Texture Representations for Image and Video Synthesis

Georgios Georgiadis[1], Alessandro Chiuso[2], Stefano Soatto[1],
[1]Department of Computer Science, University of California, Los Angeles. [2]Department of Information Engineering, University of Padova.

"Visual textures" are regions of images that exhibit some form of spatial regularity. In applications such as texture synthesis and classification, algorithms require a small texture to be provided as an input, which is assumed to be representative of a larger region to be re-synthesized or categorized. We aim to characterize and infer such representatives automatically. We construct a new representation that compactly summarizes a texture, while using significantly less storage, that can be used for texture compression and synthesis.

To characterize visual textures we use the notions of Markovianity, stationarity and ergodicity. A texture is then defined as *a region* Ω *of an image* I *that can be rectified into a sample of a stochastic process that is stationary, ergodic and Markovian*. It is parametrized by (a) The Markov neighborhood ω and its Markov scale $r = |\omega|$, (b) the stationarity region $\bar{\omega}$ and its stationarity scale $\sigma = |\bar{\omega}|$, (c) a sufficient statistic θ_ω defined on ω, and (d) Ω, the texture region. Note that $\omega \subset \bar{\omega} \subset \Omega$. In describing a texture, we seek the *smallest* ω, in the sense of minimum area ("scale") $|\omega| = r$, so the corresponding θ_ω is a *minimal (Markov) sufficient statistic*.

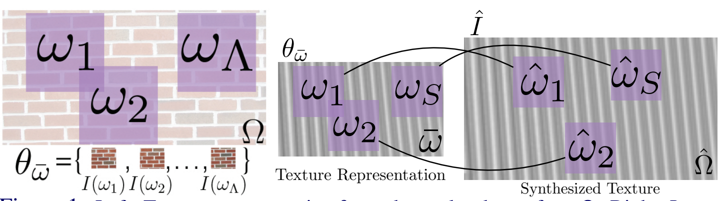

Figure 1: *Left: Texture representation $\theta_{\bar{\omega}}$ and samples drawn from Ω. Right: Image Texture Synthesis. For each neighborhood $\hat{\omega}_s$ in the synthesized texture, we find its nearest neighbor in $\bar{\omega}$.*

In a non-parametric setting, θ_ω is a collection of intensity values. $\bar{\omega} \doteq \bigcup_{\lambda=1,\ldots,\Lambda} \omega_\lambda$ is the union of Λ sample regions ω_λ. Collectively the neighborhoods capture the variability of the texture. A texture is represented by (a) ω_λ, chosen as a square for all λ with unknown area r, (b) $\bar{\omega}$, to be determined and (c) $\theta_{\bar{\omega}} \doteq \{\theta_{\omega_\lambda}\}_{\lambda=1}^{\Lambda} \doteq \{I(\omega_\lambda)\}_{\lambda=1}^{\Lambda}$ that is uniquely specified by the image given r and ω_λ (Fig. 1).

Given a representation $\{\omega, \bar{\omega}, \theta_{\bar{\omega}}\}$, we can synthesize novel instances of the texture by sampling from $dP(I(\omega))$ within $\bar{\omega}$. We choose a subset of neighborhoods from $\bar{\omega}$ that satisfy the compatibility conditions and by construction also respect the Markov structure. We perform this selection and simultaneously also infer $\hat{I}$ by minimizing [1],

$$E(\hat{I}, \{\omega_s\}_{s=1}^{S}) = \sum_{\hat{\omega}_s \in \hat{\Omega}_S} \nu_{\hat{\omega}_s} \|\hat{I}(\hat{\omega}_s) - I(\omega_s)\|^2. \quad (1)$$

An illustration of the quantities involved is shown in Fig. 1. $\nu_{\hat{\omega}_s}$ is used to reduce the effect of outliers. The process is performed in a multi-scale and multi-resolution fashion.

We extend synthesis to video, by performing synthesis using a causal approach. We use the already synthesized frames from previous time steps as a boundary condition and extend the textures to the next frame. Using a causal approach we also synthesize multiple textures simultaneously for video and images without computing a segmentation map. This is useful for applications such as video compression, hole-filling and frame interpolation (see Fig. 2). Boundary conditions are implicitly defined by the computed "structure" regions of the videos.

