WEB CARTOON VIDEO HALLUCINATION

Zhiwei Xiong[1*], *Xiaoyan Sun*[2], *Feng Wu*[2]

[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research Asia, Beijing, China

ABSTRACT

This paper addresses the super-resolution problem for low quality cartoon videos widely distributed on the web, which are generated by downsampling and compression from the sources. To effectively eliminate the compression artifacts and meanwhile preserve the visually salient primitive components (e.g., edges, ridges and corners), we propose an adaptive regularization method depending on the degradation grade of each frame, followed by learning-based pair matching to further enhance the primitives in the upsampled frames. In addition, temporal consistency is considered a directive constraint in both the regularization and enhancement processes. Experimental results demonstrate our solution achieves a good balance between artifacts removal and primitive enhancement, providing perceptually high quality super-resolution results for various web cartoon videos.

Index Terms— adaptive regularization, web video, compression artifacts, primitive enhancement, super-resolution

1. INTRODUCTION

With the flourish of internet, video is becoming more and more popular on the web. Especially, online video websites have experienced an explosion of visits during the past few years, due to their convenient access, rich contents and good interactivity (e.g. open upload and download privileges). However, since the server storage and network bandwidth are limited, web videos are usually in a low quality version degraded from the sources.

The video quality degradation process includes two aspects: downsampling and compression. Downsampling exploits the correlation in the spatial domain while compression for that in the frequency and temporal domains. Degradation inevitably leads to information loss which behaves as various artifacts, e.g., blurring, ringing and blocking. Quality degradation helps lower the storage and bandwidth cost of web videos, but at the expense of impairing the perceptual experience of users.

Tremendous works have been devoted before on video quality improvement in respect of compression artifacts removal and resolution enhancement. The former belongs to the field of post-processing while the latter falls into the scope of super-resolution. Combinations of these two techniques have also been reported in several literatures [1, 2] to solve the super-resolution problem of compressed video. These methods generally assume a prior distribution of the quantization noise and then integrate this knowledge into the Bayesian super-resolution framework, which requires multi-frame low-resolution observations to be aligned in sub-pixel accuracy. However, in practical applications such as web video super-resolution, the compression artifacts caused by quantization noise are largely dependent on the video content and difficult to be modeled with an explicit distribution. Moreover, since the performance of Bayesian super-resolution heavily relies on the accuracy of motion estimation and frame registration, these methods are not capable of reconstructing high frequency details of dynamic videos that contain fast and complex object motions.

In this paper, we propose a combination of adaptive regularization and learning-based primitive enhancement for the super-resolution of web videos. Similar idea has been applied to enlarge web images in our previous work [3]. Extending that to video, two main contributions are presented in this work. First, we investigate the energy change in the primitive and non-primitive fields during iterative regularization, from which an appropriate regularization degree is determined to best balance the artifacts removal and detail preservation. Second, primitives stably existing in the regularized frame are enhanced with learning-based pair matching from a prepared database, and the final results are produced through compatibility optimization in both the temporal and spatial domains. Different from conventional methods, our solution does not require any assumption on the quantization noise or the object motion. Since our scheme mainly focuses on the primitive components, we use cartoon videos which are typically made up of primitives to evaluate its performance.

The rest of this paper is organized as follows. Section 2 gives an overview of the proposed solution. The adaptive regularization and primitive enhancement are detailed in Section 3 and Section 4. Experimental results are presented in Section 5, and Section 6 concludes the paper.

