2D+T AUTOREGRESSIVE FRAMEWORK FOR VIDEO TEXTURE COMPLETION

F. Racape* *D. Doshkov** *M. Köppel**† *P. Ndjiki-Nya**

* Fraunhofer Institute for Telecommunications Heinrich Hertz Institute (HHI)
Image Processing Department,
Einsteinufer 37. 10587 Berlin, Germany
† Berlin Institute of Technology,
Department of Telecommunication Systems, School of Electrical Engineering and Computer Science,
Einsteinufer 17. 10587 Berlin, Germany

ABSTRACT

In this paper, an improved 2D+t texture completion framework is proposed, providing high visual quality of completed dynamic textures. A Spatiotemporal Autoregressive model (STAR) is used to propagate the signal of several available frames onto frames containing missing textures. A Gaussian white noise classically drives the model to enable texture innovation. To improve this method, an innovation process is proposed, that uses texture information from available training frames. The proposed method is deterministic, which solves a key problem for applications such as synthesis-based video coding. Compression simulations show potential bitrate savings up to 49% on texture sequences at comparable visual quality. Video results are provided online to allow assessing the visual quality of completed textures.

Index Terms— Texture completion, parametric method, autoregressive model.

1. INTRODUCTION

Texture completion is the art of filling unknown regions in images and videos. Texture completion or extrapolation is addressed by two main techniques: texture synthesis and inpainting. Texture synthesis [1] is based on reproducing the statistics of textures whereas the inpainting methods [2] aim at propagating structures. Two main approaches can be distinguished: *parametric* methods estimate a model of the texture distribution and then fill the output surface, whereas *non-parametric* approaches build the output signal on the fly by matching known information with input samples. Parametric methods aim at filling the output surface by approximating the probability density function (PDF) of the source texture. Among existing parametric approaches [3, 4, 5, 6], autoregressive (AR) methods have shown good results when synthesizing stationary textures which can be approximated as Markov Random Fields (MRF). Although the Autoregressive Moving Average (ARMA) method [6] provides a better temporal modeling, it is hardly adaptable to spatial and temporal boundaries. Many works, such as Woods [7], Chellappa et al. [4], Deguchi [5] and Tugnait [8], adapted the original 1D AR model to synthesize 2D textures. Kokaram and Rayner [9] proposed to fill blotches in old movies by interpolating the signal via 3D AR-based interpolation, using a model bigger than the missing region. Szummer [10] extended the idea to video extrapolation with a spatio-temporal AR (STAR) model. Concerning synthesis-based compression, the parametric approaches are faster than non-parametric ones [11]. The work of Khandelia et al. [12] presents a hybrid video codec based on STAR synthesis. Bitrate savings are promising, however the output images are of small size and difficult to evaluate. Bao et al. [13] proposed a scheme in which AR training is done for each frame in order to synthesize them from H.264 reference frames. However, AR parameters must be computed and transmitted to the decoder for each region at each frame since the distance between the processed frame and references changes.

In this paper, a 2D+t parametric approach has been chosen for its low computational cost, in particular when dealing with videos. We propose to improve the STAR method by changing the innovation term computation, which results in high quality completed textures. According to expert viewers subjective assessment, the proposed framework visually outperforms existing texture completion approaches in this context. The resulting sequences are provided online, enabling the reader to assess their quality. Moreover, important bitrate savings are measured compared to the last standard for video compression HEVC [14].

The following Sec. 2 presents the framework and details the proposed improvements. Experimental results are then presented in Sec. 3.

This work was carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 246016.

This work was funded in part by the German Research Foundation [Deutsche Forschungsgemeinschaft (DFG)], under the grant WI 2032/4-1.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

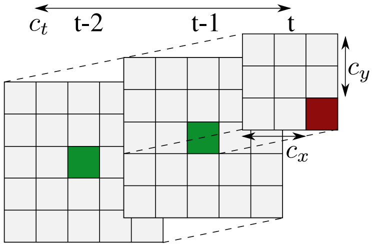

Fig. 1. STAR model in the case of a spatial size of 5×5 and a temporal size of two past frames $(c_x = 2, c_y = 2, c_t = 2)$.

2. PROPOSED FRAMEWORK

The proposed scheme aims at synthesizing missing dynamic textures, using available reference samples which can come from past but also future frames. Its input is a video with some frames containing marked missing textured regions. It outputs a processed video where the missing parts are filled in. In the rest of the paper, reference groups of pictures are called GOP_{Ref} and groups of pictures with missing textures to be synthesized are called $\text{GOP}_{\text{Synth}}$. The proposed texture completion framework contains the following steps.

