Hypothesis Testing Details

We used the hypothesis testing to verify the significance of the improvements provided by EGA optimized quantization thresholds. Hereafter, we explain the details of this test.

Part I. Hypothesis Testing

Our major concern was to test, against zero, the difference $P(n,\mathbf{\Theta}_{\text{opt}})-P(n,\mathbf{\Theta}_{\text{man}})$ between the average precisions of the wavelet correlogram indexing algorithm with the EGA-optimized ($\mathbf{\Theta}_{\text{opt}}$) and original ($\mathbf{\Theta}_{\text{man}}$) quantization thresholds. The improvements of other evaluation measures including average weighted precision, average recall, and average rank were verified in the same manner.

To decide about the closeness of the two average precisions, we tested for the following hypotheses:

$$\begin{aligned} H_1 &: \quad \Delta P = P(n,\mathbf{\Theta}_{\text{opt}})-P(n,\mathbf{\Theta}_{\text{man}}) \neq 0 \\ H_0 &: \quad \Delta P = P(n,\mathbf{\Theta}_{\text{opt}})-P(n,\mathbf{\Theta}_{\text{man}}) = 0 \end{aligned} \quad . \tag{i}$$

To this end, we computed the following test statistic (Theodoridis *et al* 2003[1]):

$$q_P = \left[P_{\text{ave}}(n,\mathbf{\Theta}_{\text{opt}}) - P_{\text{ave}}(n,\mathbf{\Theta}_{\text{man}})\right] \Big/ \sqrt{\frac{P_{\text{std}}(n,\mathbf{\Theta}_{\text{opt}})^2 + P_{\text{std}}(n,\mathbf{\Theta}_{\text{man}})^2}{|\mathbf{D}|}} \tag{ii}$$

where $|\mathbf{D}|$ is the cardinality of the reference imagebase. $P_{\text{ave}}(n,\mathbf{\Theta}_{\text{opt}})$ and $P_{\text{ave}}(n,\mathbf{\Theta}_{\text{man}})$ are computed by (63), and $P_{\text{std}}(n,\mathbf{\Theta}_{\text{opt}})$ and $P_{\text{std}}(n,\mathbf{\Theta}_{\text{man}})$ are given by (64) in the revised paper. The random variable q_P follows the *t*-distribution with $2|\mathbf{D}|-2$ degrees of freedom.

[1] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. San Diego: Academic Press, 2003.

For a specified significance level ρ, if the following criterion is satisfied, the hypothesis H_0 will be accepted which means $P_{\text{ave}}(n,\mathbf{\Theta}_{\text{opt}})$ and $P_{\text{ave}}(n,\mathbf{\Theta}_{\text{man}})$ are not statistically different with significance level ρ:

$$q_P \in C_\rho = [-c_\rho, c_\rho] \tag{iii}$$

where $C_\rho = [-c_\rho, c_\rho]$ is the confidence interval at level $1-\rho$:

$$P\left(-c_\rho \le q_P \le c_\rho\right) = 1-\rho \tag{iv}$$

where P indicates the probability. Otherwise, the hypothesis H_1 will be accepted which means $P_{\text{ave}}(n,\mathbf{\Theta}_{\text{opt}})$ and $P_{\text{ave}}(n,\mathbf{\Theta}_{\text{man}})$ are statistically different with significance level ρ.

Part II. Validating the Improvements

As stated in Subsection II-A (in the revised paper), we have $|\mathbf{D}|=1000$ therefore, the confidence intervals for significance levels $\rho = 0.01$ and $\rho = 0.05$ are respectively as follows:

$$C_{0.01} = [-2.58, 2.58] \tag{v}$$

$$C_{0.05} = [-1.96, 1.96] \tag{vi}$$

Using (ii) and the results demonstrated in Tables I and II, we obtained:

$$q_P = 3.06 \notin C_{0.01} \tag{vii}$$

$$q_{\overline{P}} = 2.42 \notin C_{0.05} \tag{viii}$$

$$q_R = 3.17 \notin C_{0.01} \tag{ix}$$

$$q_C = 2.10 \notin C_{0.05} \tag{x}$$

According to (iii), the above test confirmed that the average precision and recall were improved with statistically significance level ρ=0.01. It also demonstrated that the average weighted precision and rank were improved with statistically significance level ρ=0.05.