2. OVERVIEW OF THE PROPOSED SOLUTION

The framework of our solution for web video super-resolution is shown in Fig. 1. It consists of three steps. Firstly, a k-th frame $F_{L,k}$ from a low-resolution video is upsampled to the desired resolution through bicubic interpolation, and the upsampled frame $F_{U,k}$ is then divided into primitive field and non-primitive field. Secondly, an iterative PDE-based regularization [4] is performed on $F_{U,k}$, during which the energy change in both the primitive and non-primitive fields are recorded. When the ratio of these two energy change converges (judged by a parameter R_k, which is related to that of the previous frame R_{k-1}), regularization stops and the accumulated noise image $I_{N,k}$ is subtracted from $F_{U,k}$, resulting in a regularized frame $F_{R,k}$. Thirdly, the primitive components in $F_{R,k}$ are enhanced with learning-based pair matching [5]. Each primitive patch extracted from $F_{R,k}$ finds its corresponding enhancing patch in a prepared database D consisting of a large collection of patch pairs. Meanwhile, the temporal consistency is enforced by involving the

[1] *This work was done during Z. Xiong's internship at Microsoft Research Asia.

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

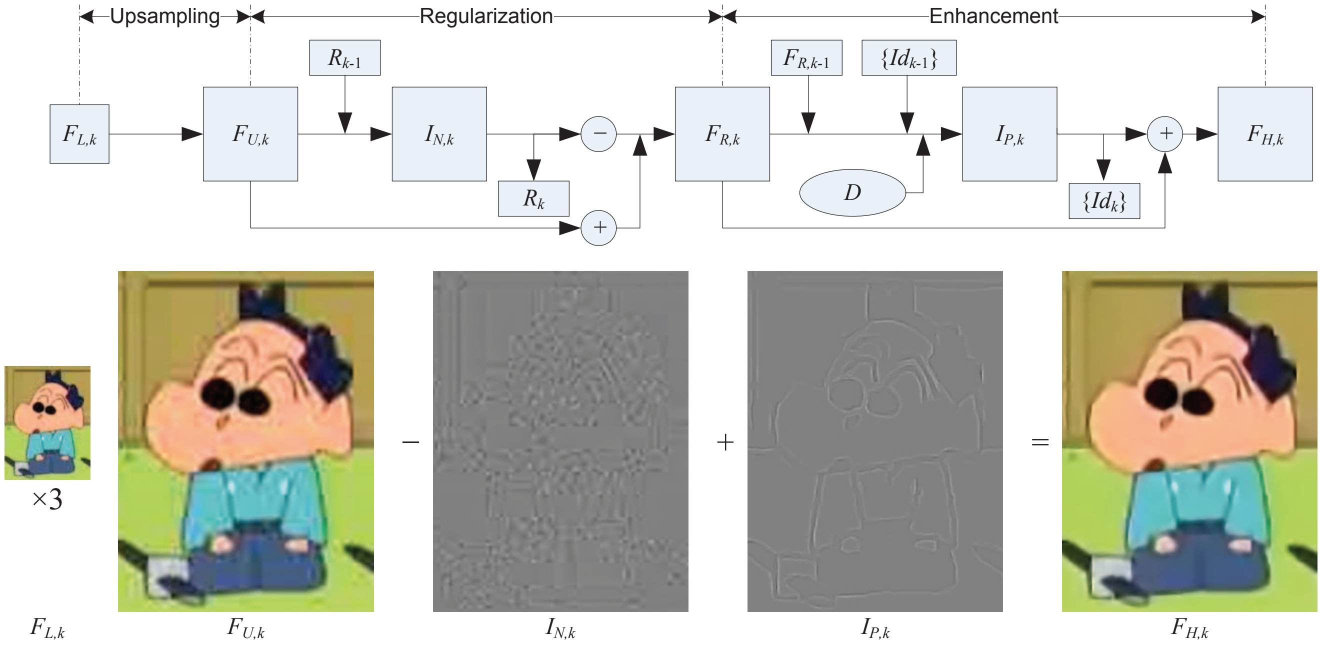


Fig. 1. Framework of our solution for web video super-resolution (Note $I_{N,k}$ is only one component of a vector-valued image)

previous regularized frame $F_{R,k\text{-}1}$ and its pair matching indices $\{Id_{k\text{-}1}\}$. Adding the primitive enhancing image $I_{P,k}$ back to $F_{R,k}$, the final high-resolution frame $F_{H,k}$ is generated. A practical example is given in Fig. 1 to visualize this framework.

