GENERIC VIDEO CODING WITH ABSTRACTION AND DETAIL COMPLETION

Zhe Yuan[1]*, Hongkai Xiong*[1,2]*, Li Song, and Yuan F. Zheng*[1,3]

[1] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, P.R. China
[2] Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Dept. of Electrical and Computer Engineering, Ohio State University, Columbus, OH 43210 USA

ABSTRACT

This paper presents a generic video coding framework with the texture abstraction and completion, inspired by a strong grouping bias of local elements in Gestalt psychology. Abstracting imagery by grouping perceptual salience from anisotropic diffusion, it decomposes video images into two layers composing of semantic components and residual detail. The similarity between textures of abstraction layer is motivated to infer the restoration of missing detail, under the spatio-temporal variation regularity. Through a motion and spatial context of moton, hence, a group of pictures (GOP) is divided into key frames and abstracted frames to form the final compressed data. An abstraction refinement is tuned to improve matching of detail restoration based on bilateral filtering. The proposed approach is more generic without incurring any specific side information, and achieves up to 20% bit saving versus standard H.264 at similar visual quality levels.

Index Terms—Video abstraction, perceptual video coding, bilateral filtering, min-cut, texture synthesis.

1. INTRODUCTION

The state-of-the-art H.264 compression engine has achieved a vital efficiency by exploiting pixel-wise statistical redundancy. Hereafter, perceptual or visual redundancy is investigated to improve the current performance of hybrid video coding. Inspired by advances in computer vision, image statistics is extended from correlations of pixel intensities to visual modeling by Zhu [1]. Zhu further considered natural images to consist of an overwhelming number of visual patterns generated by very diverse stochastic processes in nature. A pixel is an instance of a visual pattern, and a visual pattern is equalized by a set of features with statistical models (e.g. descriptive model, generative model, etc.).

From a hierarchical perspective of visual representation, images are generally decomposed into texture and piecewise smooth parts called cartoon (e.g. object hues and sharp edges like boundaries). Texture as homogeneous visual patterns, facilitates texture synthesis with inherent repeatability [2]. Image inpainting likewise aims to fill-in small missing region with high structure in a visually plausible way [3]. By edge-based side information, a compression-oriented image inpainting framework was herein proposed for restoration [4-5]. A space-time completion algorithm was posed as a global optimization problem with global tempo-spatial consistency [6]. Thereby, given corresponding side information of detail-irrelevant texture regions (e.g. water, smoke, sand, etc.), a closed-loop analysis-synthesis video coding approach was proposed [7]. In natural video sequences, however, there exist not only stochastic pixel intensity models (e.g., scaling, rotation, luminous change and local motion), but also attributes (e.g. combination of different texture or textons). Although we may detect semantically homogeneous parts, it will actually be much more difficult to extract a proper candidate patch for video reconstruction. The coding burden from the side information is also a critical issue for generic video coding.

The human visual perception from Gestalt psychology suggests that there is a strong bias toward forming global percept by grouping local elements [8]. It shows that an observer would first abstract the input and combine the topological information to a whole idea, while the detail of the objects is rather obscure. This paper is motivated to present a generic video coding framework with the texture abstraction and completion. Abstracting imagery by grouping perceptual salience with anisotropic diffusion, it decomposes video images into two layers composing of semantic components and residual detail. The similarity between textures of abstraction layer is used to infer the restoration of missing detail, under the spatio-temporal variation regularity. Through a motion and spatial context of moton, hence, a group of pictures (GOP) is divided into key frames and abstracted frames to form the final compressed data. An abstraction refinement is tuned to improve matching of detail restoration from bilateral filtering. The proposed approach is more generic without incurring any specific side information, and achieves up to 20% bit saving in contrast with standard H.264 at similar visual quality.

2. FRAMEWORK OF THE PROPOSED SCHEME

The diagram of the proposed framework is depicted in Fig. 1, where it exploits motion and spatial context as a cue for layer separation. Given a input video sequence

The work has been partially supported by the NSFC grants No. 60632040, No. 60772099, No. 60702044 and the National High Technology Research and Development Program of China (863 Program) (No. 2006AA01Z322).