- The STAR model is estimated, which requires to train it on available texture samples. In the video completion context, the training area (TA) can contain pixels from past and future frames when available.

- The main improvement of the proposed STAR scheme lies in a new step for computing the innovation term.

- $\text{GOP}_{\text{Synth}}$ are filled using the STAR method.

- Contrary to AR interpolation [9], the causal model is smaller than the missing region. Seam artifacts may then appear at bottom and right borders due to the raster scan filling. A low computational cost post-processing approach, based on Poisson editing [15], is therefore proposed.

2.1. Classical STAR method

The STAR synthesis is performed one pixel at a time using a linear combination of its causal neighbors plus an additive innovation term. In the case of 2D+t, each synthesized pixel can be expressed as

$$\hat{I}_p = \sum_{q \in V} \alpha_q I_q + \epsilon(p), \tag{1}$$

where I_q represents a source sample at location q and $\hat{I}_p$ represents the completed sample at location $p = (x, y, t)$. V contains the known spatial and temporal neighboring samples of p. An example of this neighborhood is depicted in Fig. 1. As in [12], squares of size $(2c_x + 1, 2c_y + 1)$ centered on the position collocated with the current pixel are used in past frames since all their samples are expected to be known. The coefficients α_q correspond to the STAR parameters and $\epsilon(p)$ denotes the innovation term at the current location p. The STAR parameters are estimated by means of solving the following least square problem

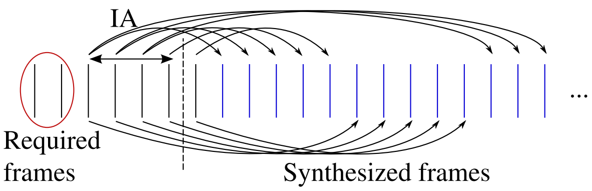

Fig. 2. Frame selection for current innovation term. Example with a IA of 4 frames. When the temporal STAR model size $c_t = 2$, 2 more past frames are required.

$$\boldsymbol{\alpha}_{C\times1} = \arg\min_{\alpha} \|\mathbf{y}_{S\times1} - \mathbf{X}_{S\times C}\boldsymbol{\alpha}_{C\times1}\|^2, \tag{2}$$

where $\boldsymbol{\alpha}$ $\left(\boldsymbol{\alpha} \in \mathbb{R}^C\right)$ is a vector containing the STAR parameters in a raster scan order. $\mathbf{y}$ $\left(\mathbf{y} \in \mathbb{R}^S\right)$ denotes the known samples in the TA and $\mathbf{X}$ $\left(\mathbf{X} \in \mathbb{R}^{S\times C}\right)$ represents the neighborhood matrix for each of the samples $\mathbf{y}$. C is the number of STAR parameters and $S = s_x s_y s_t$ the size of the TA which contains s_t frames. Eq. 2 can be solved with the closed-form solution:

$$\boldsymbol{\alpha} = (\mathbf{X}^T\boldsymbol{\alpha})^{-1}(\mathbf{X}^T\mathbf{y}). \tag{3}$$

When Eq. 2 has no solution, a pseudo inverse of $\mathbf{X}^T\mathbf{X}$ is determined [16]. Classically, the function ϵ is a Gaussian white noise process with zero mean and variance :

$$\sigma^2 = \frac{\|\mathbf{y}_{S\times1} - \mathbf{X}_{S\times C}\boldsymbol{\alpha}_{C\times1}\|^2}{S}, \tag{4}$$

and denotes the innovation term which drives the STAR model. In this work, we propose another way of computing this innovation term to better capture and reproduce known textures in the missing region.

2.2. New innovation term

The STAR model has been estimated from the TA which has been designed at the first step based on known samples. We propose to reuse this data. To compute the innovation term, a new area can be defined which contains known samples in available frames and is collocated with the missing region. In the following, we call this new area the Innovation term Area (IA). The IA is of the same spatial size as the missing region and of temporal size s'_t and may coincide with the TA but this is not mandatory. The previously computed STAR model is then applied on the IA, which results in a residual when comparing to the source signal. This residual per frame of the IA defines the innovation term:

$$\epsilon(p') = \sum_{q \in V} \alpha_q I_q - I_{p'}, \tag{5}$$

with $p' = (x, y, t')$ corresponding to the spatial position of $p = (x, y, t)$ in Eq. 1 but at different temporal location among

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

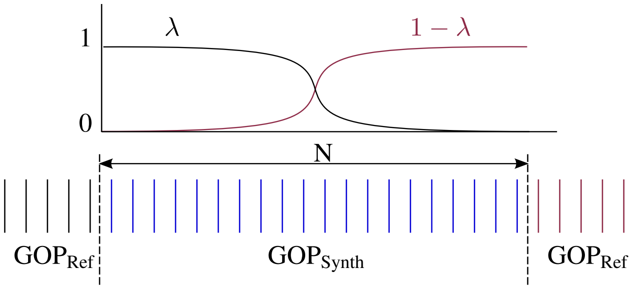

Fig. 3. Merge of forward and backward synthesis using the function λ.