3. ADAPTIVE REGULARIZATION

3.1. PDE-based image regularization

Anisotropic regularization PDE's have played an important role in the field of image processing, due to their ability to smooth data while preserving visually salient features (primitives) in images. For example, the PDE proposed in [4] has been demonstrated a common framework for many different applications, such as image restoration, image inpainting and compression artifacts removal. Our adaptive regularization is also based on this PDE, and here we only give a brief restatement.

Suppose $\mathbf{I}:\Omega\subset R^2\rightarrow R^N$ is a vector-valued image, it is updated in an iterative regularization process. The PDE velocity can be locally estimated by applying a spatially varying Gaussian smooth mask **G** over **I**:

$$\Delta I_n=\sum\nolimits_{i,j=-1}^{1}\mathbf{G}(i,j)I_n(u-i,v-j),\quad n=1,\cdots,N \tag{1}$$

where u, v are the pixel indices, I_n's are the vector components of **I**, and **G** is calculated from a local anisotropic tensor field of **I** (refer to [4] for detail). After each regularization period, there is

$$\mathbf{I}_{t+1}=\mathbf{I}_t+\lambda\Delta\mathbf{I}_t \tag{2}$$

where λ is a positive constant representing the updating step, and the subscript t denotes the regularization period.

Usually, the total regularization times are manually determined according to the image degradation grade. For compressed videos, however, the quantization noise can vary much even in a short sequence due to fast motion or scene switch. One problem arises here that how to adaptively control the regularization degree to effectively eliminate the compression artifacts while best preserving the primitive components? In other words, we need to find the turning point when regularization loses its efficacy in distinguishing primitives from artifacts.

3.2. Primitive and non-primitive fields

As mentioned above, we analyze the regularization adaptivity from the energy change point of view. For that we first divide an image/frame into primitive field and non-primitive field. In our scheme, this partition is determined by the Canny edge detection result. Suppose $\Gamma_{\mathbf{I}}:\Omega\subset R^2\rightarrow\{0,1\}$ is a bitmap storing the detected edge pixels in **I**

$$\Gamma_{\mathbf{I}}(u,v)=\begin{cases}1, & (u,v)\text{ is an edge pixel}\\ 0, & \text{else}\end{cases} \tag{3}$$

Then the primitive field $P_{\mathbf{I}}$ is defined as

$$P_{\mathbf{I}}(u,v)=\begin{cases}1, & \Gamma_{\mathbf{I}}(s,t)=1,\ (s,t)\in N_\rho(u,v)\\ 0, & \text{else}\end{cases} \tag{4}$$

$N_\rho(u,v)$ refers to a ρ-th order neighborhood of (u,v)

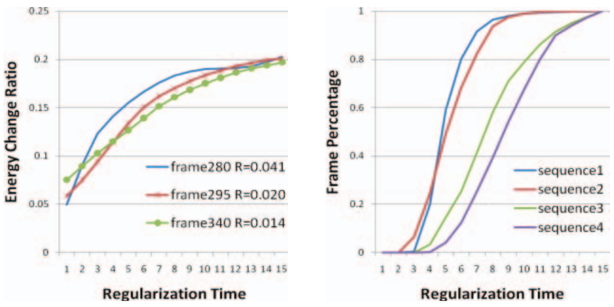
$$N_\rho(u,v)=\{(s,t):(s-u)^2+(t-v)^2<\rho^2\} \tag{5}$$

Consequently, the non-primitive field $Q_{\mathbf{I}}$ is defined as

$$Q_{\mathbf{I}}(u,v)=\bar{P}_{\mathbf{I}}(u,v)=1-P_{\mathbf{I}}(u,v) \tag{6}$$

After each regularization period, the image energy change in the primitive and non-primitive fields are calculated as

$$\begin{aligned}\Delta E_{P,t}&=\sum_u\sum_v|\Delta\mathbf{I}_t(u,v)|^2\ P_{\mathbf{I}}(u,v)\\ \Delta E_{Q,t}&=\sum_u\sum_v|\Delta\mathbf{I}_t(u,v)|^2\ Q_{\mathbf{I}}(u,v)\end{aligned} \tag{7}$$



