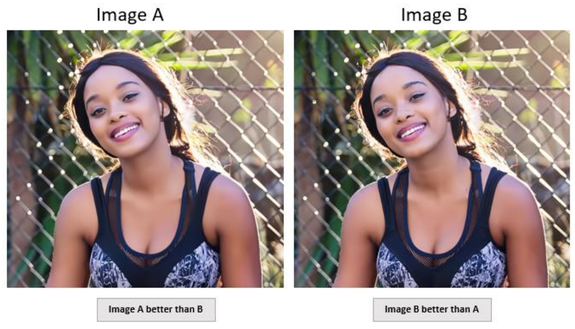
Predictive Sampling for Efficient Pairwise Subjective Image Quality Assessment

Shima Mohammadi, João Ascenso
Instituto de Telecomunicações – Instituto Superior Técnico, Lisbon, Portugal

1. Context and Objectives

- Pairwise comparison (PC) is a subjective test methodology to compare every pair of images side by side in terms of quality



- PC test provides robust, and reliable results, due to simple to answer questions.
- Despite reliable results it is expensive to conduct due to large number of trials (pairs).
- Number of pairs grows quadratically with the number of test images.
- To overcome expensive test, only a subset of pairs goes to subjective test

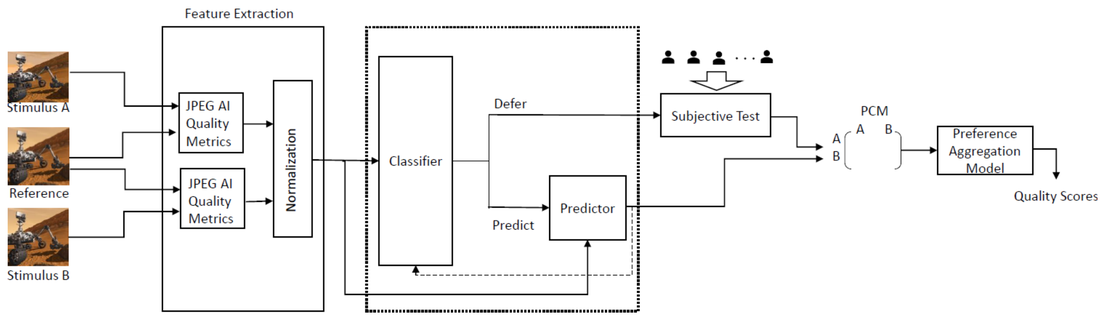
Objectives

- Exploit image characteristics for the first time, namely with objective quality assessment models, to sample the most informative image pairs and to estimate subjects' preferences

Contributions

- Propose a novel solution to perform pairwise sampling before the actual subjective assessment test starts
- Exploit image characteristics for the first time, namely with objective quality assessment models, to sample the most informative image pairs and to estimate subjects' preferences

2. Predictive Sampling for Pairwise Sampling (PS-PC)



- Feature extraction and normalization: Several full reference objective image quality metric are normalized and used as features. JPEG AI quality metrics (MS-SSIM, IW-SSIM, VMAF, VIF, NLPD, FSIM, PSNR-HVS) are used.
- Classifier: Classifier receives as input the set of features for each stimulus of a pair. The objective of the classifier is to perform a binary decision which is to classify as defer: pair must be evaluated by subjects, or as predict: subjects' preference is obtained using a predictor.
- Predictor: The predictor is responsible to estimate the probability of preferring one stimulus over the other in a pair and is only used when a pair is classified as predict.
  - The output of the predictor is also used during the training process to enable the classifier to perform better decisions

