Multi-Regions Texture Substitution for Image and Video

Jinliang Wan
College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, China
E-mail: wanqiq_2006@sohu.com

Yanhui Liu
State Key Laboratory for Seismic Reduction/Control & Structural Safety (Cultivation), Guangzhou University, Guangzhou, China
Email: Liuyanhui19@163.com

Abstract—Perceptual coding should take full advantage of the results from human visual system studies. Image can be parsed into two main categories of representation: contour curve and region texture. In this paper, we introduce a multi-regions texture substitution (MRTS) algorithm for images and videos. MRTS algorithm can achieve medium-level-visual texture substitution through selecting the smallest region texture sample. Specifically combining with dynamic texture learning and synthesis techniques, we explore it in 2D and 3D to perform multi-regions texture substitution for video sequences. Experimental results confirm MRTS algorithm can be done successfully for images with massive textures. Compared with JPEG2000 compression, MRTS can not only get more coding gain but remain most image information people care. MRTS algorithm also can get good video reconstruction quality.

Index Terms—human visual system; texture substitution; multi-regions; dynamic texture

I. INTRODUCTION

The mainstream of video coding scheme is based on hybrid coding framework integrated with three classes of traditional technologies. Transform coding, prediction coding and entropy coding are used to reduce and remove spatial redundancy, temporal redundancy and statistic redundancy. But human eyes are the terminal receptors for images and videos, and there still exist a lot of perceptual redundancy redundancies according to the specialty of human visual system. However, it is difficult to achieve additional significant progress for compression efficiency because of the limitation of pixel-based image and video representation [1]. Therefore, how to find a novel image and video representation method becomes a hot topic.

Actually speaking, many images contain massive textures like grass, sand, brick, etc., which compose the still backgrounds. In 2003, Ndjiki-Nya [2] assumed that the textures in an image could be classified into two categories: textures with unimportant subjective details and the remainder. In 2008, Sun [3] utilized this idea and considered textures as unimportant parts in terms of the imperfections inherent to the Human Visual System (HVS). Details of textures in these parts can not generally draw our attentions. These parts can be considered as detail-irrelevant textures if the original image is not shown to the viewer. As a result, changes of texture details will not affect the subjective comprehension toward the original texture. It indicates that vision character can be utilized to remove perceptual redundancy and represent images effectively and efficiently.

In addition, Dynamic textures are a kind of visual motion pattern. They can be described as a spatially varying repetitive, time-varying visual pattern that forms an image sequence with certain temporal stationarity. Examples of dynamic textures are flames, grass, smoke, water etc. The underlying local motion in dynamic texture sequences is often governed by a complex stochastic model. Dynamic texture learning and synthesis techniques aim to learn the non-deterministic temporal and spatial variation in sequences of images.

In recent years, aiming at dynamic texture analysis, learning and synthesis, some parametric models have come into being. Szummer [5] proposed a spatial temporal autoregressive (STAR) model to recognize textures. Bar-Joseph [6] proposed a multi-resolution analysis (MRA) model to synthesize 1D sound textures, 2D texture images and 3D texture movies. Doretto [7] modeled dynamic textures as Linear Dynamic System (LDS) that has been applied to recognition, compression and segmentation. Above methods have two disadvantages: (1) the synthesis quality of complex dynamic textures is not gratifying enough, since the model is linear and stable, (2) computational expense is significant, since the raw sequence is applied as observation sequence directly. To improve visual quality, Liu [8] modeled dynamic textures using subspace mixtures. To reduce computation complexity of the learning process, Abraham [10] proposed to identify LDS using a set of Fourier descriptors of the frames rather than the raw sequence. In addition, Schödl [9] introduced video texture theory which has qualities somewhere between those of a photograph and a video.

Therefore, in this paper, we propose an effective multi-regions texture substitution (MRTS) algorithm for image...
and video, which can simply and reliably represent images or frames using regional contour structure and texture information. Because number of regions is variable and many, the MRTS algorithm can select the smallest region texture sample and obtain optimal restoration images and videos. Essentially, the MRTS algorithm can be regarded as a new, simple and reliable method for dynamic background reconstruction and achieve different dynamic background of arbitrary length. This paper is organized as follows. In Sect. 2, we explained the multi-regions texture substitution for image. The implementation of multi-regions video texture substitution will be detailed described in Sect. 3. Sect. 4 shows experiment results and analysis. Finally, we give the conclusion in Sect. 5.

