Image similarity metrics

An interface for a metric will take two images, x and $\hat{x}$, to compute a similarity metric, $\text{sim}(x, \hat{x})$ that is symmetric. We take the view that unsuccessful rendering should be counted as absolute failures. As such, our image similarity metrics computed are conditioned on the successful rendering of the code into rendered images (i.e., if the rendering fails, the score will be 0 by default). In Image2Struct, we consider normalize metrics within the unit range so that they can be interpreted easily; a score of zero implies complete dissimilarity whereas a score of 1 implies that the images are identical. Without loss of generality, we assume both x and $\hat{x}$ are of dimensions $W \times H$.

Earth Mover Similarity (EMS) We introduce an image metric named Earth Mover's Similarity (EMS). It is inspired by the Earth Mover's distance (EMD)[1], which is a measure of dissimilarity between two frequency distributions.

To compute the EMD between two images x and $\hat{x}$, we first transforms the images into signatures $\mathcal{S}(x)$ and $\mathcal{S}(\hat{x})$, which are discrete distributions of features of Q elements.

$$\mathcal{S}(x) = \{(g_k, w_k^g) : 0 \leq k < Q\} \qquad \& \qquad \mathcal{S}(\hat{x}) = \{(h_l, w_l^h) : 0 \leq l < Q\} \tag{1}$$

We define a cost matrix $C \in \mathbb{R}^{Q \times Q}$ wherein each element $C[k, l]$ represents the cost of moving probability mass between g_k and h_l. We further denote the movement of probability mass between g_k and h_l by $f_{k,l}$. The optimal flow is the set of $\{f_{k,l}^*\}$ that satisfies the following optimization problem:

$$\min \sum_k \sum_l f_{k,l} C[k, l] \qquad \text{subject to} \tag{2a}$$

$$f_{k,l} \geq 0 \qquad \forall k, \forall l \tag{2b}$$

$$\sum_l f_{k,l} \leq w_k^g \qquad \forall k \tag{2c}$$

$$\sum_k f_{k,l} \leq w_k^h \qquad \forall l \tag{2d}$$

$$\sum_l \sum_k f_{k,l} = \min\{\sum_k w_k^g, \sum_l w_l^h\} \tag{2e}$$

The EMD can then be computed with Equation (3).

$$\text{EMD}(x, \hat{x}) = \frac{\sum_k \sum_l f_{k,l}^* C[k, l]}{\sum_k \sum_l f_{k,l}^*} \tag{3}$$

The signature is typically defined as the distribution of the grayscale values of an image when one wants to compare images. In other words, $\mathcal{S}(x)$ is the probability mass function where the random variable (i.e., g_k or h_l) is one of the possible pixel values (0 to 255) and the mass (i.e., w_k^g or w_l^h) is the normalized count of the number of pixels in x with that value. In this formulation, spatial information is lost and the metric is invariant to translation, reflection, and other pixel rearrangements. We solve this problem by defining multidimensional signatures that consider the pixels' x- and y-coordinates in addition to their values.

The support of our multidimensional signature, $\mathcal{S}_p$, is all the possible combinations of the x-coordinates (x-pos), y-coordinates (y-pos), and the N possible pixel values (pix):

$$\mathcal{S}_p(x) = \{((x\text{-pos}, y\text{-pos}, \text{pix})_k, w_k) : k \in \{0, 1, \cdots, WHN\}\} \tag{4}$$

The probability mass, w_k, takes the value of either $\frac{1}{HW}$ or 0. The complexity of computing the cost matrix over $\mathcal{S}_p$ is O(W^2H^2), making it difficult to compute for high resolution images. We therefore compute an approximated patch version of it, which we denoted as $\text{EMD}_{\text{block}}$.

In $\text{EMD}_{\text{block}}$, we first split two images, x and $\hat{x}$, each into K patches of dimensions $r \times s$: $P_x^0, \cdots, P_x^{K-1}$. Our implementation sets r and s individually for every image such that there

are 8×8 patches in every image. To compare two patches P_x^t and $P_{\hat{x}}^u$, we treat each patch as separate images and compute the EMD using the multidimensional signature defined in Equation (4), which we will denote as $\text{EMD}_p(P_x^t, P_{\hat{x}}^u)$. Next, we define a separate cost matrix, C_p, such that the cost of moving one patch to another is the sum of the EMD between the patches and the Euclidean distance between them:

$$C_p[i,j] = EMD(P_x^i, P_{\hat{x}}^j) + ||(x\text{-pos}_i, y\text{-pos}_i), (x\text{-pos}_j, y\text{-pos}_j)||_2 \tag{5}$$

$\text{EMD}_{\text{block}}$ attempts to minimize the cost of moving patches by solving the optimization problem defined in Equation (2), but with the new cost function, C_p. By considering both the positions and weights of the pixels within patches (through the multidimensional signature) and the distance between patches, $\text{EMD}_{\text{block}}$ heavily penalizes random shuffling of pixels and assigns a lower score (implying greater similarity) to a rendered image that contain blocks of similar but translated elements as the input (see illustration in Figure 1). This property is useful for discerning between pairs of images that contain similar elements —even if the elements are translated— and pairs where distribution of colors in the rendered image is similar to the input image.

Finally, we define the Earth Mover Similarity (EMS), a normalized similarity version of $\text{EMD}_{\text{block}}$. It compares $\text{EMD}_{\text{block}}(x, \hat{x})$ against $\text{EMD}_{\text{block}}(x, c(x))$, the EMD between the reference image and a constant black or white image ($c(x)$), whichever is the most dissimilar to the reference image x. An EMS of 0 indicates the least similarity and an EMS of 1 indicates the identity.

$$\text{EMS}(x, \hat{x}) = 1 - \frac{\text{EMD}_{\text{block}}(x, \hat{x})}{\text{EMD}_{\text{block}}(x, c(x))} \tag{6}$$

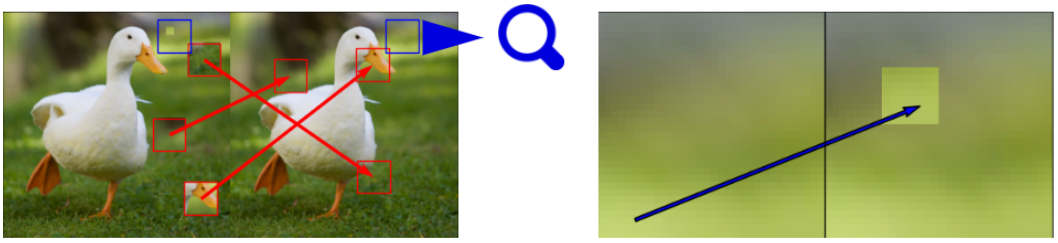

Figure 1: An illustration of the two scales at which $\text{EMD}_{\text{block}}$ operates. The left image is an altered copy of the right one in that 4 patches are manipulated. $\text{EMD}_{\text{block}}$ computes an optimal flow where 3 of these patches (in red) are moved completely without modification. For the blue patch, it decides that it incurs a lower cost to move some pixels within the patch (the zoomed version on the right). On top of moving blocks or pixels, $\text{EMD}_{\text{block}}$ can change the pixel colors at a cost (we do not illustrate color modification in this example for simplicity).

References

[1] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998.