3. Training Procedure

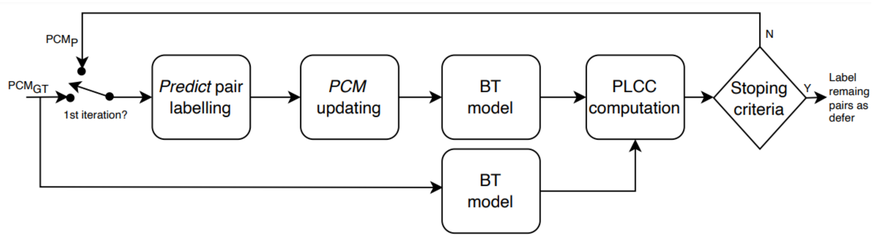
Labelling

- Training the classifier requires to label a dataset with values of defer or predict
- This process is iterative by selecting predict pairs until a stopping point is reached, and then all the remaining pairs are labeled as defer.
- The stopping point is defined for eta (η) reduction in the Pearson Linear Correlation Coefficient (PLCC) between the predicted and inferred scores. (edited)
- Several different labeling procedures were compared
  - Random-based: To randomly select some pairs as predict until reaching a stopping criteria.
  - Entropy-based: To select some pairs as predict based on the maximum entropy

$$H_{ij} = -p_{ij} log p_{ij} - (1 - p_{ij}) \log(1 - p_{ij})$$

  - Kullback Lieber Divergence (KLD)-based: To select pairs based on the minimum KLD. The KLD measure the divergence between the prior and posterior distribution after a pair is labeled as predict

$$KLD(\pi_{gt}, \hat{\sigma}_{gt}, \pi_p, \hat{\sigma}_p) = \sum_{i=1}^{n} log \frac{\hat{\sigma}_p}{\hat{\sigma}_g} - d \sum_{i=1}^{n} \frac{\hat{\sigma}_{gt}}{\hat{\sigma}_p} + \frac{1}{\hat{\sigma}_p} (\pi_{gt} - \pi_p)^2$$



Classifier Training

- Classifier receives the input features and predict a label. Two different classifiers were examined:
  - Support vector machine (SVM)
  - Extreme Gradient Boosting (XGBoost)

Predictor Training

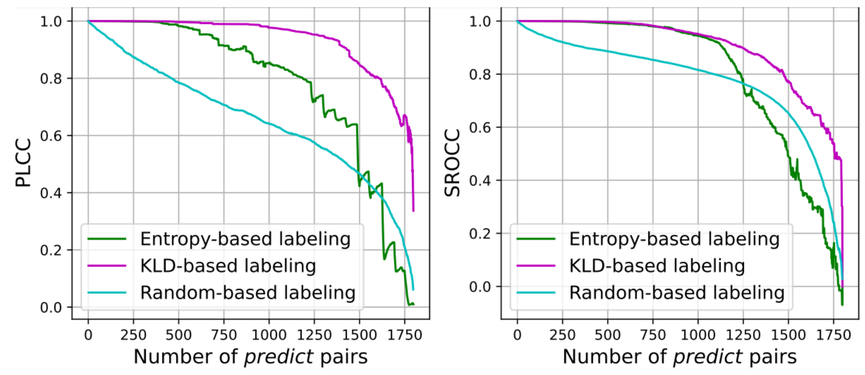
- Predictor is responsible to estimate the probability of preference for predict labeled pairs with Support vector regression (SVR).

4. Performance Evaluation

- **Dataset:**
  - PC-IQA dataset used for training and testing
  - TID2013 used for cross dataset evaluation
  - PieAPP dataset used for cross dataset evaluation

Labeling Evaluation

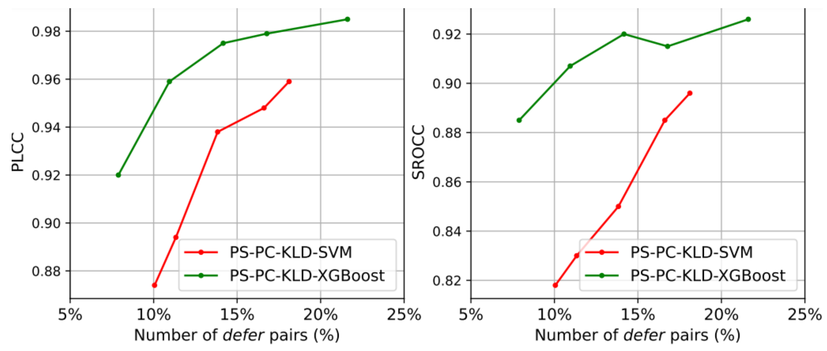
- The labeling algorithms presented are evaluated on the entire PC-IQA dataset without any stopping criteria to better understand the performance for a wide range of predict pairs.



