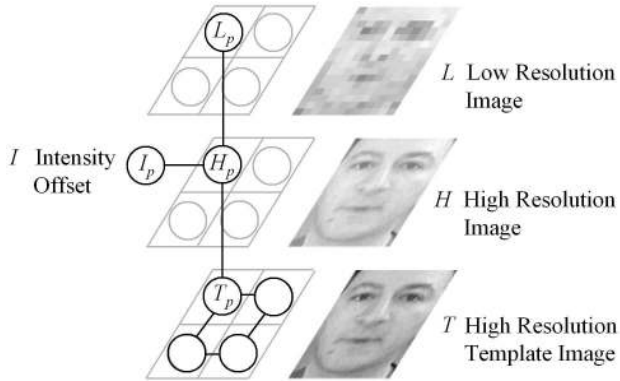


Figure 2: Model of blur and degradation (See section 2.1)

tially varying prior (as in [2]), we keep the computational requirements relatively low. Inspired by the use of spatial couplings in [12], we model both spatial *and* temporal consistencies in the super-resolved videos. In contrast to [4], 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.

## 2 Modeling the High-Zoom Problem

In this section, we present a model for the high-zoom problem, through which we integrate our domain knowledge about the videos of interest with the physical principles of image formation.

### 2.1 Generative Image Model

A graphical model is a concise tool for expressing causal and statistical dependence relationships between random variables of interest. We now introduce our graphical model for the formation of low-resolution observations. For clarity, we describe this generative model for the static image case, then extend it to the temporal dimension for videos in subsection 2.2.

Our model for low-resolution observations comprises three steps: organized upwards in Fig. 2, 1) Generation of template image  $T$ , 2) addition of illumination offset  $I$  to generate a noisy high-resolution image  $H$ , and 3) downsampling 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

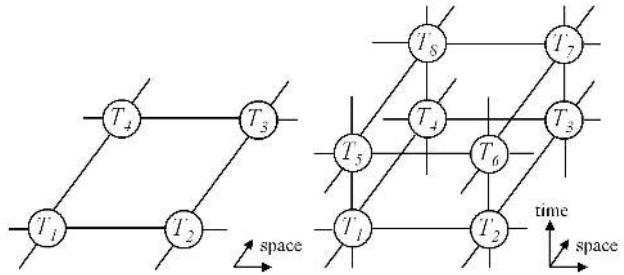


Figure 3: Spatial (left) and spatio-temporal (right) coupling between neighboring template patches is shown in the Markov random field graphs for image (2-d, left) and video (3-d, right).

them independently, as in [2]. 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-occurrence model, we will impose compatibility constraints only between neighboring patches. In particular, we will use a Markov Random Field (MRF) (Fig. 3, left) to model spatial interactions, allowing us to compose face template images without 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  $T$  to produce 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].

### 2.2 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 MRF framework into the time dimension (Fig. 3, right), 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. 4 (bottom), 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 traditional approaches to super-resolution: In the literature, the relative motion between frames is estimated, then eliminated via warping or optical flow. These approaches 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

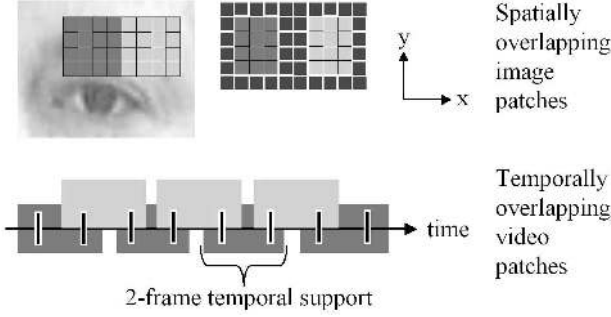


Figure 4: Implementation details of spatial (top) and temporal (bottom) overlap consistencies. The black pixels (top) indicate the locations where neighboring patches must have similar intensity. Whole image frame for overlapping video patches must agree as well for the video case (bottom)

fact, the very small size of inputs ( $8 \times 6$  pixels) considered in this work would make the recovery of facial motions (e.g., opening and closing of the eyelids and mouth, the appearance of pupils and teeth, etc.) 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 Theory of High-Zoom Inference

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  $L$ :

$$(H_{MAP}, I_{MAP}) \triangleq \arg \max_{H, I} \log P(H, I | L).$$

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

$$P(H, I | L) = \sum_T P(H, I, T | L).$$

By applying the chain rule<sup>1</sup> twice, the posterior becomes

$$\begin{aligned} & \sum_T \left\{ P(H | I, T, L) P(I, T | L) \right\} \\ &= \sum_T \left\{ P(H | I, T, L) P(I | T, L) P(T | L) \right\}. \end{aligned}$$

<sup>1</sup>The chain rule asserts that  $P(X, Y) = P(X | Y)P(Y)$ .