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

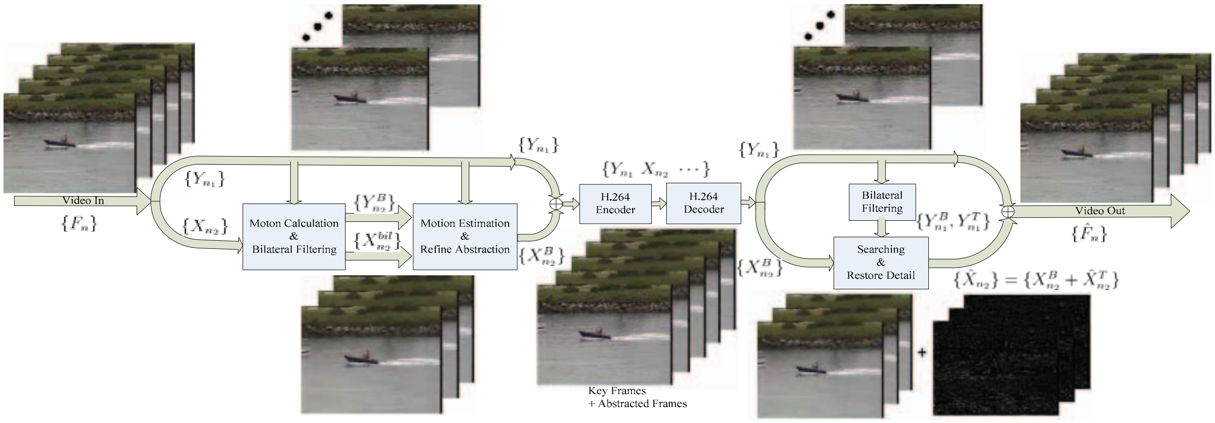

Video In
$\{F_n\}$
$\{Y_{n_1}\}$
$\{X_{n_2}\}$
Moton Calculation & Bilateral Filtering
$\{Y^B_{n_2}\}$
$\{X^{bil}_{n_2}\}$
Motion Estimation & Refine Abstraction
$\{Y_{n_1}\}$
$\{X^B_{n_2}\}$
$\{Y_{n_1}\ X_{n_2}\ \cdots\}$
H.264 Encoder
H.264 Decoder
Key Frames + Abstracted Frames
$\{Y_{n_1}\}$
$\{X^B_{n_2}\}$
Bilateral Filtering
$\{Y^B_{n_1}, Y^T_{n_1}\}$
Searching & Restore Detail
$\{\hat{X}_{n_2}\} = \{X^B_{n_2} + \hat{X}^T_{n_2}\}$
+
Video Out
$\{\hat{F}_n\}$

Fig. 1. The overview of the generic video coding framework with abstraction and detail completion.

$(F_1, F_2 \cdots F_n \cdots)$, it is supposed to decompose a picture as two layers F^B_n and F^T_n : $F_n = F^B_n + F^T_n$. F^B_n is abstracted imagery by grouping perceptual salience from anisotropic diffusion, preserving semantic components such as structure, location and color of objects. F^T_n as residual detail may refine the scale and resolution of homogeneous patterns. By a tempo-spatial consistency measure, a group of pictures (GOP) $\{F_n\}$ is divided into key frames $\{Y_{n_1}\}$ and abstracted frames $\{X_{n_2}\}$ to form the final compressed data. The detail reconstruction of $\{X^T_{n_2}\}$ in the decoder will be generalized in terms of energy minimization. The goal is to find a restoration $\{\hat{X}^T_{n_2}\}$ that are piecewise smooth and consistent with the observed data (the neighboring key frames).

$$\hat{X}_{n_2} = X^B_{n_2} + \hat{X}^T_{n_2}, \quad \text{s.t.} \ \underset{\hat{X}^T_{n_2}}{\arg\min} \, ||E_{data}(\hat{X}^T_{n_2}) + E_{smooth}(\hat{X}^T_{n_2})||_p \tag{1}$$

where E_{data} measures the disagreement between $\{\hat{X}^T_{n_2}\}$ and the observed data, while E_{smooth} measures the extent to which $\{\hat{X}^T_{n_2}\}$ is not piecewise smooth.

It is equivalent to a learning-based optimization problem from the given set of sparse data. Such a problem is ill-posed as there exists an infinity of functions that pass through the data $\{X^B_{n_2}\}$. The common way with regard to regularization theory is by means of a stabilizer assuming that the function presents some intrinsic properties, e.g. smoothness. It induces the underlying problem in Eq. (1) of finding the function that minimizes the functional combination of the empirical convex loss and prior information, associating with different approximation on the balance between fitness and prior constraints.