- KLD-based labeling provides the best selection of pairs since PLCC gradually reduces and is always above the other labeling approaches.

Classifier Evaluation

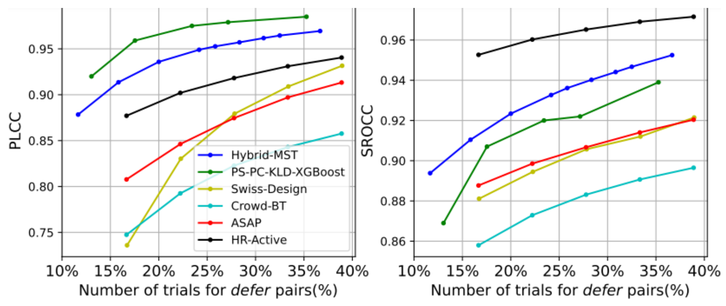
- Five different models for each classifier are trained
- In each model, the ground-truth labels were obtained by having different eta (η) in the labeling algorithm which defines the tradeoff between the number of pairs selected as defer (and thus the subjective test length) and the accuracy of the scores in the subjective test..



- The horizontal axis is the percentage of the number of defer pairs each classifier selects for the complete PC-IQA dataset (total of 1800 pairs).
- The experimental results show that XGBoost is the best choice.

State-of-the-art Evaluation

- PS-PC is compared with other random-based, sorting-based, and active-based approaches.



- Regarding PLCC: PS-PC is the best choice followed by Hybrid-MST
- Regarding SROCC: HR-Active followed by Hybrid-MST has higher performance than PS-PC.

Ablation Study

- The performance of each module of the PS-PC framework is measured individually.
- The performance of PS-PC is measured by replacing the classifier with a random classifier.

Module	Metric	Models				
		$\eta = 0.97$	$\eta = 0.98$	$\eta = 0.985$	$\eta = 0.99$	$\eta = 0.995$
Classifier only	PLCC	0.67	0.85	0.91	0.93	0.94
	SROCC	0.44	0.60	0.63	0.68	0.75
Predictor only	PLCC	0.85				
	SROCC	0.83				
RndClass+predict	PLCC	0.89	0.88	0.89	0.87	0.88
	SROCC	0.85	0.85	0.85	0.86	0.86
PS-PC	PLCC	0.92	0.96	0.97	0.98	0.99
	SROCC	0.87	0.90	0.92	0.922	0.94

- Both classifier and predictor play an important role in the overall performance
- Random classifier and predictor barely improve the performance compared with the predictor only case

Cross Dataset Evaluation

- The performance of the PS-PC framework is measured using TID2013, and PieAPP dataset

Dataset	Metric	Models				
		$\eta = 0.97$	$\eta = 0.98$	$\eta = 0.985$	$\eta = 0.99$	$\eta = 0.995$
TID2013	PLCC	0.83	0.85	0.90	0.92	0.95
	SROCC	0.82	0.83	0.90	0.93	0.95
	Defer Pairs	7%	17%	35%	44%	62%
PieAPP	PLCC	0.80	0.84	0.86	0.89	0.89
	SROCC	0.44	0.49	0.53	0.60	0.62
	Defer Pairs	9%	22%	26%	38%	46%
PC-IQA	PLCC	0.92	0.96	0.97	0.98	0.99
	SROCC	0.87	0.90	0.92	0.922	0.94
	Defer Pairs	8%	11%	15%	17%	22%

- Performance is lower for TID2013 and PieApp but not very significantly.

5. Conclusions

- PS-PC framework outperforms relevant state-of-the-art
- Both predictor and classifier contribute to the final performance of the proposed solution.
- The proposed sampling algorithm is acquired a priori to the subjective test.