Active Sampling Exploiting Reliable Informativeness for Subjective Image Quality Assessment Based on Pairwise Comparison

Zhiwei Fan, Tingting Jiang, *Member, IEEE*, and Tiejun Huang, *Senior Member, IEEE*

***Abstract*—Subjective image quality assessment (IQA) based on pairwise comparison (PC) overcome the shortcomings of IQA based on category rating, such as an ambiguous scale definition. However, the testing scale of PC tests can be very large, as the number of image pairs for comparison is a quadratic form of the number of images. To conduct PC tests on a large-scale image set with limited budget, an active sampling strategy to reduce testing scale is required. The conventional active sampling strategies usually select the most informative sample and assume that any image pair's correct label can be obtained from any subjects who are attentive. However, this is not true for IQA, because of human visual system's limitation. If two images are similar, their difference can be too subtle for some subjects to perceive. It means that it takes subjects more effort to obtain correct preference labels of two similar images, and that it is even impossible to obtain the correct preference labels of two images that are too similar. To address this issue, we study the reliability of preference labels. Based on the combination of reliability and informativeness, we design a new active sampling framework. It not only considers the informativeness, but also adjusts the effort spent on an image pair according to its ambiguity. Experiments show that this adjustment can effectively improve the performance of sampling strategies only based on informativeness. Besides, the proposed method is expected to be applied to more general subjective tests based on PC beyond IQA.**

***Index Terms*—Active sampling, pairwise comparison, quality of experience, subjective image quality assessment (IQA).**

I. Introduction

With the development and popularization of multimedia technologies, people nowadays are enjoying more and more multimedia services and contents. And the evaluation of these services or contents' Quality of Experience (QoE) has become an important topic. For instance, subjective image quality assessment (IQA), which studies the QoE of images, is of great importance to set gold-standard quality measures of images.

There are mainly two classes of methods for subjective IQA tests [1]. One is based on Category Rating (CR) test [1], [2], and the other is based on Pairwise Comparison (PC) test [1], [3]. Some work combined these two methods as well [4]. The CR test is a category judgement where the test images are rated on a category scale, including absolute category rating (ACR), absolute category rating with hidden reference (ACR-HR) [1] and degradation category rating (DCR) [1]. The CR test is widely used because it is simple and intuitive. However, its drawback is also obvious [5]. Significantly, it is likely to cause subjects' confusion about the rating scale, even including DCR where reference images exist. Subjects are not able to exactly repeat their prior scores for items they have rated [6]. In contrast to the CR test, the PC test, which is also popular [7], is of high discriminatory power. And it is particularly valuable when some of the test items are very similar in quality. Some experiments based on simulated data have investigated the behaviour of results of the PC test [8]. On the other hand, as there are $\binom{N}{2}$ image pairs for N images, the PC test costs much more time than the CR test, which only takes $O(N)$ time. When N is large, it is usually beyond budget to have every image pair fully compared. Therefore, a wise sampling strategy, which selects valuable image pairs for comparison, plays an important role. There have been many research efforts devoted to this direction [9]–[17]. A flow chart illustrating how these sampling methods work for such a subjective IQA test is as shown in Fig. 1. There is an image set for quality assessment. During the subjective test, pairs of images are selected from the image set sequentially through some sampling strategy for comparison. At the same time, the preference labels are collected from subjects and then translated into

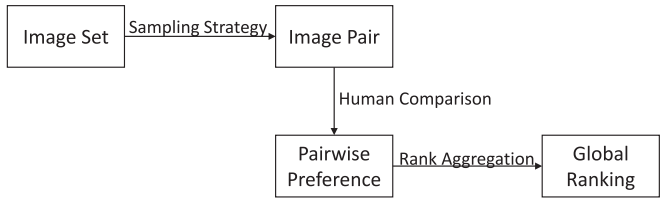

Fig. 1. Paradigm of a subjective IQA test based on PC. In the test, image pairs are selected from the image set through a sampling strategy. Then the selected image pair is judged by subjects, and the pairwise preference provided by the subjects is used to update the images' global ranking through a rank aggregation algorithm.