Let $X^i_{n_2}$ is the block indexed by position i, and $Y^j_{n_1}$ is a block in a key frame. We attempt to use the similarity between abstraction layer $X^{B,i}_{n_2}$ and $Y^{B,j}_{n_1}$ to infer the missing detail $X^{T,i}_{n_2}$, under the spatio-temporal variation regularity:

$$\hat{X}^i_{n_2} = X^{B,i}_{n_2} + Y^j_{n_1}, \quad \text{s.t.} \ \underset{\{n_1, j\}}{\arg\min} \int_{\cdots F_n \cdots} ||X^{B,i}_{n_2} - Y^{B,j}_{n_1}||_p d\mathbf{S} + \lambda | \bigtriangledown \hat{X}^i_{n_2} | d\mathbf{S} dt \tag{2}$$

At the decoder, we might treat decoded abstracted frame $X^B_{n_2}$ as an entry to search a proper patch in a dictionary of detail information $\{Y^T_{n_2}\}$ for restoration.

3. PROPOSED METHOD

3.1. Spatio-temporal consistency measurement

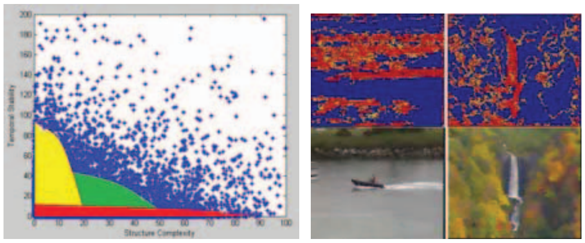
To sample the abstracted pictures of high correlation within a GOP, restrictions on a spatio-temporal feature called "moton" are imposed to simultaneously capture motion and spatial context [9]. The temporal derivatives of block i in frame F_n is denoted as $\dot{\mathbf{z}}^i_n = ||G(F^i_n) - G(F^j_{ref})||_2$, where $G(\cdot)$ is a Gaussian kernel at the scale of σ_n pixels, and F^j_{ref} is the block indexed j in the reference to match F^i_n with MAD criteria. Note that F_{ref} is a block-wise motion-compensation prediction frame from neighboring key frames. Also, we add a parameter v^i_n as variance of temporal derivatives: $\mathbf{v}^i_n = \mathrm{Var}(\dot{\mathbf{z}}^i_n)$ which describes the temporal stability of visual pattern. The spatial gradients $\mathbf{g}^i_n = | \bigtriangledown F^i_n |$, is computed by convolving with first-order derivative of Gaussian kernels (DOG) with standard deviation σ_s. The combination of temporal derivatives and spatial gradients will form a 2-D feature "moton", which has attributes for the motion consistency and structure complexity of the visual patterns.

$$\mathbf{M}^i_n = \left(\mathbf{g}^i_n, \ \dot{\mathbf{z}}^i_n \cdot \log_e(1 + \mathbf{v}^i_n)\right) \tag{3}$$

Similar to textons, the 2D vectors $\{\mathbf{M}^i_n\}$ are constructed into N clusters via Expectation Maximization for a training set. It enables efficient indexing and correlation of the joint space between structure and motion. Fig. 2(a) shows a sampling distribution with $N = 4$ motons from "coastguard" sequence. The stationary content is labeled with red color through the video sequence. The areas colored yellow and green, respectively, identify weak texture and strong texture with little motion, and they are supposed to incur detail restoration. The remaining highly structured and non-rigid texture with blue color, is approximately attained by edge-preserving filtering of video abstraction.

3.2. Video Abstraction

3.2.1 Bilateral Filtering



(a) (b)

Fig. 2. (a) A sampling distribution of "moton" in “Coastguard” sequence; (b) top: classification (blue for temporally stable and rigid texture); bottom: corresponding abstraction result (“coastguard” frame 3, “waterfall” frame 5).

The video abstraction aims to simplify the visual patterns, while preserve or even emphasize most of the perceptually salient features. Anisotropic diffusion could prevent averaging across edges, while still averaging within smooth regions. The bilateral filtering combining domain and range filtering for edge-preserving smoothing, serves as a link between the extended nonlinear diffusion and mean shift filtering [10].

