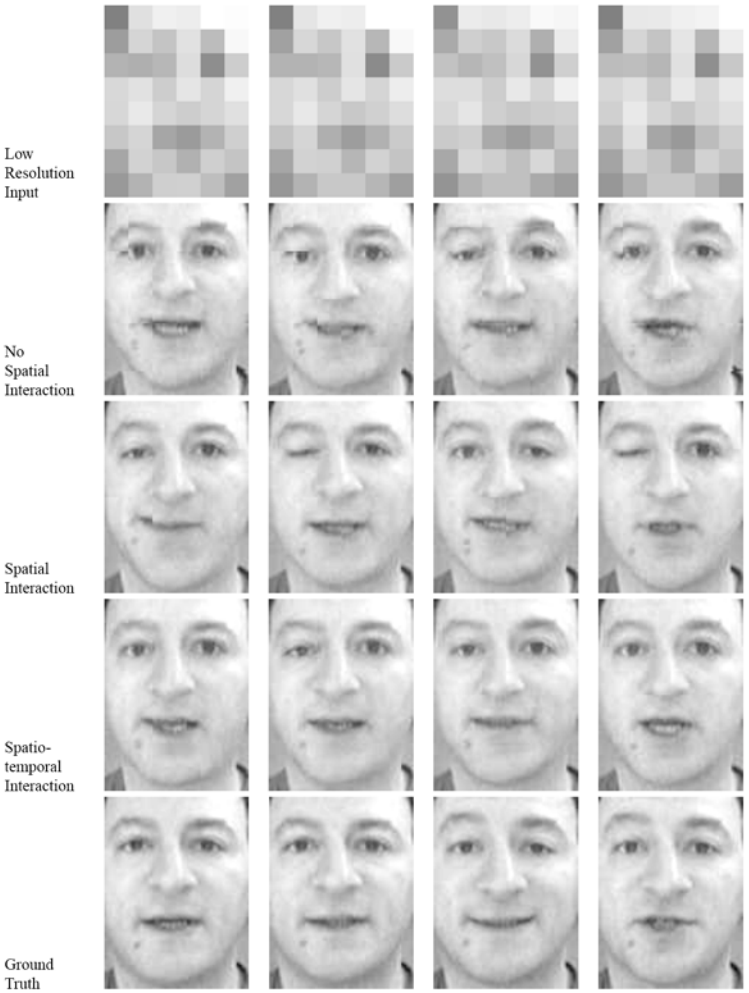
Hallucination of low resolution images

High-Zoom Video Hallucination by exploiting Spatio-Temporal Regularities

This scientific article exploits the possibilities of a new image processing area called Super Resolution. Prepared by Goksel Dedeoglu, Takeo Kanade and Jonas August from The Robotics Institute, Carnegie Mellon University, in Pittsburgh, the so-called "hallucination" concept might be very interesting for CCTV in general, forensic imaging and investigation of low quality video images as it achieves almost unbelievable reconstructing (hallucinated) factors. Movie clips showing the results from this can be downloaded from http://www.ri.cmu.edu/pubs/pub_4639.html.



Abstract

In this paper, we consider the problem of super-resolving a human face video by a very high (x16) 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 8x6 pixel-wide face video, subject to translational jitter and additive noise, gets magnified to a 128x96 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.

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), 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. Mathematically, such problems are highly ill-posed, motivating the use of Bayesian techniques and generic smoothness assumptions about high resolution images (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. The essence of these techniques is to use a training set of high resolution images and their low resolution counterparts to build a co-occurence 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 the work by W. T. Freeman, E. C. Pasztor, and O. T. Carmichael: Learning low-level vision, in the International Journal of Computer Vision, 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 the work by C. M. Bishop, A. Blake, and B. Marthi: Super-resolution enhancement of video, in the magazine of the Society for Artificial Intelligence and Statistics, 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, the work by S. Baker and T. Kanade: Limits on super-resolution and how to break them, in the IEEE Transactions on Pattern Analysis and Machine Intelligence; 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 spatially varying prior, we keep the computational requirements relatively low. Inspired by the use of spatial couplings, we model both spatial and temporal consistencies in the super-resolved videos. In contrast to some of the previously mentioned works, 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.

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.