(a) (b)

Fig. 2. Regularization degree control

3.3. Regularization degree control

Due to the primitive-preserving property of PDE-based regularization, the energy change in $Q_\mathbf{I}$ is more intensive than that in $P_\mathbf{I}$ in the beginning stage of regularization. On the other hand, there is a large probability that compression artifacts such as ringing and blocking appear in $Q_\mathbf{I}$ when ρ is small. So regularization removes compression artifacts first. However, since the variation in $Q_\mathbf{I}$ decreases faster than that in $P_\mathbf{I}$, the energy change ratio

$$r_t = \frac{\Delta E_{P,t}}{\Delta E_{Q,t}} \tag{8}$$

will increase with t. When regularization is performed to a certain degree, PDE loses its efficacy in distinguishing primitives from quantization noise, and r_t will then remain at a stable level. Fig. 2(a) shows several $r_t - t$ curves for example.

In practice, we stop regularization at the T-th period when

$$r_T - r_{T-1} \le \varepsilon R, \quad R = \max_t (r_t - r_{t-1}), \quad 0 < \varepsilon < 1 \tag{9}$$

where ε is a constant and R represents the fastest increasing speed of r_t. Fig. 2(b) gives the distribution of T for four test web cartoon videos (randomly downloaded from YouTube [7], each with 1500 frames). The minimum T is 2 and the maximum T is set to 15. It can be observed that most of the frames require less than 15 regularization times, which validates the convergence property of r_t.

3.4. Inter-frame interaction

The above regularization degree control can adapt to the degradation grade within a single image/frame. On the other hand, in a video sequence, compression artifacts in consecutive frames can vary much due to fast motion or scene switch. An adaptive regularization should also take the inter-frame interaction into consideration. For two frames with similar contents, the one with heavier quantization noise requires larger regularization times, and this adaptivity is mainly reflected by the parameter R. One can easily find in Fig. 2(a) that frames with heavier quantization noise have smaller R (refer to Fig. 3 for the frames).

To further enhance the regularization adaptivity, we record R_{k-1} of the (k-1)-th frame and for the k-th frame, R_k is calculated as

$$R_k = \frac{R_{k0}^2}{R_{k-1}} \tag{10}$$

where R_{k0} is the initial quantity measured from the current frame. If $R_{k0} < R_{k-1}$, it suggests the quantization noise in the current frame is more severe than that in the previous frame. Then R_{k0} is further diminished to penalize the regularization times, and vice versa. In this way, the regularization degree can also adapt to the variable degradation grade between consecutive frames, making the quality improvement on the whole video sequence more stable.

4. CONSISTENT PRIMITIVE ENHANCEMENT

After the adaptive regularization, compression artifacts in each frame are effectively reduced while the primitive components are best preserved. The remaining primitives are then further enhanced with learning-based pair matching as implemented in [3] for images. (Also refer to [5], the detailed procedures are not presented here due to space limitations.) The main problem when applying this method to video is that, how to make the enhanced primitives temporally consistent to avoid flicker, especially for the sequence with slight motion? To solve this problem, we propose to optimize the compatibility in both the temporal and spatial domain.

In the temporal domain, we record the pair matching indices of primitive patches in the (k-1)-th frame. Then, for each primitive patch p_{ik} extracted from the location i of the k-th frame, it is first compared with the primitive patch $p_{ik\text{-}1}$ in the same position of the (k-1)-th frame (in case $p_{ik\text{-}1}$ exists). If p_{ik} is judged the same as $p_{ik\text{-}1}$, the matching index of $p_{ik\text{-}1}$ is directly assigned to p_{ik}; or else pair matching for p_{ik} is conducted in the database.