For each pixel denoted as 2D coordinates vector $\mathbf{p}$, a continuous version of Gaussian bilateral filtering can be written as $F^{bil}(\mathbf{p}) = \frac{\int c(\xi, p)s(\xi, \mathbf{p})F(\xi)d\xi}{\int c(\xi, \mathbf{p})s(\xi, \mathbf{p})d\xi}$. The convolution mask is the product of the functions $c(\cdot)$ and $s(\cdot)$, which represent closeness (domain) and similarity (range), respectively. Within a window of size S, the discrete expression of abstracted frames is:

$$X_{n_2}^{bil}(\mathbf{p}) = \frac{\sum_{\xi_1=-S}^{+S}\sum_{\xi_2=-S}^{+S} X_{n_2}(p_1+\xi_1, p_2+\xi_2)w(\mathbf{p}, \xi)}{\sum_{\xi_1=-S}^{+S}\sum_{\xi_2=-S}^{+S} w(\mathbf{p}, \xi)} \tag{4}$$

where $\xi = (\xi_1, \xi_2)$ is the offset of neighboring pixels, and the weight w is given by:

$$w(\mathbf{p}, \xi) = \exp(\frac{-\|\xi\|_2}{2\sigma_D^2})\exp(\frac{-(X_{n_2}(\xi+\mathbf{p}) - X_{n_2}(\mathbf{p}))^2}{2\sigma_R^2}) \tag{5}$$

The filtered output $Y_{n_1}^{B}$ of a key frame is likewise computed, as references for searching at both encoder and decoder sides. The final abstraction $X_{n_2}^{B}$ of abstracted frames will be further refined based on both $X_{n_2}^{bil}$ and references.

3.2.2 Abstraction Refinement

With the abstraction $X_{n_2}^{bil}$ from bilateral filtering, it is anticipated to improve matching between a refined abstraction $X_{n_2}^{B}$ and the reference during detail restoration.

Assume the reference Y_{ref}^{B} which is obtained as a combination of abstracted blocks from neighboring key frames. The refined abstraction block $X_{n_2}^{B,i}$ is supposed to have a matching block $Y_{ref}^{B,j}$ according to Eq. (2), which is tuned by minimizing all the differences of abstraction residue and corresponding detail residue ($X_{n_2}^{T,i}$ and $Y_{ref}^{T,j}$):

$$\begin{aligned} X_{n_2}^{B,i} = \underset{\mathbf{p}\in X_{n_2}^{B,i}}{\arg\min}\ & \lambda_1 \left\| X_{n_2}^{B,i} - Y_{ref}^{B,j} \right\|_2 \\ & +\lambda_2 \left\| (X_{n_2}^{i} - X_{n_2}^{B,i}) - (Y_{ref}^{j} - Y_{ref}^{B,j}) \right\|_2 + \lambda_3 \left\| X_{n_2}^{B,i} - X_{n_2}^{bil,i} \right\|_2 \end{aligned} \tag{6}$$

$$\triangleq \underset{\mathbf{p}\in X_{n_2}^{B,i}}{\arg\min}\ \lambda_1 \left\| X_{n_2}^{B,i} - A_1 \right\|_2 + \lambda_2 \left\| X_{n_2}^{B,i} - A_2 \right\|_2 + \lambda_3 \left\| X_{n_2}^{B,i} - A_3 \right\|_2 \tag{7}$$

$A_1 = Y_{ref}^{B,j}$, $A_2 = X_{n_2}^{i} + Y_{ref}^{B,j} - Y_{ref}^{j}$, $A_3 = X_{n_2}^{bil,i}$, and it could be simplified as: $X_{n_2}^{B,i} \approx (\lambda_1 A_1 + \lambda_2 A_2 + \lambda_3 A_3)/(\lambda_1 + \lambda_2 + \lambda_3)$.

With different choices on the value of $\{\lambda_i\}$, we can control abstraction state based on the spatial-temporal consistency. For example, the second item can be emphasized in stable and rigid areas; and the first two items are set to zero for regions containing structure. Fig.2 (b) shows the abstraction result of two video sequences, which can be seen that the amount of detail is greatly reduced, while preserving the boundary and structure of objects.

3.3. Detail Reconstruction

As shown in Fig.1, it is required to restore the detail $\{X_{n_2}^{T}\}$ omitted at the encoder. The operations in both encoder and decoder sides are, respectively, concluded as follows. Once obtaining the matching detail $Y_{ref}^{T,j}$ as an proper estimation for $\hat{X}_{n_2}^{T,i}$, we will recover the video combining with $X_{n_2}^{B,i}$. Here, an importance map based on the remaining edge information will be used to restore regions with more reliable and important information with high priority, followed by homogeneous area.