Using Bayes rule in the first term, the posterior becomes

$$\sum_T \left\{ \frac{P(L | H, I, T) P(H | I, T)}{P(L | I, T)} P(I | T, L) P(T | L) \right\}.$$

Observing the conditional independence<sup>2</sup>  $P(L | H, I, T) = P(L | H)$  entailed by our graphical model, and capturing the denominator by a constant  $C$ , we rewrite the posterior as

$$C \sum_T \left\{ P(L | H) P(H | I, T) P(I | T, L) P(T | L) \right\}. \quad (1)$$

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., uniquely, and to arbitrary resolution) implies that the underlying distribution  $P(T | L)$  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  $P(T | L)$  over all possible configurations of  $T$  is highly concentrated around  $T^* = T^*(L)$ , i.e.,

$$P(T | L) \approx \delta(T - T^*). \quad (2)$$

Deferring the computation of  $T^*$  until section 4, we substitute (2) into (1) so that  $P(H, I | L)$  is approximately

$$C P(L | H) P(H | I, T^*) P(I | T^*, L). \quad (3)$$

Using (3),  $H_{MAP}$  and  $I_{MAP}$  approximately maximize

$$\log P(L | H) + \log P(H | I, T^*) + \log P(I | T^*, L). \quad (4)$$

The individual terms of this objective function have natural interpretations: The first term states that through  $H$  we should aim to increase the likelihood of the *reconstructed* observation  $L$ . The second term encourages those  $H$  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* to be consistent with the assumed template  $T^*$  and observation  $L$ .

### 4 Finding the Peak Template $T^*$

Now we describe our method for computing the peak template  $T^*$  in (2) by estimating the maximum of  $P(T | L)$ . Using Bayes rule, we first rewrite this posterior in terms of likelihood and prior terms. Observing that nodes in  $L$  are

<sup>2</sup>Two nodes which are not connected by a link are independent when conditioned upon their neighbors.

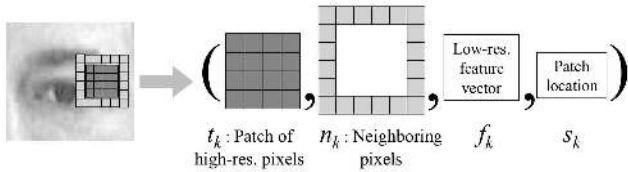


Figure 5: Each database entry contains an image patch, the neighboring pixels (for enforcing consistency), a feature vector (for matching to the low-resolution image), and its location (for supporting non-homogeneous spatial statistics). This structure is repeated for all frames within the temporal support considered.

conditionally independent given the high-resolution template  $T$ , we obtain a factorized likelihood term

$$P(T | L) \propto P(L | T) P(T) = \prod_{p=1}^N P(L_p | T_p) P(T). \quad (5)$$

Unfortunately, in the case of extremely blurred images, the likelihood term  $P(L_p | T_p)$  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 (5) will no longer be valid. In section 4.2, we will present the details of such a feature vector, and expose our assumptions for achieving a computationally tractable algorithm.

#### 4.1 The Template Prior

We restrict the space of possible  $T$ 's to a domain-specific collection of example templates. To 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. 5, 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 the 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  $\phi$  using similarity between pixel values in the overlapping areas of example patches.

$$\phi(T_p = t_k, T_q = t_i) \propto \exp\left(-\sum_{\text{overlap}} (t_k(u) - t_i(v))^2 - \sum_{\text{overlap}} (n_k(u) - t_i(v))^2\right).$$

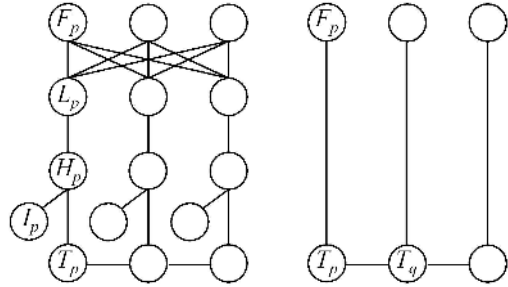


Figure 6: Interactions involved in determining optimal template  $T^*$ . For illustration purposes, a 1-dimensional version of the model in Fig. 2 is shown on left. After applying the factorization assumption, the resulting graph structure (right) is tractable enough to apply inference methods such as ICM.