II. MULTI-REGIONS TEXTURE SUBSTITUTION FOR IMAGE

Figure 1. Architecture of MRTS for image. The region shape feature representation, region texture sample selection and multi-regions image reconstruction should be repeated several times until all regions are restored successfully.

The architecture of MRTS for image is shown in Fig.1. The implementing procedure of this approach consists of four stages: multi-regions extraction, region shape feature representation, region texture sample selection and multi-regions image reconstruction.

A. Multi-Regions Extraction

An original image is first segmented into some regions in terms of homogeneous color and texture. The segmentation method is the so-called JSEG algorithm [11, 12]. Considering that texture sample selection is sensitive to segmentation results, we merge connected small regions if they are nearby in space so that bigger regions are resulted for each image [4]. And, the number of regions of the same image may be different. Through multi-regions extraction, every region obtains a unique index number. Every pixel belonged to one region is assigned the same region index number. These index numbers play an important role in region shape feature representation.

B. Region Shape Feature Representation

Shape and texture are salient features of the appearance of objects in natural scenes. The mid-level region is image representing and reconstructing unit which contains shape and texture characteristic.

The implementing procedure of region shape feature representation is described as the following steps:

- Contour tracking. An ordered sequence of contour points approximating the regional shape is extracted. For this purpose, a contour tracing algorithm [13] is applied to the input binary image.

- Down sampling. A vector composed of equally distributed points along the regional contour is extracted from the ordered sequence of contour pixels. Contour points are extracted in such a way that they will be approximately equally distributed along the regional contour curve.

- Piecewise iterative curve fitting. It [3, 14] combined with top-down split algorithm and curve-fitting principle is implemented to get the polynomial coefficients used to represent regional shape feature.

C. Region Texture Sample Selection

In one region, a big texture sample contains a majority of local and global characteristics of the inner texture. The inner texture is analyzed by a 2D-autocorrelation statistic analysis method [17]. The autocorrelation function of an image can be used to assess the amount of regularity as well as the fineness/coarseness of the texture present in the image. The autocorrelation function of an image is defined as follows:

\[ \rho(x, y) = \frac{MN}{(M-x)(N-y)} \sum_{u=1}^{M-x} \sum_{v=1}^{N-y} (I(u, v)I(u+x, v+y) - \frac{\sum_{u=1}^{M-x} \sum_{v=1}^{N-y} I^2(u, v)}{MN}) \]

For regular structural textures, the function will exhibit peaks and valleys. We can easily determine the scale of the texture primitives from the given texture. The exemplar should contain 2–5 texture primitives. For stochastic textures, 90% above values of the function are bigger than 0.9 generally. Correspondingly, the size of exemplar may be set smaller.

D. Multi-Regions Image Reconstruction

The implementing procedure of multi-regions image reconstruction is described as the following stages:

- Reconstruct region contour. Contour of one region can be described as a polynomial function. In order to make the connection between two neighboring curves perfect and guarantee the contour closed, a generic algorithm of dilation is applied.

- Synthesize region texture. Texture synthesis is performed in the manner of Efros and Freeman’s image quilting [15]. Many implementation details are introduced in [16]. For instance, a bigger patch means better capturing of texture characteristics in the texture patches and thus more similarity between the original texture and the synthesized one.

- Restore region. Synthetic region texture is filled to the reconstructed region contour.

Finally, multi-regions image reconstruction is achieved when all regions are successfully restored.
III. MULTI-REGIONS TEXTURE SUBSTITUTION FOR VIDEO

Fig.2 shows the architecture of MRTS for video. The multi-regions extraction, region dynamic texture learning and synthesis, multi-regions texture substitution and regions merge should be repeated several times until all frames are restored successfully.

Where, $\nu(t) \in \mathbb{R}^n$ is an IID realization from the density $q(\cdot)$; For some choice of matrices, $A_i \in \mathbb{R}^{m \times n}$, $i = 1, ..., k$; $B \in \mathbb{R}^{m \times n}$, and initial condition $x(0) = x_0$.