In view of block-wise reconstruction, partial overlap and image quilting [11] are introduced to reduce annoying artifacts. It aims to achieve an optimal seam by minimizing the cumulative error energy as:

$$\varepsilon = \sum_{\mathbf{p}\in overlap} \left\| \hat{X}_{n_2}^{i}(\mathbf{p}) - \hat{X}_{n_2}^{j}(\mathbf{p}) \right\|_2, \{i, j\} \in Neighbour \tag{8}$$

The operations of the proposed scheme are concluded as:

```
Encoder side:
1. Determine the groups to abstracted frames $\{X_{n_2}\}$ and key frames $\{Y_{n_1}\}$;
2. For one frame in $\{X_{n_2}\}$
   2.1 Calculate spatial-temporal consistency $\mathbf{M}_n^i$.
   2.2 Abstract using bilateral filtering to get $X_{n_2}^{bil}$.
   2.3 Refine the abstraction to get $X_{n_2}^{B}$.
3. Arrange all the frames into a sequence.
Decoder side:
1. Determine the group of frames that needs to be reconstructed $\{X_{n_2}^{B}\}$ and select reference $\{Y_{ref}^{B}\}$;
2. For one frame in the current $\{X_{n_2}^{B}\}$
   2.1 Determine the reconstruction order.
   2.2 For current block $X_{n_2}^{B,i}$, search in the reference frames $Y_{ref}^{B}$ in order to find a patch satisfies Eq. (2).
   2.3 Recover the block by the estimated detail $\hat{X}_{n_2}^{T,i}$ and $X_{n_2}^{B,i}$.
   2.4 Allow overlap between blocks and using image quilting to find an optimal seam according to Eq. (8). Then go to step 2.2.
3. Repeat step 2 for other groups of frames.
```

4. EXPERIMENTAL RESULT

We evaluate the proposed approach in H.264 engine for versatile video sequences. It enables two B frames, and CABAC entropy coding with a frame rate of 25 fps, CIF (352x288) resolution, and a GOP size of 9 frames. Rate control over quantization parameter (QP) is enabled to

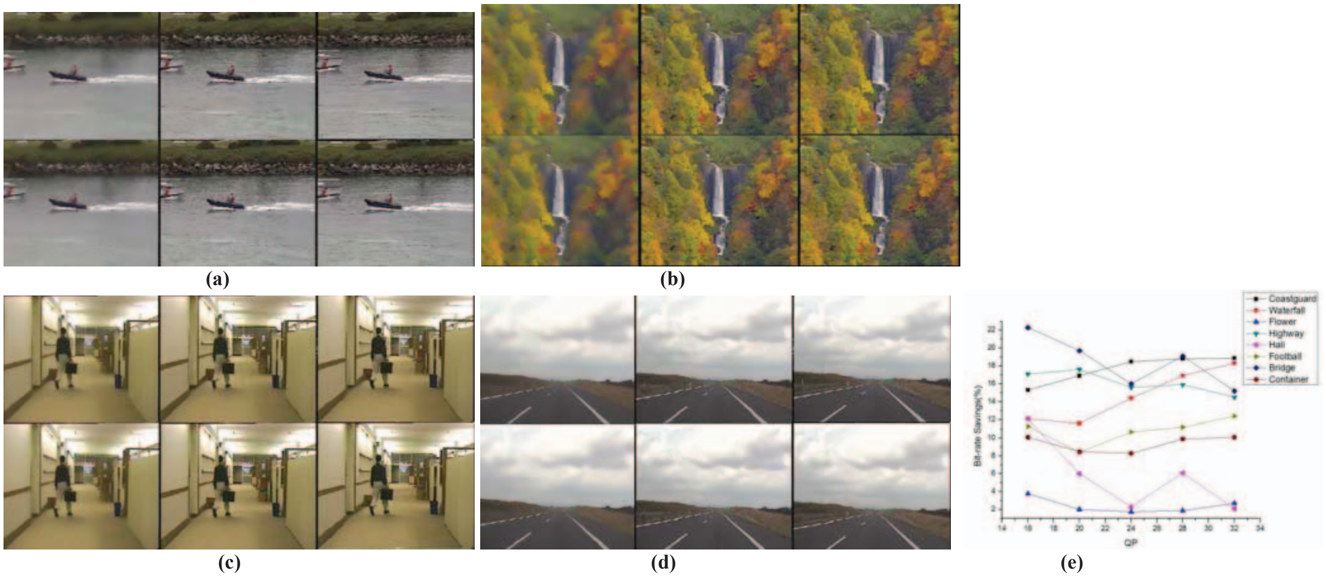


Fig. 4. (a) The 12th frame from "coastguard", using frame 9 and 19 as reference. Top: QP=28; Bottom: QP=16; (b) 14th frame from "waterfall", using frame 9 and 19 with QP=28, 20; (c) 48th frame from "Hall", using frame 45 and 55 with QP=32, 20; (d) 12th frame from "Highway", using frame 11 and 19 with QP=28, 20; (e) Coding performance (bit-rate savings) w.r.t. QP versus H.264.