In the spatial domain, some of the pair matching indices come from the previous frame and others from the current frame, so the compatibility of learned enhancing patches should be optimized. Specifically, for each p_{ik} whose enhancing patch p_{ik}^* will be generated from the current frame, we take M matching results $\{p_{ik,1}^*, p_{ik,2}^*, \ldots p_{ik,M}^*\}$ as candidates for p_{ik}^* and the optimum one is found as

$$\hat{m} = \arg\min_{1 \le m \le M} d(p_{ik,m}^*, p_{jk}^*) \tag{11}$$

where p_{jk}^* is the previous selected enhancing patch in raster-scan order, and function d measures the difference in the overlapped region of two patches. In brief, the enhancing patch corresponding to p_{ik} can be denoted as

$$p_{ik}^* = \begin{cases} p_{ik-1}^*, & p_{ik} = p_{ik-1} \\ p_{ik,\hat{m}}^*, & \text{else} \end{cases} \tag{12}$$

Finally, the primitive enhancing image is generated by assembling all enhancing patches, where pixel values in the overlapped regions are averaged.

5. EXPERIMENTAL RESULTS

We test our solution on a variety of web cartoon videos downloaded from YouTube. They are generally in a decreased 320×240 resolution but with different compression degrees. We perform a uniform 3× super-resolution on them using a 10M database. The database is trained from two 1536×1024 resolution images. In the regularization stage, the PDE updating step λ is set to 5.0, the neighborhood order of the primitive field ρ is 2, and ε is 0.25 for regularization degree control. In the enhancement stage, the candidate number of the enhancing patch M is 16, and the pair matching process is speed up by the ANN tree searching algorithm [6]. The overall complexity of our solution is relatively low, with averagely less than 1 second per frame on an Intel Xeon 2.33 GHz CPU.