Manuscript received November 30, 2016; revised March 5, 2017 and May 15, 2017; accepted May 20, 2017. Date of publication June 5, 2017; date of current version November 15, 2017. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2015CB351803, and in part by the Natural Science Foundation of China under Grant 61572042, Grant 61390514, Grant 61421062, Grant 61210005, and Grant 61527084. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abdulmotaleb El Saddik. *(Corresponding author: Zhiwei Fan.)*

The authors are with the Institute of Digital Media, School of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Cooperative Medianet Innovation Center, Shanghai 200240, China (e-mail: fanzw@pku.edu.cn; ttjiang@pku.edu.cn; tjhuang@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2711860

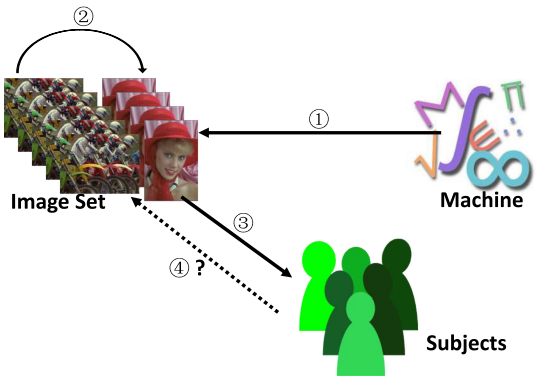

Fig. 2. System of subjective IQA, consisting of three components, the image set, the machine (rank aggregation algorithm), and the subjects.

a global ranking of all the images through a rank aggregation algorithm. The final goal is to obtain an as accurate as possible global ranking of the images with a limited budget for testing.

To achieve an efficient sampling strategy, there have been several different approaches proposed from different perspectives. As shown in Fig. 2, a system of subjective IQA test is made up of three components, the images for test, the subjects providing preference labels, and the machine (*e.g.,* the preference aggregation algorithm) for modeling. Each arrow in Fig. 2 represents one approach of active sampling. The component indicated by the head of the arrow is what is actively selected, while the component indicated by the tail of the arrow is the basis of the active sampling. To be specific, the arrow ① shown in Fig. 2 represents one approach [10]–[15] working on selecting images from the perspective of machines. Such an approach usually designs an uncertainty/information measure for samples based on a specific machine and therefore select the most uncertain/informative sample for annotation. Another approach considers the information density measure [10]–[12], leading subjects to compare image pairs in different regions of the sample space, which means that it selects image pairs from the perspective of the image set as the arrow ② shown in Fig. 2. Both of the above approaches focus on the difference between samples. Combining these two approaches, the conventional active sampling strategies define the informativeness of image pairs *assuming that their correct preference labels can be obtained from any attentive subjects*. Besides, there is a different approach, which works on selecting reliable subjects who are either attentive or of high expertise. Reliable subjects are defined as subjects who provide preference labels consistent with the majority. Thus the approach selects subjects from the perspective of the image set [15]–[17], as the arrow ③ shown in Fig. 2. This approach focuses on the difference between subjects and evaluates the quality of subjects in order to improve the efficiency of subjective tests. However, all the above three approaches ignore the arrow ④in Fig. 2, which is how to evaluate the difference between samples from the perspective of human subjects. As we all know, there are limitations of human visual system (HVS). In the subjective IQA task, some limitation is extremely important, as Just Noticeable Difference (JND) theory [18] tells that human can only perceive the difference between two images above some threshold. This implies that if the difference between two images is below the threshold of a subject, the subject is not able to definitely provide the correct preference label. This incorrect label is not noise caused by subjects' carelessness, but their limitation of vision. Therefore, the more similar two images are, the more likely a subject is not able to tell the difference, and the more subjects are needed to obtain the correct label. Extremely, for two very similar images like Fig. 3(c), even increasing the number of subjects to judge them does not help. This fact contradicts the assumption taken by the conventional active sampling strategies, which is as long as the image pair is sampled, its ambiguity will be resolved eventually.

To solve this problem, a wise sampling strategy should consider the difference between images from the perspective of human and spend appropriate effort on an image pair accordingly. The optimal effort to obtain an image pair's correct preference label varies with the similarity of the two images. Specifically speaking, for two images that differ a lot in quality like the two images in Fig. 3(a), it is very easy for subjects to judge, and we can acquire the correct preference label of the two images from a small number of subjects. On the contrary, for two images that are similar in quality, like the two images in Fig. 3(b), it might be somewhat confusing for subjects to judge. A correct preference label should be obtained based on more subjects' judgements than that of the images in Fig. 3(a). In extreme cases, for two images that are very similar in quality, such as the two images in Fig. 3(c), even an excellent expert in image quality is not able to distinguish them, and it is just beyond limitation of human vision and cognition. Thus it is impossible to get the correct preference label of their quality, which means having subjects compare such extremely similar image pairs makes no contribution to obtaining the global ranking of the total image set. On the other hand, without predicting the effort it costs to obtain the correct preference label, we may spend too much effort on two images as Fig. 3(a), too little effort on two images as Fig. 3(b), or useless effort on two images as Fig. 3(c), none of which is wise. Therefore, the factor of human subjects should be considered in selecting image pairs for comparison, as it is directly related to the effort it costs to obtain image pairs' correct preference labels.