Without loss of generality, we can assume $k = 1$ since we can redefine the state of the above model $x(t)$ to be $[x(t)^T x(t-1)^T ... x(t-k)^T]^T$. Therefore, a linear dynamic texture is associated to an auto-regressive moving average (ARMA) model with unknown input distribution expressed by (3):

$$x(t+1) = Ax(t) + Bv(t)$$
$$y(t) = \phi(x(t)) + \omega(t)$$

(3)

Where, $x(0) = x_0$; $\nu(t) \sim q(\cdot)$ unknown and $\omega(t) \sim p_w(\cdot)$ given, such that $I(t) = \phi(x(t))$. The output of an ARMA model can represent the characterization of a dynamic texture. This model effectively captures a wide range of visual phenomena from which dynamic textures arise.

The definition of dynamic texture above entails a choice of filters $\phi_\alpha$, $\alpha = 1...n$. These filters are also inferred as part of the learning process for a given dynamic texture. There are several criteria for choosing a suitable class of filters, ranging from biological motivations to computational efficiency. In the simplest case, we can take $\phi$ to be the identity, and therefore look at the dynamics of individual pixels $x(t) = I(t)$ in (3).

We view the choice of filters as a dimensionality reduction step and seek for a decomposition of the image in the simple (linear) form shown as (4):

$$I(t) = \sum_{i=1}^{n} x_i(t) \theta_i = Cx(t)$$

(4)

Where, $C = [\theta_1, ..., \theta_n] \in \mathbb{R}^{m \times n}$ and $\{\theta_i\}$ can be an orthonormal basis of $L^2$, a set of principal components or a wavelet filter bank, for instance. An alternative nonlinear choice of filters can be obtained by processing the image with a filter bank and representing it as the collection of positions of the maximal response in the passband [18].

B. Region Dynamic Texture Learning

Given a sequence of noisy images $\{y(t)\}_{t=1}^{T}$, the dynamic texture learning amounts to identifying the model parameters $A, B, C$ and the distribution of the input $q(\cdot)$ in (3). This is a system identification problem [19] where one has to infer a dynamical model from a
time series. But in the literature of dynamical systems, it is commonly assumed that the distribution of the input is known. In the context of dynamic textures, we have the additional complication of having to infer the distribution of the input along with the dynamical model.

It is well known that a second-order stationary process with arbitrary covariance can be modeled as the output of a linear dynamical system driven by white, zero-mean Gaussian noise. With this choice of linear filters, the model representation can be written as (5):

\[
\begin{align*}
\dot{x}(t+1) &= A x(t) + B v(t) \\
y(t) &= C x(t) + o(t)
\end{align*}
\]

(5)

Where, \(x(0) = x_0\) is initial condition, \(x_t \in \mathbb{R}^n\); \(\{x(t)\}_{t=0}^{\tau}\) is a sequence of \(\tau\) images, \(x(t) \in \mathbb{R}^n\); \(v(t) \in N(0,Q)\) and \(o(t) \in N(0,R)\) ;symmetric positive-definite matrices \(Q \in \mathbb{R}^{n \times n}\) and \(R \in \mathbb{R}^{m \times m}\); For some Matrices \(A \in \mathbb{R}^{nxn}\) and \(C \in \mathbb{R}^{mxn}\). The problem of system identification consists in estimating the model parameters \(A, C, Q, R\) from the measurements \(y(1), \ldots, y(\tau)\). Note that parameters \(B\) and \(v(t)\) are such that \(BB^T = Q\) and \(v(t) \in N(0, I_{n_x})\), where \(I_{n_x}\) is the identity matrix of dimensions \(n_x \times n_x\).

The above definition of time-varying textures can easily be generalized to a non-linear model of the form \(x(t+1) = f(x(t), v(t))\) to capture the non-linear components of the dynamic texture behavior. With a non-linear model, learning the unknown parameters becomes more computationally expensive. In this paper, we use an efficient and approximate methods based on singular value decomposition to estimate parameters in (5). In the dynamic texture learning process, we reduce the dimensionality of image data in order to improve the learning speed.