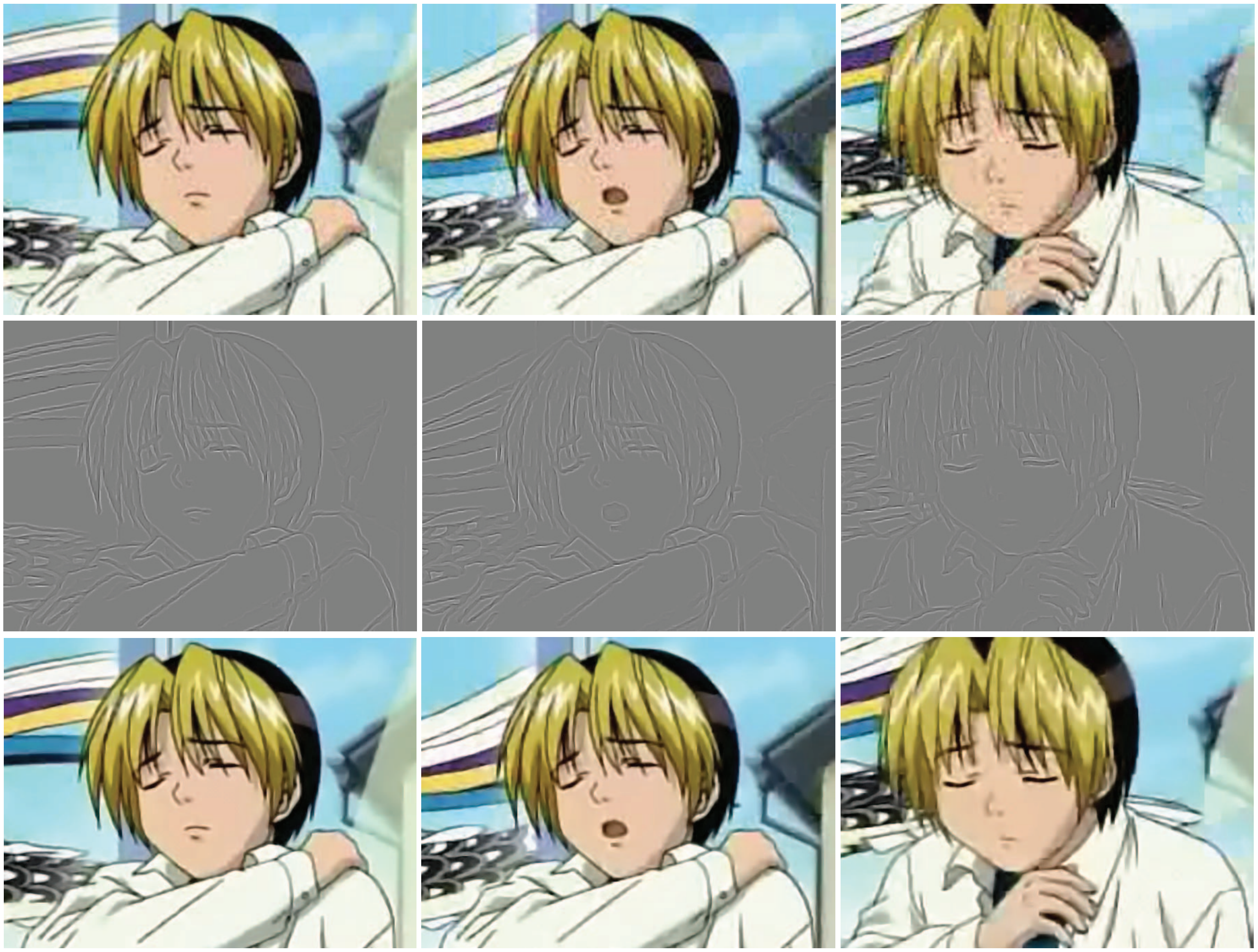


Fig. 3. Super-resolution of a web cartoon video. Top: bicubic upsampling, middle: primitive enhancing images; bottom: our results.

Fig. 3 shows three frames extracted from a super-resolved web cartoon video (sequence 3 in Fig. 2). This result demonstrates the effectiveness of our solution in three aspects. First, the regularization times, as enclosed in the caption, are appropriately dependent on the degradation grade of each frame. Second, the primitive enhancing images preserve both temporal and spatial consistency due to compatibility optimization. Last, the combination of adaptive regularization and primitive enhancement steadily improve the perceptual quality of directly interpolated videos, even when severe compression artifacts and fast motions are presented. (Please see the electronic version for better visualization.)

6. CONCLUSION

This paper proposes a practical solution for the super-resolution of web cartoon videos. An adaptive PDE regularization in combination with learning-based primitive enhancement greatly improves the perceptual quality of video sequences with variable degradation grade. In addition, temporal consistency is enforced by inter-frame interaction and compatibility optimization. Besides cartoons, other types of web videos will be considered in our future work.

REFERENCES

[1] B. K. Gunturk, Y. Altunbasak, and R. Mersereau, "Bayesian resolution enhancement framework for transform-coded video," in *ICIP* 2001, pp. 41-44.

[2] C. A. Segall, A. K. Katsaggelos, R. Molina, and J. Mateos, "Bayesian resolution enhancement of compressed video," *IEEE Trans. Image Process*., vol. 13, no. 7, pp. 898–911, July 2004.

[3] Z. Xiong, X. Sun and F. Wu, "Super-resolution for low quality thumbnail images," in *ICME* 2008, pp. 181-184.

[4] D. Tschumperlé and R. Deriche, "Vector-valued image regularization with PDE's: a common framework for different applications", in *CVPR* 2003, pp. 651-656.

[5] J. Sun, N. Zheng, H. Tao, and H. Shum, "Image hallucination with primal sketch priors," in *CVPR* 2003, pp. 729–736.

[6] D. Mount and S. Arya. Ann: Library for approximate nearest neighbor searching. *http://www.cs.umd.edu/ mount/ANN/.*

[7] *http://youtube.com.*