In this paper, to optimize the effort spent on each image pair, we introduce a probabilistic model to estimate the probability that a subject is able to distinguish an image pair of a certain similarity. With this estimation, the reliability of an image pair's preference can be calculated according to two parameters. One is the number of times that the image pair has been compared, and the other one is the similarity between the two images. Note that this reliability is not to measure the quality of individual subjects, but to measure the quality of the preference labels collected for an image pair during the PC test, considering the similarity of the two images. Taking the reliability and informativeness both into consideration, we propose an active sampling framework based on expected reliable informativeness gain. An image pair's reliable informativeness is a combination of the image pair's reliability and informativeness. And its expected reliable informativeness gain is the gain

Fig. 3. Three image pairs of different similarities. (a) Two images that differ a lot in quality, and subjects can easily judge the preference. (b) Two images that are similar in quality, and it may take subjects some effort to judge the preference. (c) Two images that are very similar, which are beyond most humans' ability to judge. (a) Low similarity. (b) Medium similarity. (c) High similarity.

of the image pair's reliable informativeness if the two images are compared by one more subject. By actively sampling the image pair of the largest expected reliable informativeness gain, the proposed active sampling framework overcomes the shortcomings of previous active sampling strategies purely based on informativeness. The results of the experiments based on two different ranking algorithms show that the consideration of reliability does improve the performance of the sampling strategies purely based on informativeness.

The proposed strategy can adapt to both the laboratory environment and the crowdsourcing environment. As it can effectively reduce the testing scale of a large testing image set, and it considers ordinary unspecific subjects' limitation to make correct judgements, the proposed strategy is specially suitable for crowdsourcing environment. It should be noted that although this paper investigates the problem of IQA, the proposed active sampling strategy can be applied to more general evaluation problems based on pairwise comparison. It can be directly ap-

plied to the QoE assessment of general multimedia contexts and applications, including images, videos, games and so on. And it can even be applied to the evaluation of higher-level properties such as relative attribute [10] and aesthetic value [19].

The remainder of the paper is organized as follows. Section II discusses the related work. Section III concretely illustrates the proposed active sampling strategy, including the definition of reliable informativeness in Section III-B, the estimation of reliability in Section III-C, and the selection of informativeness in Section III-D. Section IV illustrates the settings and the results of the conducted experiments. And Section V is a conclusion of the whole paper.

II. Related Work

A. Crowdsourced QoE

Compared to conventional QoE assessment test conducted in laboratory environment, crowdsourced test conducted via the Internet is cheaper and more time-saving, and it enjoys a large and diverse panel of subjects across international and regional block, who are much closer to the realistic users. With the popularization of paid crowdsourcing platform such as Amazon Mechanical Turk, the interest of conducting QoE assessment test by crowdsourcing is rising. Hossfeld *et al.* [20] talked about the key issues of moving from the laboratory to the crowd, particularly including the reliability of user ratings, and a collection of best practices addressing these issues are provided, which demonstrates the feasibility and advantage of crowdsourced QoE assessment test. Ribeiro *et al.* [21] performed the ACR test using crowdsourcing, taking the detection of inaccurate scores into consideration. As discussed above, one of the crowdsourcing test's shortcomings is that subjects are more likely to provide inaccurate ratings because of their inexpertise and the less controlled environment. Therefore, some researchers tried to conduct PC tests in crowdsoucing environment, as PC test is of high discriminatory power, and it is much easier for ordinary subjects. Chen *et al.* [5] proposed a crowdsourceable QoE assessment framework based on pairwise comparison. And they later designed a crowdsourcing platform for QoE assessment based on pairwise comparison [22]. However, Chen *et al.*'s work is based on a complete set of pairwise comparison. When the scale of an image set is large, the cost of comparison is high even by crowdsourcing. Therefore, this method does not suit the assessment of large image sets. In that case, Xu *et al.* [23], [24] introduced the HodgeRank, which is able to recover the ordinal rank from an uncomplete and unbalanced set of pairwise comparisons. Xu *et al.*'s work increases the likelihood of conducting PC tests on large scale datasets. But its testing scale can still be large. One way to further reduce the testing scale is to replace the random sampling strategy with active sampling. Our proposed strategy, particularly considering subjects' inexpertise and ratings' reliability, can adapt to the crowdsourcing environment well.