C. Implementation Procedure

The implementation procedure of multi-regions texture substitution for video is as follows:

- Extract regions. Every frame of video sequence is divided into static regions and moving regions.
- Implement multi-regions texture substitution in those static regions.
- Implement learning and synthesis of dynamic texture in those moving regions. At the same time, with reference to H.264 compression coding theory, the static regions all are set to 0 so as to effectively improve the compression ratio and time efficiency before learning and synthesizing dynamic texture.
- Frame is successfully restored through merging reconstructed moving regions and static regions.

Finally, multi-regions texture substitution for video is achieved when all frames are successfully restored.

IV. EXPERIMENTAL RESULTS

We realize the proposed multi-regions texture substitution algorithm for image and video by C code, and test it on some natural images and videos with massive still textures and dynamic textures.

A. MRTS for Image

For multi-regions extraction, we specify three parameters as follows: \(TQUAN = -1\), \(NSCALE = -1\), \(threshcolor = 0.8\), which make the performance of homogeneity detection perfect. For the piecewise iterative curve fitting, the split threshold amounts to 5~10 and the degree of polynomial of curve fitting is specified as 4 generally. Here, the JPEG2000 version used is the Kakadu_V2.2.3. The general viewing conditions for subjective assessments in laboratory environment completely comply with the ITU-R BT.500-11 [20] standard. We assess visual quality of MRTS images using the double-stimulus impairment scale (DSIS) method. The 25 students selected are the observers.

There are reconstructed results with different number of regions in Fig.3. Images in the first row are original images, and images in the second row are the synthesized ones.

### TABLE I.

<table>
<thead>
<tr>
<th>Original Image</th>
<th>Fig3 (a)</th>
<th>Fig3 (b)</th>
<th>Fig3 (c)</th>
<th>Fig3 (d)</th>
<th>Fig3 (e)</th>
<th>Fig3 (f)</th>
<th>Fig3 (g)</th>
<th>Fig3 (h)</th>
<th>Fig3 (i)</th>
<th>Fig3 (j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>3.78</td>
<td>2.89</td>
<td>2.33</td>
<td>1.47</td>
<td>5.81</td>
<td>8.52</td>
<td>5.81</td>
<td>1.47</td>
<td>1.42</td>
<td>4.15</td>
</tr>
<tr>
<td>Bytes</td>
<td>30288</td>
<td>6782</td>
<td>29472</td>
<td>12527</td>
<td>9999</td>
<td>14481</td>
<td>11155</td>
<td>3964</td>
<td>7034</td>
<td></td>
</tr>
<tr>
<td>MRTS</td>
<td>1664</td>
<td>1398</td>
<td>1526</td>
<td>1248</td>
<td>7548</td>
<td>1344</td>
<td>3391</td>
<td>1760</td>
<td>3072</td>
<td>736</td>
</tr>
<tr>
<td>Bytes</td>
<td>15556</td>
<td>5060</td>
<td>8795</td>
<td>4935</td>
<td>5721</td>
<td>6806</td>
<td>5721</td>
<td>6806</td>
<td>5721</td>
<td>6806</td>
</tr>
<tr>
<td>MRTS/GS200000 (%)</td>
<td>30.24</td>
<td>74.94</td>
<td>39.64</td>
<td>39.81</td>
<td>61.02</td>
<td>45.92</td>
<td>34.93</td>
<td>55.7</td>
<td>75.23</td>
<td></td>
</tr>
</tbody>
</table>

© 2012 ACADEMY PUBLISHER
At the same time, we assess visual quality of MRTS images in Fig.3 (a)-(d), (f), (h), (k)-(m) using DSIS method, as is shown in Fig.4. Experimental results show these images reconstructed by MRTS algorithm have visual impairment. But visual impairment of JPEG2000 images is almost imperceptible. The main reason why mean option score (MOS) of MRTS images is not high is the loss of non-texture detailed information.

According to the characteristics of human visual system (HVS), the visual sensitivity of smooth region is much higher than that of texture intensive region, so the visual impairment of MRTS images to some extent can be reduced. In addition, if we assume that these MRTS images containing massive textures are used to compose the still background of images or videos and there are some persons, or some animals, etc. as the foreground, the visual impairment in these still background regions should not be noticeable because foreground objects can receive much more visual attention than the still background, especially texture background. That is our motivation of proposing the MRTS algorithm for image. Our algorithm can be used to reconstruct still background of images or videos.

