# Scene Completion Using Millions of Photographs

**Matei Negulescu**

James Hays and Alexei A. Efros. Scene completion using millions of photographs. In *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, page 4, New York, NY, USA, 2007. ACM

Hays and Efros present a technique that uses gist scene descriptors to find semantically similar scenes that can be used to fill in unwanted/unknown regions of an image. The authors motivate their work by noting that algorithms that extend textures from the same image into the target region tend to either produce bland looking images or fail completely because of lack of data. The main contribution of this work is to demonstrate that effectively modelling the visual world can be approximated by a brute force search of sufficiently large number of indexed images.

The system presented takes as input a source image, an outline of the target area, and a corpus of 2 million images with similar resolutions. A possible composite image is judged by adding different metrics: the scene matching distance, the local context distance, the local texture distance and the cost of the proposed graph cut required in the final blending stage. The 20 lowest scored composites are presented to the user.

In order to find images from the database that are semantically similar to the source image, the authors use the concept of a gist descriptor. The descriptor is constructed by decomposing the image using 6 oriented edge responses that analyze the intensity gradients (rate of change of the monochromatic image over a particular orientation) over 5 scales (to extract lower frequency information). Each of the resulting 30 matrices are downsampled to 4x4 spatial bins – the gist of the image. The gist of each image in the database is compared to the gist of the image minus the target region (the spatial bins are weighted based on how many valid pixels are in the bin) and the sum of square distance is recorded as the **scene matching distance**.

The local context of the target region (all pixels closer than 80px to the hole) is compared to the translated and scaled corpus image and the distance is recorded as the **local context distance**. The scales are restricted to .81,.90,1 and there is an increasing penalty for matches at high offsets. The texture score is computed as the image gradient magnitude at each pixel in the local context and downscaling the results to a 5x5 grid. The **local texture distance** is the sum of square distances of the source and database image's texture score.

Lastly, the proposed region must be cut on some seam. The optimal seam is found by minimizing the **graph cut cost** function $C(L) = \sum_p C_d(p, L(p)) + \sum_{p,q} C_i(p, q, L(p), L(q))$. $C_d(p, L(p))$ is the unary cost of assigning the pixel $p$ a label $L(p) = \{patch, exist\}$ corresponding to either patching from the new image or leaving the existing value. The target region has $C_d(p, exist)$ set to a high value and portions of the image not covered by the scene match have a high $C_d(p, patch)$. Otherwise, $C_d(p, patch) = (0.002 * Dist(p, hole))^3$ where $Dist$ is $p$'s distance from the hole. $C_i(p, q, L(p), L(q))$. $C_d(p, L(p))$ is only non-zero for adjacent pixels $p, q$ that don't share the same label. In this case, the cost is the magnitude of the gradient of the difference between the two images at points $p$ and $q$.

The authors admit that the algorithm fails to complete atypical scenes or those with uniform textures. In order to compare to existing techinques, the author designed a study wherein they asked 20 participants to judge whether each of 51 images were real or fake. Three even randomly assignmed subsets of images were either real, produced by the algorithm, or produced by Criminisi's single image fill-in technique. The authors report that 37% of their fake images were mislabeled real versus 10% for Criminisi's images. Lastly, the authors note that after 10 seconds, only 34% of their generated images were labeled as fake.

The presented system uses gist descriptors in a novel way by filtering out potential images allowing for a brute force. The implementation uses the ever expanding online image library to produce generally pleasing images that never existed. Though the authors admit that Criminisi's approach is better than theirs for certain cases, it is difficult to gauge what would be deemed 'unbiased' choices for image completion. Though the present system is more flexible, asking for completion of a wall versus merely the graffiti can be better for one method than another. Overall the authors spent too long introducing the paper instead of focusing more on a properly designed study and a time/space analysis of their algorithm. Moreover, the authors did not mention whether they precompute the gist descriptor for their entire dataset and how long that would take. The authors didn't mention the value for the local context translation penalty, but they included the 3 curiously selected scale values. Possible improvements involve allowing the user to specify a subset of images (people already structure similar images into folders). Additionally, the algorithm can presegment the images using rough structures such as ground, vertical planes and sky (Geometric Context - Hoiem et al, 2005).