

A Conditional Random Field Model for Video Super-resolution

Dan Kong[†] Mei Han[‡] Wei Xu[‡] Hai Tao[†] Yihong Gong[‡]

Department of Computer Engineering [†]
University of California, Santa Cruz
Santa Cruz, CA 95064
{kongdan,tao}@soe.ucsc.edu

NEC Labs America[‡]
10080 North Wolfe Road
Cupertino, CA 95014
{meihan,xw,ygong}@sv.nec-labs.com

Abstract

In this paper, we propose a learning-based method for video super-resolution. There are two main contributions of the proposed method. First, information from cameras with different spatio-temporal resolutions is combined in our framework. This is achieved by constructing training dictionary using the high resolution images captured by still camera and the low resolution video is enhanced via searching in this customized database. Second, we enforce the spatio-temporal constraints using the conditional random field (CRF) and the problem of video super-resolution is posed as finding the high resolution video that maximizes the conditional probability. We apply the algorithm to video sequences taken from different scenes using cameras with different qualities and promising results are presented.

1. Introduction

Super-resolution has become an active research topic in computer vision lately. Super-resolution techniques have many applications ranging from video quality enhancement, object recognition to image compression.

Most of the previous super-resolution algorithms employ the reconstruction-based approach [6][9][5][1] which requires multiple low resolution images to be aligned in sub-pixel accuracy. Therefore, an accurate alignment is the key to the success of reconstruction-based methods. However, in practice, it is challenging for this type of methods to handle arbitrary video sequences because they commonly use simple global parametric transformations (e.g., affine transform), which are not capable of modeling dynamic videos that contain complex object motions.

Learning-based methods have been successfully applied to both image and video super-resolution recently [4][11][8][12][2][3]. The basic idea is to assemble a large

database of patch pairs from high resolution images and their corresponding smoothed and down-sampled ones. The dictionary can then be used to increase the resolution of images and videos by adding appropriate high frequency information via a search procedure.

Inspired by image hallucination with primal sketch priors [11], we propose a novel video super-resolution method with two main contributions. First, the training dictionary is constructed from the scene itself instead of general image pairs. Second, we enforce the spatio-temporal constraint based on conditional random field (CRF) to obtain much smoother and more continuous results.

2. The approach

2.1. Motivation

The motivation of the proposed method is based on two observations. The first observation is that cameras of different spatio-temporal resolutions can provide complementary information. Therefore, we can combine information obtained by still cameras (which have very high spatial-resolution, but extremely low temporal-resolution), with information obtained from standard video cameras (which have low spatial-resolution but higher temporal resolution), to obtain an improved video sequences of higher resolution. This principle is employed in [10] for space-time video super-resolution.

Indeed, with the rapid progress in image and video capturing devices, combining information captured by cameras with different spatio-temporal resolutions has become easier than ever before. Many latest digital camcorders, webcams, and cellphone cameras support dual-mode operations that can take high-resolution photos in photo-shooting mode, but low-resolution videos in video-shooting mode. We can take advantage of this functionality by operating the camera in video-shooting mode for most of the time,

and switching the camera to photo-shooting mode regularly to capture high-resolution photos of the target scene.

The second observation is that learning-based methods can be much more powerful when images are limited to a particular domain. Due to the intrinsic ill-posed property of super-resolution, the prior model of images plays an important role in regularizing the results. The modeling complexity can be reduced remarkably if we construct the prior model on image patches instead of full-size images, and on image patches from a particular domain instead of arbitrary images.

2.2. Scene-specific priors

Scene-specific priors is an extension to previous primal sketch-based priors for image hallucination [11]. The basic idea is to represent the priors of image primitives (edge, corner, etc.) using examples and the hallucination is only applied to the primitive layer.

