The Prague Texture Segmentation Benchmark Criteria Set

The segmentation benchmark criteria is divided into four subsets:

- Region-Based Criteria
- Pixel-Wise Weighted Average Criteria
- Consistency Error Criteria
- Clustering Comparison Criteria

Symbols $\uparrow$ / $\downarrow$ denote required increase / decrease of the corresponding criterion.

The criteria with the exception of d_{VI} are displayed after multiplication by 100.

Contents

1 Region-Based Criteria [1]

The region-based criteria mutually compare the machine segmented regions $R_i \;\; i = 1, \ldots, M$ with the correct ground truth regions $\bar{R}_j \;\; j = 1, \ldots, N$ where $|R|$ is the corresponding set cardinality. The regions overlap acceptance is controlled by the threshold $(< 0.5; 1 >) \;\; k = 0.75$. Single region-based criteria are defined as follows:

$\uparrow$ **CS** (correct detection): $[R_m; \bar{R}_n]$ *iff*

(i) $|R_m \cap \bar{R}_n| \geq k \; |R_m|$

(ii) $|R_m \cap \bar{R}_n| \geq k \; |\bar{R}_n|$

$\downarrow$ **OS** (over-segmentation): $[R_{m1}, \ldots, R_{mx}; \bar{R}_n]$, $2 \leq x \leq M$ *iff*

(i) $\forall i \in < 1; x >, \; |R_{mi} \cap \bar{R}_n| \geq k \; |R_{mi}|$

(ii) $\sum_{i=1}^{x} \; |R_{mi} \cap \bar{R}_n| \geq k \; |\bar{R}_n|$

$\downarrow$ **US** (under-segmentation): $[R_m; \bar{R}_{n1}, \ldots, \bar{R}_{nx}]$, $2 \leq x \leq N$ *iff*

(i) $\sum_{i=1}^{x} \; |R_m \cap \bar{R}_{ni}| \geq k \; |R_m|$

(ii) $\forall i \in < 1; x >, \; |R_m \cap \bar{R}_{ni}| \geq k \; |\bar{R}_{ni}|$

$\downarrow$ **ME** (missed): $[\bar{R}_n]$ *iff*

(i) $\bar{R}_n \notin$ correct detection

(ii) $\bar{R}_n \notin$ over-segmentation

(iii) $\bar{R}_n \notin$ under-segmentation

$\downarrow$ **NE** (noise): $[R_m]$ *iff*

(i) $R_m \notin$ correct detection

(ii) $R_m \notin$ over-segmentation

(iii) $R_m \notin$ under-segmentation

Single region-based criteria are also available as the corresponding performance curves $CS(k), OS(k), US(k), ME(k), NE(k)$. The curves allow to compare sensitivity of different segmenters to the changing threshold value $(k \in < 0.5; 1 >)$.

Finally the last five region criteria are the approximations of the performance curves integrals

$$\bar{f} = \int_{0.5}^{1} f(k) \, dk \;\; ,$$

where $f(k)$ is some curve from $\{CS(k), OS(k), US(k), ME(k), NE(k)\}$.

2 Pixel-Wise Weighted Average Criteria

Let us denote

$$n_{i,\bullet} = \sum_{j=1}^{N} n_{i,j} \ , \qquad n_{\bullet,i} = \sum_{j=1}^{M} n_{j,i} \ ,$$

where N, M are the correct number of classes and the interpreted number of classes (or regions), respectively. $K = \max\{M, N\}$, n is the number of pixels in the test set, $n_{i,j}$ is the number of pixels interpreted as the i-th class but belonging into the j-th class. The error matrix $(\{n_{i,j}\})$ extended into $K \times K$ is obtained by padding missing entries with zeros. $\hat{\imath}$ is either i for supervised tests or mapping of the i-th class ground truth into an interpretation segment based on the Munkres algorithm (for unsupervised test). The following pixel-wise criteria were implemented:

$\downarrow$ **O** (omission error – the overall ratio of wrongly interpreted pixels):

$$O = \text{median} \left\{ \frac{O_i}{n_{\bullet,i}} \right\}_{i=1}^{N} = \text{median} \left\{ \frac{(n_{\bullet,i} - n_{\hat{\imath},i})}{n_{\bullet,i}} \right\}_{i=1}^{N} \qquad \langle 0; 1 \rangle \ ,$$

where O_i is the i-th class omission error

$\downarrow$ **C** (commission error – the overall ratio of wrongly assigned pixels):

$$C = \text{median} \left\{ \frac{C_i}{n_{\hat{\imath},\bullet}} \right\}_{\hat{\imath}=1}^{M} = \text{median} \left\{ \frac{(n_{\hat{\imath},\bullet} - n_{\hat{\imath},i})}{n_{\hat{\imath},\bullet}} \right\}_{\hat{\imath}=1}^{M} \qquad \langle 0; 1 \rangle \ ,$$

where C_i is the i-th class commission error

$\uparrow$ **CA** (the weighted average class accuracy):

$$CA = \frac{1}{n} \sum_{i=1}^{K} \frac{n_{\hat{\imath},i}\, n_{\bullet,i}}{n_{\bullet,i} + n_{\hat{\imath},\bullet} - n_{\hat{\imath},i}} \qquad \langle 0; 1 \rangle$$

$\uparrow$ **CO** (recall, the weighted average correct assignment):

$$CO = \frac{1}{n} \sum_{i=1}^{K} n_{\bullet,i}\, CO_i = \frac{1}{n} \sum_{i=1}^{K} n_{\hat{\imath},i} \qquad \langle 0; 1 \rangle$$

$\uparrow$ **CC** (precision, object accuracy, overall accuracy):

$$CC = \frac{1}{n} \sum_{i=1}^{K} n_{\bullet,i}\, CC_i = \frac{1}{n} \sum_{i=1}^{K} \frac{n_{\hat{\imath},i}\, n_{\bullet,i}}{n_{\hat{\imath},\bullet}} \qquad \langle 0; 1 \rangle$$

$\downarrow$ **I**. (type I error, the weighted probability of wrong assignment of classes pixels):

$$I = \frac{1}{n}\sum_{i=1}^{K}(n_{\bullet,i} - n_{\hat{\imath},i}) = 1 - CO \qquad \langle 0;1\rangle$$

$\downarrow$ **II**. (type II error, the weighted probability of commission error):

$$II = \frac{1}{n}\sum_{i=1}^{K}\frac{n_{\hat{\imath},\bullet}\, n_{\bullet,i} - n_{\hat{\imath},i}\, n_{\bullet,i}}{n - n_{\bullet,i}} \qquad \langle 0;1\rangle$$

$\uparrow$ **EA** (mean class accuracy estimate):

$$EA = \frac{1}{n}\sum_{i=1}^{K}\frac{2n_{\hat{\imath},i}\, n_{\bullet,i}}{n_{\bullet,i} + n_{\hat{\imath},\bullet}} \qquad \langle 0;1\rangle$$

$\uparrow$ **MS** (mapping score – emphasizes the error of not recognizing the test data):

$$MS = \frac{1}{n}\sum_{i=1}^{K}(1.5\, n_{\hat{\imath},i} - 0.5 n_{\hat{\imath},\bullet}) \qquad \langle -0,5;1\rangle$$

$\downarrow$ **RM** (root mean square proportion estimation error):

$$RM = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\frac{n_{\hat{\imath},\bullet} - n_{\bullet,i}}{n}\right)^2} \qquad \geq 0$$

indicates unbalance between the omission O_i and commission C_i errors, respectively

$\uparrow$ **CI** (comparison index – includes both these types of errors):

$$CI = \frac{1}{n}\sum_{i=1}^{K} n_{\hat{\imath},i}\sqrt{\frac{n_{\bullet,i}}{n_{\hat{\imath},\bullet}}} = \frac{1}{n}\sum_{i=1}^{K} n_{\bullet,i}\sqrt{CC_i CO_i} \qquad \langle 0;1\rangle \ ,$$

where CC_i, CO_i are the object precision and recall. CI reaches its maximum either for the ideal segmentation or for equal commission and omission errors for every region (class)

The F measure curve (see *Region-Based Criteria*)

$$F = \frac{1}{n}\sum_{i=1}^{K} n_{\bullet,i}\,\frac{CC_i\, CO_i}{\gamma CO_i + (1-\gamma)CC_i} \qquad \langle 0;1\rangle$$