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

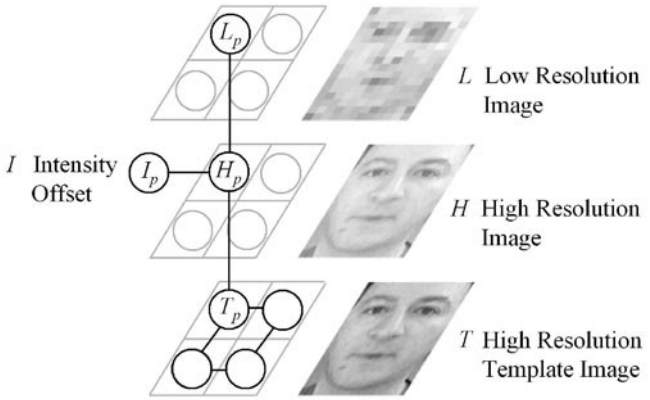


Figure 2: Model of blur and degradation

adopt local models, defined over image patches, and treat them independently.

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 neighbouring 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.

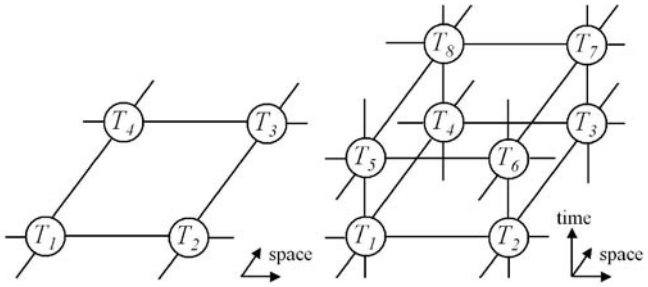


Figure 3: Spatial (left) and spatio-temporal (right) coupling between neighboring template patches is shown in the Markov random field graphs for image (2-d, left) and video (3-d, right).

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.

Exploiting Time

Just as neighbouring 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

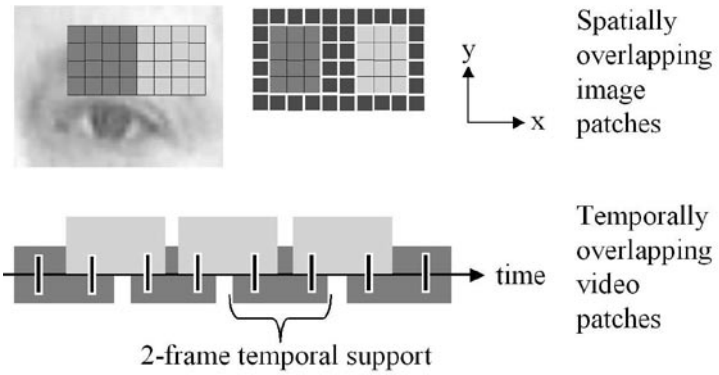


Figure 4: Implementation details of spatial (top) and temporal (bottom) overlap consistencies. The black pixels (top) indicate the locations where neighboring patches must have similar intensity. Whole image frame for overlapping video patches must agree as well for the video case (bottom)

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 fact, the very small size of inputs (8x6 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.

Results

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 x 96

www.cctv-focus.com

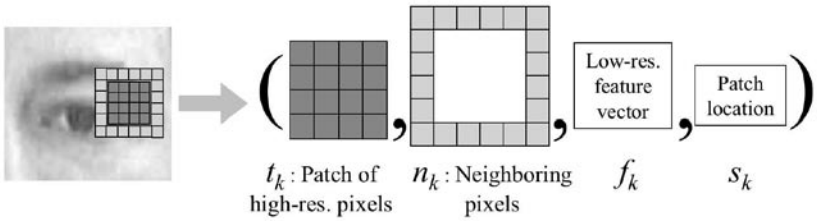


Figure 5: Each database entry contains an image patch, the neighboring pixels (for enforcing consistency), a feature vector (for matching to the low-resolution image), and its location (for supporting non-homogeneous spatial statistics). This structure is repeated for all frames within the temporal support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

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 x 16 pixel-wide high resolution patches in both T and H. The neighbouring 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 x 6 pixels (examples of such images can be seen in the top row of Fig. 6). 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.

Spatial Interaction

Fig. 6 displays a selected subset of frames corresponding to time instants t=2, 4, 14, and 19, for three such settings 3. In the first row, 8 x 6 input images are displayed whereas the last row shows the underlying 128 x 96 pixel-wide ground truth images.

The second row shows hallucination results with no interaction among patches Tp (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).