The generalization capability of training data determines the successfulness of example-based super-resolution. To measure the generalization capability of training data, we define two terms. The first is **sufficiency**, which determines whether or not an input sample can find a good match in the training dictionary. The advantage of primal-sketch over arbitrary image patch is demonstrated by the statistical analysis on an empirical data set in [11]. The conclusion is that primal sketch priors can be learned well from a number of examples that we can computationally afford. The second is **predictability**, which determines whether or not the high resolution patch corresponding to the input sample's nearest neighbor in the dictionary is a good prediction of what we want to infer from the input sample. For super-resolution, since many high resolution patches, when smoothed and down-sampled, will give the same low resolution patch, higher **predictability** also means lower randomness of this mapping relationship. In our approach, we improve both **sufficiency** and **predictability** by constructing the high resolution training data from the same scene as the input low resolution videos. The scene-specific dictionary gives us a customized prior and is adaptively updated over time. Thus, fewer examples are required to achieve the **sufficiency** and to increase **predictability**.

As described above, capturing video sequences with periodic insertions of high-resolution frames can be accomplished by using a dual-model camera, and by periodically switching the camera between photo-shooting and video-shooting modes.

2.3. Spatio-temporal Constraint

To make the super-resolution video sequence smooth, we use the Conditional Random Field (CRF) model [7] to en-

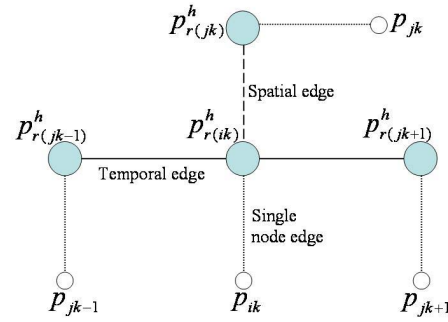


Figure 1. Conditional random field used in the algorithm.

force spatio-temporal constraints. With this framework, the problem of video super-resolution is posed as finding the high resolution video V_H that maximizes the conditional probability $P(V_H|V_L, D)$, given the input low resolution video V_L and the scene-specific dictionary D . The conditional probability is typically defined by a set of potential functions $\psi(c; \Lambda)$ as follows:

$$P(V_H|V_L, D) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c; \Lambda) \quad (1)$$

where G is the undirected graph defining the spatio-temporal dependencies among frame patches, $C(G)$ is the set of cliques in the graph, Λ is the parameter set of the CRF model, and $Z_\Lambda = \sum_{V_H} \prod_{c \in C(G)} \psi(c; \Lambda)$ is the partition function that normalizes the probability distribution. Generally, a potential function takes the form $\psi(c; \Lambda) = \exp(\lambda_c f(c))$. We construct the undirected graph as follows. Each node ik in the graph G corresponds to a primitive patch p_{ik} from a video frame k ; each edge connects either two overlapping patches in the same frame (spatial edge), or two overlapping patches (p_{ik}, p_{jm}) , where $m = k-1, k+1$ in adjacent frames (temporal edge) (See Figure 1). The temporal edges are determined by computing the optical flows between current frame k and adjacent frames $k-1, k+1$. Three types of cliques are defined on the graph: (1) single node clique that is composed of each node i only, (2) spatial edge clique that is formed by a spatial edge, and (3) temporal edge clique that contains a temporal edge from the graph. For these cliques, we define the following three potential functions: (1) The potential function defined on single node cliques

$$f_1(i, k, D) = d_1(p_{ik}, p_{r(i k)}^l) \quad (2)$$

where p_{ik} is the patch at location i in frame k , $p_{r(i k)}^l$ is a low frequency patch in dictionary D , whose corresponding high frequency patch $p_{r(i k)}^h$ is to be used to add upon p_{ik} in constructing the high resolution video frame k , and $d_1(\cdot, \cdot)$ is the Sum of Squared Difference (SSD) between the two

image patches. This potential function serves to assign the high frequency patch from the dictionary that is the most similar to p_{ik} . (2) The potential function defined on spatial edge cliques

$$f_2(i, k, D) = \sum_{jk \in SE(ik)} d_2(p_r^h(i_k), p_r^h(j_k)) \quad (3)$$

where $SE(ik)$ denotes the set of nodes connected by spatial edges to node ik , $p_r^h(i_k), p_r^h(j_k)$ are the high resolution patches from dictionary D assigned to the nodes ik and jk , respectively, and $d_2(\cdot, \cdot)$ is the SSD of the overlapping region between the two image patches. This potential function enforces spatial compatibility by encouraging the assignment of similar high resolution patches to neighboring nodes. (3) The potential function defined on temporal edge cliques