To evaluate the quality of the texture synthesis algorithm, we need a criterion that measures the similarity of the input, I, and synthesized, $\hat{I}$, textures. We introduce the Texture Qualitative Criterion (*TQC*), represented by E_{TQC}, which is composed of two terms. The first, $E_1(\hat{I}, I)$, penalizes structural dissimilarity, whereas $E_2(\hat{I}, I)$ penalizes statistical dissimilarity. We let $\hat{\omega}_s/\omega_i$ be patches within $\hat{\Omega}/\Omega$, the domains of $\hat{I}/I$, and their nearest neighbors be $\omega_s/\hat{\omega}_i$, which are selected within the domains of $I/\hat{I}$:

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

Figure 2: *Video texture synthesis in natural images (Hole-filling). From left to right: (i) Last frame (5th) of input video, (ii) Structure regions, (iii) Structure / Texture regions, (iv) Synthesized frame (our result).*

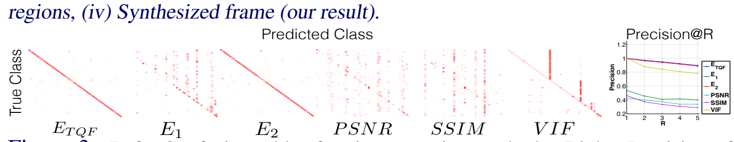

Figure 3: *Left: Confusion tables for six competing methods. Right: Precision of methods for various values of retrieved nearest neighbors.*

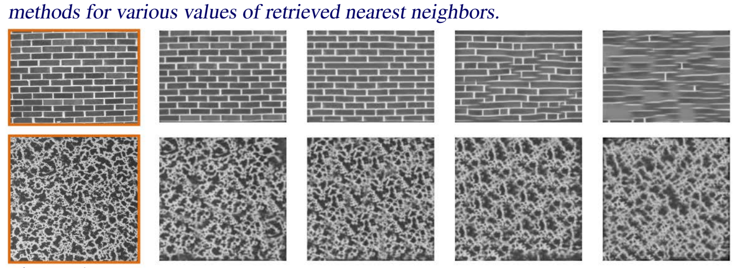

Figure 4: *Ordered synthesized textures using TQC. Left: Input texture. Right: Synthesized textures, left is the most similar to the input.*

$$E_1(\hat{I}, I) = \frac{1}{2N_I} \sum_{i=1}^{N_I} \frac{1}{|\omega_i|} \|\hat{I}(\hat{\omega}_i) - I(\omega_i)\|^2 + \frac{1}{2N_S} \sum_{s=1}^{N_S} \frac{1}{|\hat{\omega}_s|} \|\hat{I}(\hat{\omega}_s) - I(\omega_s)\|^2, \quad (2)$$

$$E_2(\hat{I}, I) = \frac{1}{L} \sum_{l=1}^{L} \|\phi(g_l(I)) - \phi(g_l(\hat{I}))\|_{\chi^2}, \quad (3)$$

where $\|.\|_{\chi^2}$ is the χ^2 distance, $\phi(.)$ is a histogram of filter response values and $g_l(I), l = 1, \ldots, L$ are the responses of the L filters. *TQC* is given by:

$$E_{TQC}(\hat{I}, I) = E_1(\hat{I}, I) + E_2(\hat{I}, I). \quad (4)$$

To evaluate *TQC*, we have constructed a dataset made out of 61 classes of textures, with 10 samples in each class. Each sample is compared against the other 609 texture images using six different quantities: E_{TQC}, E_1, E_2, *PSNR*, *SSIM* [3] and *VIF* [2]. We show confusion tables for $R = 5$ nearest neighbors for all competing methods and also plot the precision of each of the six methods in Fig. 3, for $R = 1, \ldots, 5$. To qualitatively evaluate *TQC*, we synthesized a number of textures and ordered them according to their similarity with the input texture using *TQC* (Fig. 4).

Our Contributions. (i) We summarize an image/video into a representation that takes significantly less space to store than the input, (ii) we use our representation for synthesis on images using the texture optimization technique, (iii) we extend this framework to video using a causal scheme and show results for multiple time-varying textures, (iv) we synthesize multiple textures simultaneously on video without explicitly computing a segmentation map useful for hole-filling and video compression, and (v) we propose a criterion ("Texture Qualitative Criterion" (*TQC*)) that measures structural and statistical dissimilarity between textures.

[1] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. Texture optimization for example-based synthesis. *Proc. of ACM SIGGRAPH*, 2005.

[2] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE TIP*, 2006.

[3] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 2004.