maintain comparable subjective visual quality with traditional H.264 standard. The comparison effects can be found in Fig. 4 (a) ~ (d), where the left column displays the abstracted frame; the middle our reconstructed frame, and the right the standard decoded frame by H.264. Moreover, the generic coding performance of eight video sequences is fully investigated for different QP in Fig. 4 (e), where the bit-rate saving (%) versus standard H.264 is used. It shows that up to 20% bit saving is achieved at the similar visual quality levels. Within common QP interval [16, 32], "Hall" saves 2.1~12.1 %, "waterfall" 12.1~18.3%, "coastguard" 15.3~18.9% bits and "highway" 14.5~17.1 %. Table I evaluates the coding fluctuation under different GOP sizes, too. It proves the proposed scheme generic for all kinds of configurations. Also, it can be seen that the proposed scheme performs well for dynamic weak texture, e.g. "coastguard" and "waterfall", because the textures involving local motion are difficult to code using MSE-based ME-MC.

TABLE I Bit-rate savings w.r.t GOP size versus H.264.

GOP Size Sequence	6	9	12	15
Coastguard	19.9	18.8	23.7	24.7
Waterfall	18.6	16.9	17.4	17.5
Highway	9.6	15.9	10.6	13.1
Bridge	21.2	20.1	19.7	22.3

5. CONCLUSION

From a global percept by grouping local elements from Gestalt psychology, we present a generic video coding framework with texture abstraction and completion. It abstracts imagery by grouping perceptual salience from anisotropic diffusion, and decomposes images into two layers composing of semantic components and residual detail. The similarity between textures of abstraction layer is used to infer the restoration of missing detail, under the spatio-temporal variation regularity. Through motion and spatial context, a group of pictures is divided into key frames and abstracted frames to form the compressed data. The proposed approach does not incur any specific side information like in [7], and achieves a distinctive bit-rate savings at similar visual quality levels.

6. REFERENCES

[1] S.-C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans. Pattern. Analysis and Machine Intelligence*, vol. 25, no. 6, Jun. 2003.

[2] V. Kwatra, A. Schodl, I. Essa, et al., "Graphcut textures: image and video synthesis using graph cuts," *SIGGRAPH*, Jul. 2003.

[3] M. Bertalmio, G. Sapiro, V. Casselles and C. Ballester, "Image Inpainting", *SIGGRAPH*, pp.417-424, 2000.

[4] C. Wang, X. Sun, F. Wu and H. Xiong, "Image compression with structure-aware inpainting," Proc. of *IEEE Symposium on Circuits and Systems*, Greece, pp. 21-24, 2006.

[5] D. Liu, X. Sun, F. Wu, et al., "Image compression with edge-based inpainting," *IEEE Trans. Circuits and Systems for Video Technology*, vol.17, no. 10, Oct. 2007.

[6] Y. Wexler, E. Shechtman and M. Irani, "Space-time completion of video", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 463-476, 2007.

[7] P. Ndjiki-Nya, T. Hinz, and T. Wiegand, "Generic and robust video coding with texture analysis and synthesis," *IEEE International Conference on Multimedia*, Jul. 2007.

[8] T. Leung and J. Malik. "Representing and recognizing the visual appearance of materials using three-dimensional textons," *IJCV*, vol. 43, no. 1, 2001.

[9] P. Yin, A. Criminisi, J. Winn, I. Essa, "Tree-based Classifiers for Bilayer Video Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2007.

[10] N. Sochen, R. Kimmel, and A.M. Bruckstein, "Diffusions and Confusions in Signal and Image Processing,"*Journal of Mathematical Imaging and Vision*, vol. 14, no. 3, 2001.

[11] A. Efros and W. Freeman, "Image Quilting for Texture Synthesis and Transfer," Proceedings of SIGGRAPH, California, Aug. 2001.