#### 4.2 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 [2], we adopted the *parent vector* [10] as our feature  $F_p$ , which stacks together local intensity, gradient and Laplacian image values at multiple scales. Fig. 6 (left) shows a 1-dimensional version of Fig. 2 with the feature vector nodes added.

**Factorization Assumption.** Observe that we have two random fields,  $F$  and  $T$ , which are coupled through the image degradation model. For computational tractability, we invoke the pseudo-likelihood approximation [14] to assume that  $P(F | T)$  factorizes across feature image pixels:

$$P(F | T) \approx \prod_{p=1}^N P(F_p | T). \quad (6)$$

Correspondingly, the graphical model of Fig. 6 (left) is simplified to Fig. 6 (right).

The likelihood  $P(F_p = f_p | T_p = t_k)$  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

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

Using the factorized form (6),  $T^*$  is approximately

$$\arg \max_T \prod_{p=1}^N P(F_p | T_p) \prod_{(p,q)} \phi(T_p, T_q). \quad (7)$$

#### 4.3 ICM Algorithm for $T^*$

Maximizing the joint probability of  $T$  in (7) to obtain  $T^*$  is a non-trivial task. We adopt a greedy approach commonly

taken in the field of Bayesian image estimation: The Iterated Conditional Modes (ICM) algorithm [3] 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 */
1 for all video patches  $p$  do
  |  $T_p^* \leftarrow \arg \max_{t_k} P(F_p = f_p | T_p = t_k)$ 
  end
/* choose a video patch, and update it using its neighbors */
2 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

## 5 Hallucinating $H$

Now we introduce the details of the likelihood models of image formation and observation, and show that video hallucination (computing  $H_{MAP}$  and  $I_{MAP}$ ) only requires a quadratic minimization in addition to the ICM algorithm for computing  $T^*$ . We then note that a particular case of our framework provides a probabilistic interpretation for the objective function used in [2] for static images.

### 5.1 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  $\eta_H \sim N(0, \text{diag}(\sigma_H))$ :

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

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  $L$ :

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

### 5.2 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  $-\log P(I|T^*, L)$  is a quadratic form. Combined with the likelihood models above, we can evaluate (4). Thus,  $H_{MAP}$  and  $I_{MAP}$  minimize

$$\|L - AH\|^2 + \frac{\sigma_L^2}{\sigma_H^2} \|T^* + I - H\|^2 + \|L - AT^* - I\|^2. \quad (8)$$

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 (8) is quadratic in the unknowns  $H$  and  $I$ , and employ a gradient ascent scheme for this maximization.

### 5.3 Comparison to Baker-Kanade

The algorithm (8) obtained in the previous subsection can be compared to an existing approach for static images [2]. Specializing the illumination offset to be constant across the image (i.e.,  $I(h) = I_c$  for  $h = 1, 2, \dots, MN$ ), the Euler-Lagrange equations for the maximization in (8) w.r.t. a high-resolution pixel  $h$  become

$$(A^T(L - AH))(h) + H(h) - T^*(h) - I_c = 0. \quad (9)$$

Consider two neighboring high-resolution pixels,  $h = i$  and  $h = j$ . Taking the difference of their constraints yields

$$H(i) - H(j) \approx T^*(j) - T^*(i). \quad (10)$$

Note that for pixels  $i$  and  $j$ , the first terms of (9) are approximately equal: When we backproject an error in the low-resolution reconstruction onto high-resolution pixels, we cannot distinguish the effect of high-resolution pixel  $i$  on  $L$  from that of its neighbor  $j$ . (10), in turn, suggests an approximate match of the gradients

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

In other words, our scheme encourages the *gradients* of the hallucination  $H$  to match those in  $T^*$ . Contrasting this observation with the Baker-Kanade [2] objective function

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

we observe that our graphical model setup and its subsequent specialization motivates the objective of *matching gradients* as well. Again, note that [2] dealt with static images only, without modeling spatial interactions.