3 Consistency Error Criteria [2]

Let S_1, S_2 are two segmentations, $R_{1,i}$ is the set of pixels corresponding to a region in the S_1 segmentation and containing the pixel i, $|R|$ is the set cardinality and $\setminus$ is the set difference. A refinement tolerant measure error was defined [2] at each pixel i:

$$\epsilon_i(S_1, S_2) = \frac{|R_{1,i} \setminus R_{2,i}|}{|R_{1,i}|} \quad .$$

This non-symmetric local error measure encodes a measure of refinement in one direction only. Two error measures for entire image are defined: Global Consistency Error (GCE) forces all local refinements to be in the same direction while Local Consistency Error (LCE) allows refinement in both directions.

$\downarrow$ **GCE** (global consistency error):

$$GCE(S_1, S_2) = \frac{1}{n} \min \left\{ \sum_i \epsilon_i(S_1, S_2), \sum_i \epsilon_i(S_2, S_1) \right\}$$

$\downarrow$ **LCE** (local consistency error):

$$LCE(S_1, S_2) = \frac{1}{n} \sum_i \min \left\{ \epsilon_i(S_1, S_2), \epsilon_i(S_2, S_1) \right\}$$

$$LCE, GCE \in \langle 0; 1 \rangle, \qquad LCE \leq GCE$$

4 Clustering Comparison Criteria [3]

A clustering $\mathcal{S}$ is a partition of a *data set* D into sets $R_1, R_2, \ldots R_M$ called *clusters* such that

$$R_k \cap R_l = \emptyset \quad \text{and} \quad \bigcup_{k=1}^{M} R_k = S\,.$$

Let the number of data points in D and in cluster R_k be n and n_k respectively. We have, of course, $n = \sum_{k=1}^{M} n_k$.

The number of points in the intersection of clusters R_k of $\mathcal{S}$ and $R'_{k'}$ of $\mathcal{S}'$ is denoted $n_{kk'}$:

$$n_{kk'} = |R_k \cap R'_{k'}|\,.$$

$\downarrow$ $\mathbf{d_{VI}}$ (variation of information):

$$d_{VI}(\mathcal{S}, \mathcal{S}') = H(\mathcal{S}) + H(\mathcal{S}') - 2I(\mathcal{S}, \mathcal{S}')$$

where H and I represent respectively the entropies[1] of and the mutual information[2] between the two clusterings.

$\downarrow$ $\mathbf{d_M}$ (Mirkin metric):

$$d_M(\mathcal{S}, \mathcal{S}') = \frac{d'_M(\mathcal{S}, \mathcal{S}')}{n^2}$$

$$d'_M(\mathcal{S}, \mathcal{S}') = \sum_k n_k^2 + \sum_{k'} {n'}_{k'}^2 - 2\sum_k \sum_{k'} n_{kk'}^2$$

$\downarrow$ $\mathbf{d_D}$ (Van Dongen metric):

$$d_D(\mathcal{S}, \mathcal{S}') = \frac{d'_D(\mathcal{S}, \mathcal{S}')}{2n}$$

$$d'_D(\mathcal{S}, \mathcal{S}') = 2n - \sum_k \max_{k'} n_{kk'} - \sum_{k'} \max_k n_{kk'}$$

---

[1] $H(\mathcal{S}) = -\sum_{k=1}^{M} \frac{n_k}{n} \log \frac{n_k}{n}$

[2] $I(\mathcal{S}, \mathcal{S}') = \sum_{k=1}^{M} \sum_{k'=1}^{N} \frac{n_{k,k'}}{n} \log \frac{n_{k,k'}}{n} \frac{n_k}{n} \frac{{n'}_{k'}}{n}$

References

[1] Adam Hoover, Gillian Jean-Baptiste, Xiaoyi Jiang, Patrick J. Flynn, Horst Bunke, Dmitry B. Goldgof, Kevin Bowyer, David W. Eggert, Andrew Fitzgibbon, and Robert B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(7):673–689, July 1996.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[3] Marina Meilă. Comparing clusterings: An axiomatic view. In *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pages 577–584, New York, NY, USA, 2005. ACM.