IA frames. While filling a missing texture, the remaining choice lies in the frame position to determine the current innovation term. Two main options have been experimentally tested: random and same order cycle. Fig. 2 depicts the latter which provides the best visual results. In this case, if t_0 corresponds to the first frame to be synthesized and $t'_0 = t_0 - s'_t$ the first frame in the IA,

$$\hat{I}_{p(t_0+t)} = \sum_{q \in V} \alpha_q I_q + \epsilon(p'(t'_0 + t \bmod s'_t)), \qquad (6)$$

where mod s'_t corresponds to the modulo operation to drive the frame selection, as shown with arrows in Fig. 2. Not only does this new innovation term provide higher visual quality, but also is deterministic, which is a key issue when encoding frames. This allows the encoder to perform prediction using synthesis modes, while knowing the exact future reconstructed values at the decoder side.

Furthermore, when dealing with the YUV color space, this method enables to compute the training on the Y channel only. U and V are then copied from the current IA location. The coherency between color channels is thus preserved, whereas this is impossible when separately synthesizing the 3 channels. Computation time is also reduced since the model is computed on Y only.

The following section details the proposed way of dealing with both past and future GOP_{Ref}.

2.3. Handling consistency with both temporal boundaries

In the context of video compression or transmission errors, some future frames can be available and the temporal consistency has to be saved. In this work, we propose to perform forward synthesis using past frames and backward synthesis using available frames in the future. Then, both forward and backward versions are merged using the coefficients

$$\lambda(t) = \frac{1}{1 + \exp(\frac{N}{2} - t)} \qquad (7)$$

where N corresponds to the number of synthesized frames. Then, the final values are given by

$$\hat{I}_p = \lambda(t)\hat{I}_p^{fwd} + (1 - \lambda(t))\hat{I}_p^{bwd} \qquad (8)$$

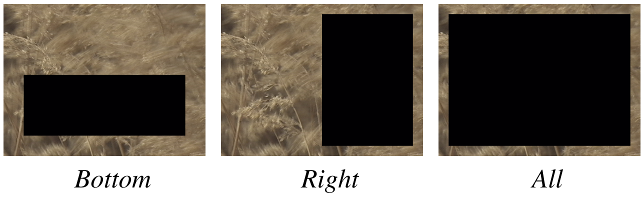
Bottom *Right* *All*

Fig. 4. Masks used in experiments. *Bottom*: 512×160, *Right*: 352×416 and *All*: 576×416.

$\hat{I}_p^{fwd}$ (resp. $\hat{I}_p^{bwd}$) being the forward (resp. backward) synthesized values at frame t. This merge function weighted by λ is depicted in Fig. 3 to prevent blurring on many frames. To save computation time, the training step can be performed once. The STAR model is then used for both forward and backward directions, while the IA contains future frames for the backward synthesis.

3. EXPERIMENTAL RESULTS

This section provides video synthesis results and potential video coding gains. Although quality metrics have received increasing attention over recent years, no satisfactory tools enable to assess the perceived quality of dynamic textures. This being so, extensive video results can be found on the following web page:
http://fabienracape.fr/hhi/TextureCartoonCompletion.html
For experiments, a test set of 4 dynamic textures has been selected, containing the HD sequences *Red Kayak*, *Riverbed*, *West Wind Easy* and *Ducks Take Off*, formatted to 8 bit 4:2:0 YUV files and cropped to 640 × 480 (73 frames). 3 regions to be synthesized have been designed as shown in Fig. 4.

Fig. 5 depicts the adopted sequence structure, which is adapted to HEVC random access GOPs [14]. The structure contains two synthesized GOPs of 23 frames each. GOP_{Ref} contain one more frame compared to HEVC GOPs (in gray) to preserve the first key frame for predicting the rest of the following GOP. One can also notice that the length of $\{\text{GOP}_{\text{Ref}} + \text{GOP}_{\text{Synth}}\}$ corresponds to the intra period, which allows to start a new encoding sequence with an Intra picture, after synthesized frames. Hence, the synthesis has no impact on the prediction of future temporally predicted frames.