$$f_3(i, k, D) = \sum_{jm \in TE(ik)} d_3(p_r^h(i_k), p_r^h(j_m)) \quad (4)$$

where $TE(ik)$ denotes the set of nodes connected by temporal edges to node ik , and $d_3(\cdot, \cdot)$ is the SSD of the overlapping region between the two image patches after the motion compensation based on optical flows. This potential function enforces temporal compatibility by encouraging the assignment of similar high resolution patches to the overlapping nodes between adjacent video frames. Substituting the above potential functions into Eq. (1), we have

$$P_\Lambda(V_H|V_L, D) = \frac{1}{Z_\Lambda} \exp\left[\sum_{k=1}^N \sum_{i \in F(k)} (\lambda_1 f_1(i, k, D) + \lambda_2 f_2(i, k, D) + \lambda_3 f_3(i, k, D))\right] \quad (5)$$

where N is the total number of frames in the input video V_L , $F(k)$ denotes the set of primitive patches in frame k , and $\lambda_1, \lambda_2, \lambda_3$ are the weights that control the relative importance of each potential function.

Finding the global maximum of (5) is not trivial. In our implementation, we select K -nearest matching pairs from the dictionary for each low resolution primitive and use Gibbs sampling with simulated annealing to approximate the global maximum by iteratively selecting high resolution patch from the K candidates. To show how the spatio-temporal constraint can improve the results, we zoom into a region of two adjacent frames shown in figure 2. As observed from the images, smoother solutions are obtained by adding spatio-temporal constraints.

3. Experimental results

We applied the algorithm to three video clips. The first one is taken by a SONY video camcorder. The scene in

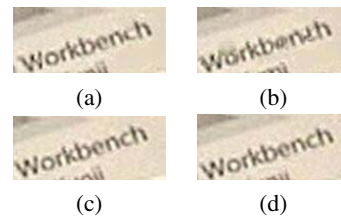


Figure 2. Comparison of video super-resolution results. (a)(b) Independent super-resolution of each frame. (c)(d) Super-resolution with temporal smoothing.

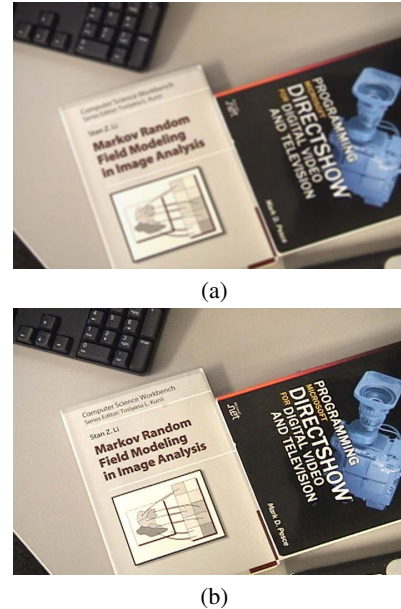


Figure 3. Super-resolution results for frame 11 of book sequence. (a) Bi-cubic interpolation (b) Our approach

this sequence consists of two books. To simulate the hybrid camera and do evaluation, we shot the video at high resolution 720×480 . We pick a high resolution image once every 15 frames and reduce the resolution of the rest to 240×160 . Approximately 10,000 primitives are extracted from every high resolution frame to create the scene-specific dictionary. Next, those low resolution frames are processed so that their resolution is increased three times in both dimensions using the temporally adaptive scene-specific dictionary. Figure 3 shows the result for one frame from the book sequence. It can be seen that our method outperforms the Bi-cubic interpolation by recovering sharp details of the scene. The second clip is taken by the same camcorder but capturing a moving human face. This time, since the relative motion between the object and the camera is non-rigid, accurate motion estimation is hard and a large face dictionary is required to produce good results. However, by using scene-specific dictionary, the spatial resolution is increased three

times in both directions, which reveals the high frequency details on the face, as shown in figure 4. We did our last experiment using a SONY USB web camera. This camera can take 30 frames/s video at 320×240 spatial resolution and still pictures at 640×480 spatial resolution. We shot a 200-frame keyboard video sequence by alternating the two modes. For each 320×240 frame, we also down-sampled it to 160×120 and applied the super-resolution algorithm to increase its resolution four times in both dimensions. The results for this sequence are shown in figure 5