## 6 Results

### 6.1 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  $128 \times 96$  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  $16 \times 16$  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. 5). Finally, 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  $8 \times 6$  pixels (examples of such images can be seen in the top row of Fig. 8). 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.

### 6.2 Spatial Interaction

Fig. 8 displays a selected subset of frames corresponding to time instants  $t=2, 4, 14,$  and  $19$ , for three such settings<sup>3</sup>. In the first row,  $8 \times 6$  input images are displayed whereas the last row shows the underlying  $128 \times 96$  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).

<sup>3</sup>Test input, output, and ground truth videos can be downloaded from <http://www.cs.cmu.edu/~dedeoglu/cvpr04>

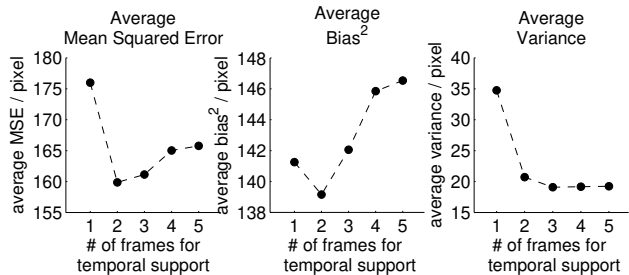


Figure 7: Bias-Variance Trade-Off: We ran 36 hallucination experiments with independent jitter and noise, and compared output videos against the ground truth. To summarize the measured bias and variance videos, we plot their value averaged spatially and temporally. Enforcing spatio-temporal couplings 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).

### 6.3 Spatio-Temporal Interaction

In the fourth row of Fig. 8, 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. 8 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, our results as *video sequences* are 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 to 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 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. 7 (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 [7] reveals a more interesting phenomenon: Temporal models dramatically reduce the variance of our video hallucinator (Fig. 7, 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 differ-