The presented image results come from the synthesis with a STAR model of size ($c_x = c_y = 7, c_t = 1$). The TA and the IA are collocated with the missing region. 9 frames are available minus $c_t = 1$ required frame (Fig. 5 and 2). The TA size is set to 4 since it drastically impacts the computation time by setting the number of equations for the least square solver. The proposed order for the innovation term computation can result in temporally periodic artifacts when $N > s'_t$, since the use of IA frames loops (cf. Fig. 4). However, these artifacts remain much less annoying than those introduced by randomly choosing the frame in the IA (cf. online videos). Therefore, the larger the IA the better the loop artifact is re-

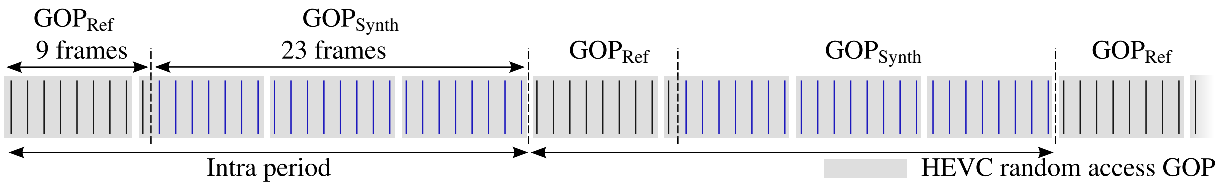

Fig. 5. Experimental GOP structure. Reference GOPs (GOP_{Ref}) and synthesized GOPs (GOP_{Synth}) do not follow the HEVC's random access GOP structure which basically contains 8 frames.

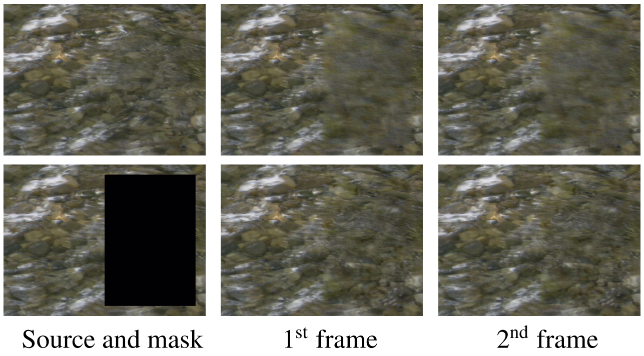
Source and mask 1st frame 2nd frame

Fig. 6. STAR Results on *Riverbed* using (top row) the classical white noise and (bottom row) the proposed method. The first two synthesized frames are shown.

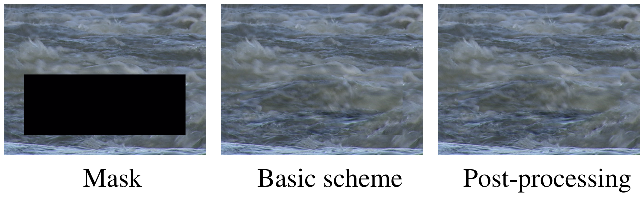
Mask Basic scheme Post-processing

Fig. 7. Results on *Red Kayak* using the proposed method without (basic) and with post-processing. One can notice the seams without post-processing in the bottom and right.

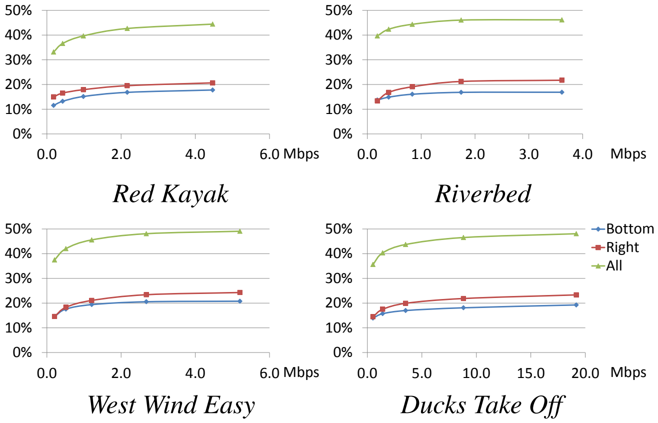

Fig. 8. Observed bitrate savings when removing the regions. The GOP structure depicted in Fig. 5.

duced. Thus $s'_t = 8$. Although these parameters should ideally be optimized for each input texture, the proposed parameters give good results in most cases.

In this case of texture prediction, the classical AR methods using white noise are not adapted to fully describe natural textures. One can see in Fig. 6 the higher quality obtained when using the proposed scheme (same STAR parameters). Moreover, to process 73 frames with the *Bottom* region, $66s$ are required in average with our scheme and $138s$ with the basic (3 channels) STAR, using our c++ implementation.