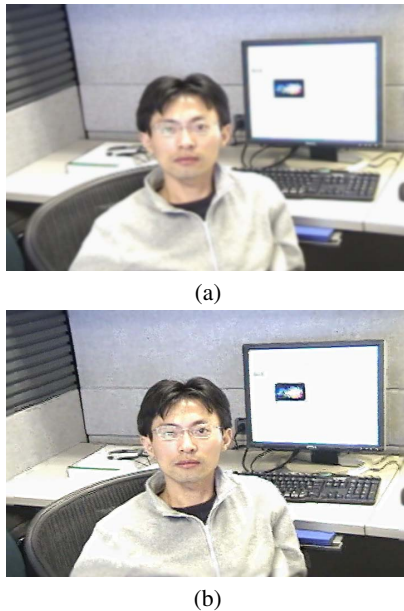


Figure 4. Super-resolution results for frame 42 from the face video sequence (a) Bi-cubic interpolation (b) Our approach

4. Conclusions

In this paper, we propose a novel learning-based video super-resolution algorithm with two key contributions. First, we combine the information from cameras with different spatial-temporal resolution by constructing scene-specific dictionary from high resolution image of the scene. Second, we integrate the spatio-temporal constraint into our super-resolution algorithm using Conditional Random Field (CRF) to obtain smooth and continuous super-resolved videos. Encouraging results are obtained for various video sequences captured using different cameras.

References

- [1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *PAMI*, 24(9):1167–1183, 2002.

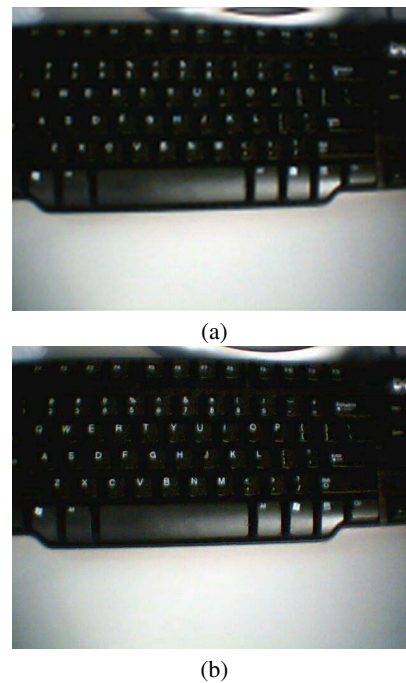


Figure 5. Super-resolution results for frame 87 from the keyboard video sequence. (a) Bi-cubic interpolation (b) Our approach

- [2] C. M. Bishop, A. Blake, and B. Marthi. Super-resolution enhancement of video. In *Proceedings Artificial Intelligence and Statistics*, 2003.
- [3] G. Dedeoglu, T. Kanade, and J. August. High-zoom video hallucination by exploiting spatio-temporal regularities. In *CVPR04*, pages II: 151–158, 2004.
- [4] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *IJCV*, 2000.
- [5] R. Hardie, K. Barnard, and E. Armstrong. On the fundamental limits of reconstruction-based super-resolution algorithms. *IEEE Trans. on Image Processing*, 1997.
- [6] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP*, (3):231–239, 1993.
- [7] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML01*, pages 282–289, 2001.
- [8] C. Liu, H. Shum, and C. Zhang. A two-step approach to hallucinating faces: global parametric model and local non-parametric model. *CVPR*, 2001.
- [9] R. Schultz and R. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Trans. on Image Processing*, 1996.
- [10] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *PAMI*, 27(4):531–545, April 2005.
- [11] J. Sun, N. N. Zheng, H. Tao, and S. H.-Y. Generic image hallucination with primal sketch prior. *CVPR*, 2003.
- [12] Q. Wang, X. Tang, and H. Shum. Patch based blind image super resolution. In *ICCV05*, pages I: 709–716, 2005.