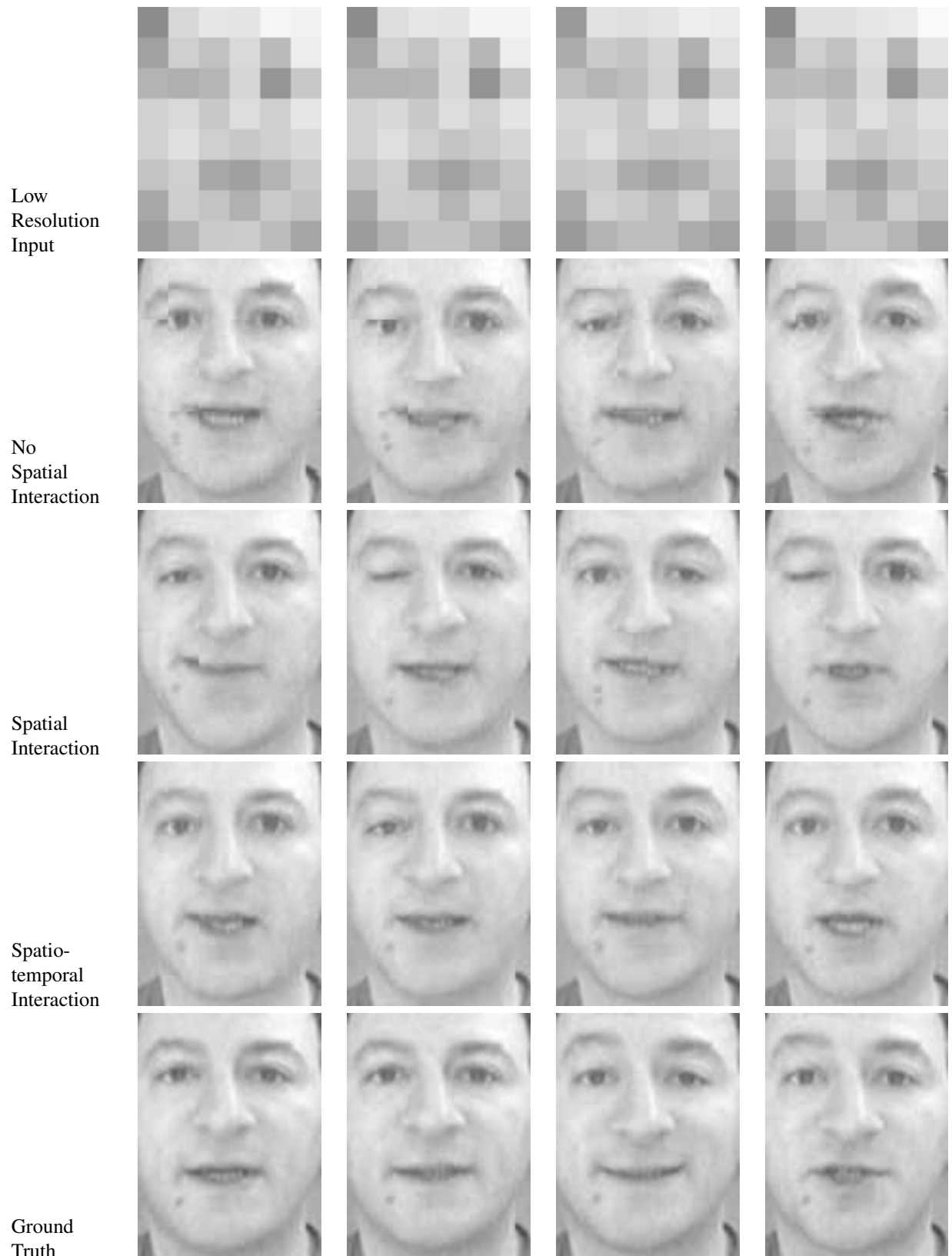


Figure 8: The regularizing role of time for video hallucination (see section 6).

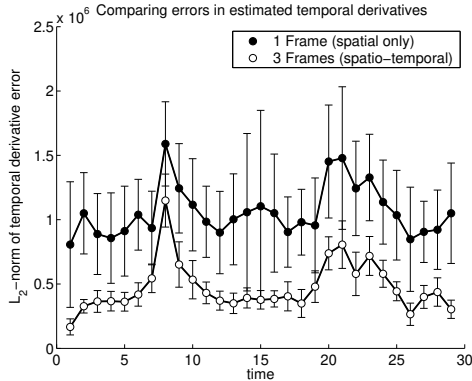


Figure 9: Incorporation of temporal couplings reduces the errors in the estimates of temporal derivatives. The two peaks observed around frames 8 and 21 are due to blinking eyes, indicating that both algorithms are challenged. Error bars indicate one standard deviation from a sample set of size 36.

ences between consecutive time instants (i.e., temporal derivatives) in videos, and we have 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).

## 7 Limitations and Conclusion

Our training and testing sets have dealt with only one subject’s videos. Yet our experimental results already expose the benefits of using spatial and temporal interactions in hallucinating high-zoom videos. In the future, we will be enlarging our database to include more subjects.

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, space-invariant priors remain to be studied. Finally, since our data set did not include illumination variations, the additional power of our intensity offset model remains to be tested.

In summary, we formulated the task of hallucinating high-zoomed face videos as one of probabilistic inference, and dealt with the temporal nature of the problem directly. Through experiments, we visually displayed and quantified the benefit of incorporating spatial *and* temporal couplings among units of estimated high-resolution videos.

## References

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

[2] 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, 2002.

[3] J. E. Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B*, 48(3):259 – 302, 1986.

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

[5] F. M. Candocia and J. C. Principe. Super-resolution of images based on local correlations. *IEEE Transactions on Neural Networks*, 10(2):372–380, March 1999.

[6] 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)*, pages II-627 – II-634, December 2001.

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

[8] B. Chalmoud. *Modeling and Inverse Problems in Image Analysis*. Springer-Verlag, New York, 2003.

[9] S. Chaudhuri, editor. *Super-Resolution Imaging*. Kluwer Academic Publisher, Boston, 2001.

[10] 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)*, volume 10. The MIT Press, 1998.

[11] M. Elad and A. Feuer. Super-resolution reconstruction of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):817 – 834, 1999.

[12] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25 – 47, 2000.

[13] 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*, 6(12):1621 – 1633, 1997.

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

[15] 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*, volume 1, pages I-192 – I-198, 2001.

[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*, 6(10):1064–1076, 1997.

[17] R. R. Schultz and R. L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996 – 1011, 1996.

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