Spatio-Temporal Interaction

In the fourth row of Fig. 6, 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

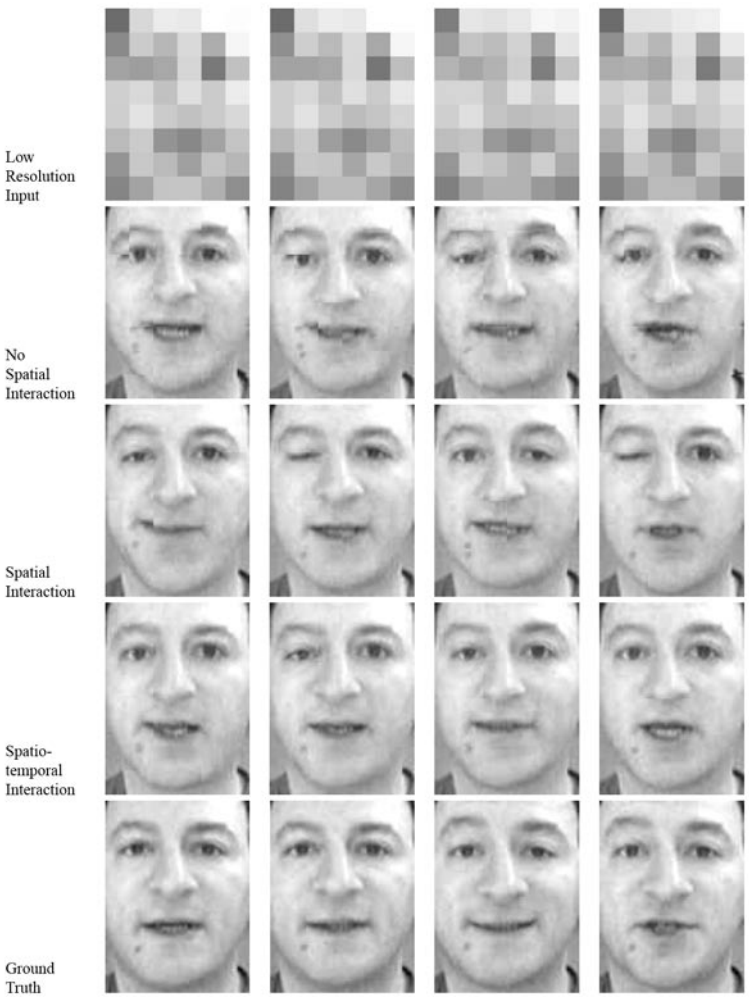


Figure 6: The regularizing role of time for video hallucination.

www.cctvfocus.com
CCTV focus

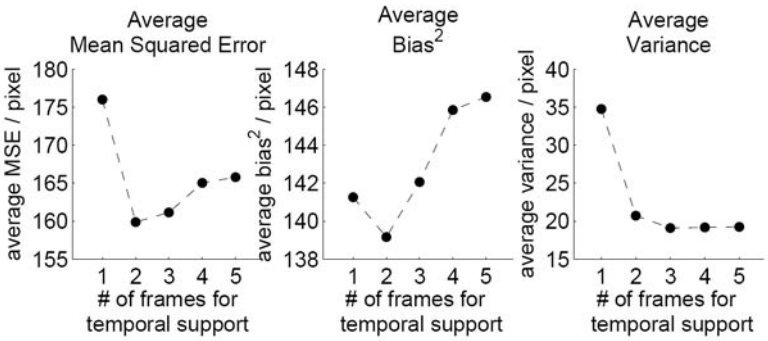


Figure 7: Bias-Variance Trade-Off: We ran 36 hallucination experiments with independent jitter and noise, and compared output videos against the ground truth. To summarize the measured bias and variance videos, we plot their value averaged spatially and temporally. Enforcing spatio-temporal couplings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

spatial-only hallucinations).

Inspected as static images, the results in Fig. 6 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 video files (download at *http://www.ri.cmu.edu/pubs/pub_4639.html*), 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 regulator in our inference.

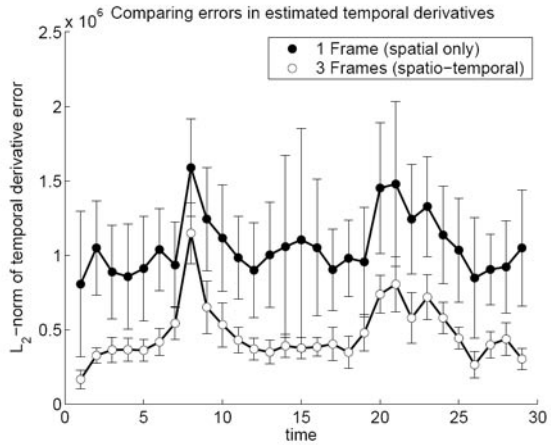


Figure 8: Incorporation of temporal couplings reduces the errors in the estimates of temporal derivatives. The two peaks observed around frames 8 and 21 are due to blinking eyes, indicating that both algorithms are challenged. Error bars indicate one standard deviation from a sample set of size 36.

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 L2-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 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 differences between consecutive time instants (i.e., temporal derivatives) in videos, and we have investigated how well these matched. Fig. 8 plots the L2-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).

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. [•]

More info at http://www.ri.cmu.edu/pubs/pub_4639.html.

www.cctv-focus.com