In our system, bytes of synthesized image are computed by (6):

$$\psi = \sum_{i=1}^{N} (E_i + G_i)$$  \hspace{1cm} (6)

Where, $\psi$ is bytes of synthesized image; $N$ is number of regions; $E_i$ is bytes of polynomial coefficients of curve of region whose index number is $i$; $G_i$ is bytes of texture sample of region whose index number is $i$.

In Table 1, we show the number of bytes used in our proposed algorithm and compare them to the JPEG2000 compression for the equivalent images. In addition, our synthesized results are fair approximations and we may do more useful attempt for establishing more sophisticated image reconstruction algorithms to synthesize very realistic images.

**B. MRTS for Video**

Dynamic texture synthesis can be performed at frame rate. The dimension of the state $n$ and input $n_v$ is given as an input argument. In our implementation, we have used $\tau$ between 100 and 150, $n$ between 10 and 50 and $n_v$ between 10 and 30.
In order to illustrate the implementing process of the proposed MRTS for videos, we show the results of each phase in Fig. 5. Fig. 5 (a) displays original video sequences. Through multi-regions extraction, Fig. 5 (b) and Fig. 5 (c) display the static regions and the moving regions, respectively. Fig. 5 (d) and Fig. 5 (e) show the restored static regions and the restored moving regions, respectively. Finally, the reconstructed video is achieved through regions merge, as is shown in Fig. 5 (f).

There are more reconstructed videos with different number of moving regions and static regions in Fig. 6. Fig. 6 shows the consecutive five frames of original videos and corresponding reconstructed videos. Frames in the first row are original frames and frames in the second row are synthesized ones.

We also can see some distortion between the original video and the synthesized one. That is mainly also caused by the loss of non-texture detailed information. However, the distortion of shapes of regions isn’t obvious.

According to the characteristics of human visual system (HVS), If these MRTS videos are used to compose the dynamic backgrounds of video sequences, and there are some steamships, or some flying birds, etc. as the moving foreground, the visual impairment in these moving regions and static regions of backgrounds should not be noticeable because moving foreground objects can receive much more visual attention than the background, especially texture background. The faster foreground target motions, the lower the visual sensitivity of background is. That also is our motivation of proposing multi-regions texture substitution for video. In particular, our proposed approach can be regarded as a new, simple and reliable method for dynamic background reconstruction and achieve infinite and different dynamic backgrounds.

V. CONCLUSION

In this paper, we have investigated the multi-regions texture substitution for image and video based on the human vision characteristic. Through selecting the smallest region texture sample, The MRTS algorithm can achieve medium-level-visual texture substitution and effectually eliminate human visual perceptual redundancy. And compared with JPEG2000 compression, MRTS for image can get more coding gain but remain most image information people care. Simulation results show that MRTS for image can obtain optimal reconstructed results. This algorithm can be used to reconstruct still backgrounds of images or videos. At the same time, MRTS for video can get good reconstructed videos of arbitrary length. This approach can be regarded as a new, simple and reliable method for dynamic background reconstruction.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation of China key project (Grant No. 60832004) and youth project (Grant No. 51008009).

REFERENCES

Yanhui Liu, Postdoctor, Certified structural engineer, 1st. grade of China, State Key Laboratory for Seismic Reduction /Control & Structural Safety (Cultivation), Guangzhou University. Research field: structure vibration control.

In 2003, 2007 and 2009, I obtained the bachelor's degree of engineering in Haiman University, master degree of engineering in Lanzhou University of Technology and doctor degree of engineering in Lanzhou University of Technology respectively. In 2009-2011, I went in for postdoctor in Beijing University of Technology. Now, I am working in State Key Laboratory for Seismic Reduction /Control & Structural Safety (Cultivation), Guangzhou University. In the past several yeas, I taken part in the design of hybrid mass dampers for Guangzhou Tv and Sighseeing Tower (600m) in china as principal designer, and achieved the shaking table test of uncontrolled scale model, AMD-control scale model and HMD-control scale model under different control strategy and the adjusting work of actual HMD control system of Guangzhou Tv and Sighseeing Tower. Now, I hold one National Natural Science Foundation of China and one Postdoctoral Science Foundation of China, obtain one patent of China and have published twenty academic paper of which most have been indexed SCI or EI.