Fig. 7 shows the Poisson editing-based post-processing result, using an overlap of 16 pixels at the bottom and right border. This method proves to be sufficient for an efficient seam removal.

Concerning the compression application, Fig. 8 shows the different bitrate savings when encoding the sequence *Red Kayak* with and without regions removal, using the HEVC software version HM 11.0 [17]. The range of QP $\{20, 25, 30, 35, 40\}$ has been tested. The Coding Units to be synthesized are skipped by the encoder and flagged to the decoder for synthesis. One can notice, for instance, gains of 10% to 25% with the *Right* region. The limits of these results lie in the hypothesis that the produced artifacts are imperceptible. Extensive video results can be assessed online but no objective metrics results can be provided. In this case, the lower the QP, the higher the bitrate and thus bitrate savings since much less information is transmitted. However, the presented results show how promising this scheme could be, coupled with an adapted prediction mode selection.

4. CONCLUSION AND FUTURE WORK

This paper proposed an improved 2D+t autoregressive method for video texture completion providing high quality video results. The synthesis is deterministic and can be controlled which is of great value for the video coding application. Avoiding the transmission of residuals of dynamic regions can drastically reduce bitrate, using the HEVC software HM 11.0. Future work will focus on merging forward and backward synthesis and on the better utilization of the information from both past and future frames. Assessing the quality of the synthesis at the encoder side also remains a key issue for mode selection : classical prediction vs. synthesis.

5. REFERENCES

[1] L.-Y. Wei, S. Lefebvre, V. Kwatra, G. Turk, et al., "State of the art in example-based texture synthesis," in *Eurographics 2009, State of the Art Report, EG-STAR*, 2009, pp. 93–117.

[2] M. Bertalmío, V. Caselles, S. Masnou, and G. Sapiro, "Inpainting," *Encyclopedia of Computer Vision, Springer*, 2011.

[3] J. Portilla and E. P Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[4] R. Chellappa and R.L. Kashyap, "Texture synthesis using 2-d noncausal autoregressive models," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 1, pp. 194–203, 1985.

[5] K. Deguchi, "Two-dimensional auto-regressive model for analysis and sythesis of gray-level textures," *Proc. of the 1st Int. Sym. for Science on Form, General Ed. S. Ishizaka, Eds. Y. Kato, R. Takaki, and J. Toriwaki*, pp. 441–449, 1986.

[6] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[7] J.W. Woods, "Two-dimensional discrete markovian fields," *Information Theory, IEEE Transactions on*, vol. 18, no. 2, pp. 232–240, Mar 1972.

[8] J. K. Tugnait, "Estimation of linear parametric models of nongaussian discrete random fields with application to texture synthesis," *Image Processing, IEEE Transactions on*, vol. 3, no. 2, pp. 109–127, 1994.

[9] A. Kokaram and P. Rayner, "Detection and interpolation of replacement noise in motion picture sequences using 3d autoregressive modelling," in *Circuits and Systems, 1994. ISCAS '94., 1994 IEEE International Symposium on*, May 1994, vol. 3, pp. 21–24 vol.3.

[10] M. Szummer and R. W. Picard, "Temporal texture modeling," in *Image Processing, 1996. Proceedings., International Conference on*. IEEE, 1996, vol. 3, pp. 823–826.

[11] P. Ndjiki-Nya, D. Doshkov, H. Kaprykowsky, F. Zhang, D. Bull, and T. Wiegand, "Perception-oriented video coding based on image analysis and completion: A review," *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 579 – 594, 2012.

[12] A. Khandelia, S. Gorecha, B. Lall, S. Chaudhury, and M. Mathur, "Parametric video compression scheme using AR based texture synthesis," in *Computer Vision, Graphics Image Processing, 2008. ICVGIP '08. Sixth Indian Conference on*, 2008, pp. 219–225.

[13] Zhihua Bao, Chen Xu, and Chong Wang, "Perceptual auto-regressive texture synthesis for video coding," *Multimedia Tools and Applications*, pp. 1–13, 2013.

[14] G.J. Sullivan, J. Ohm, Woo-Jin Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[15] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3, pp. 313–318, 2003.

[16] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, vol. 2, no. 2, pp. 205–224, 1965.

[17] HEVC reference software HM version 11.0, "https://hevc.hhi.fraunhofer.de/svn/svn_hevcsoftware/tags/hm-11.0," .

978-1-4799-5751-4/14/$31.00 ©2014 IEEE