B. Preference Aggregation

The problem of ranking through pairwise preferences has been investigated in many fields, including information retrieval [25], recommender systems [26], social opinions [27], and so on. As the problem of obtaining the optimal global ranking is a well known non-deterministic polynomial-time hard (NP-hard) problem called minimum feedback arc-set in tournaments (MFAST) [28], various approximation algorithms are proposed to get the practical solution, such as Borda Count [29], HodgeRank [30] and rank centrality [31]. Moreover, with the development of machine learning, there come many rank algorithms based on learning, which are called "Learning to Rank" (LTR) algorithms, such as RankSVM [32], [33] and RankNet [34]. Because of the generalization ability of learning based methods, LTR algorithms are usually able to obtain a practical ranking with fewer preferences than non-learning based rank aggregation algorithms. However, the performance of the LTR algorithms depends on the design of the features. They are not suitable for the case when the aim of the subjective test is to obtain the gold-standard label for a newly studied problem. For instance, in subjective IQA, if the aim is to get subjective quality of some kind of images, whose quality has not been sufficiently studied, there can be no appropriate features available to represent the quality of the new kind of images, and LTR algorithms based on learning cannot not apply.

C. Active Sampling

As human comparison is cost and time consuming, fully comparing every pair of images in a large-scale dataset is often beyond the budget, and series of active sampling strategies are proposed to reduce the number of preference labels required for ranking. Several active ranking algorithms have proposed the upper bound of the sampling complexity when regret of loss exists. Ailon [35] proposed an algorithm that queries at most $O(\varepsilon^{-6} n \log^5 n)$ preference labels for a regret of ε times the optimal loss. Jamieson *et al.* [36] reduced the sampling complexity to $O(d \log n)$, assuming each object is embedded into a d-dimensional Euclidean space and that the rankings reflect their relative distances from a common reference point in R^d.

The works of active sampling strategies are usually divided into two categories. The first selects a pair of images that are informative, and the second selects subjects that are reliable or of high expertise. As far as we know, most works on active sampling are designed for algorithms based on learning, which is called active learning, while active sampling strategies applied for non-learning based algorithms are rarely considered.

The strategy of selecting most informative samples in active learning is usually driven by two measures, uncertainty measure and information density measure. The uncertainty measure is an "exploitation" strategy [11], [12], leading subjects to annotate the samples near the boundary, which in return helps to refine the boundary. The boundary here means the region in the sample space that the learner is most uncertain about, such as the hyperplane in the problem of classification. The information density measure is an "exploration" strategy [11], [12], leading subjects to annotate samples in different regions of the sample space, which helps to avoid incorrectly predicting samples in some region. Defining the informativeness of each sample as one or a combination of these two measures, a variety of active learning strategies are successfully applied in different fields, including

annotation [9], [37], recognition [13], [14], retrieval [38], *etc*. Similarly, for the problem of ranking through preference, the informativeness of a pair is defined in this way in some active learning strategies as well. For instance, Chen *et al.* [15] defined the informativeness of each pair as the uncertainty of the pair's preference label based on the Bradley-Terry Model [39]. Liang *et al.* [10] combined the two measures in the ranking of relative attribute, by selecting to compare samples of the lowest margin, with the constraint that the selected two samples are from different clusters. This work has considered the problem of comparing two very similar images that are beyond subjects' ability, and the constraint, which was a variant of information density measure, can be partly used for solving this problem.

Another important strategy to reduce the number of needed preference labels is to query reliable subjects. The prediction of subjects' reliability or expertise has been studied in many works. Long *et al.* [17] modeled both the overall annotation noises and the expertise level of each individual annotator, which helps to find reliable annotator. Hua *et al.* [16] presented annotating quality control of each annotator. For the problem of ranking through preference, Chen [15] asked subjects to compare pairs of large margins at the beginning of the test and evaluated the reliability of subjects by calculating the consistency of each subject's answers with the final preference label.

Although the consideration of information density measure can partly solve the problem of having subjects compare two images that are too similar, such as Liang *et al.* [10], its effectiveness heavily depends on the information density measure and the feature space it adopts, which implies that it is only suited for ranking algorithms based on learning. Besides, the previous active learning strategies rarely consider the difference between image pairs from the perspective of human vision system (HVS), which introduces reliability problem for the subjective preference labels. Here the reliability is different from the reliability of subjects in previous work [15]–[17]. They focus on the difference between subjects while we take subjects as a group and focus on the credibility of the annotation for an image pair given by this group according to the similarity of the two images. By introducing reliability into an image pair's informativeness, our proposed active sampling framework solves the above problems. Furthermore, our method can be applied to both LTR algorithms and non-learning based algorithms.

TABLE I
DESCRIPTIONS OF THE VARIABLES USED IN THIS PAPER

Variable	Description
X	$X = \{x_1, x_2, \ldots, x_N\}$ is the image set for quality assessment.
N	The number of the images in the image set X.
x_i	An image in the image set X $(1 \leq i \leq N)$.
C	The set of all the image pairs that are comparable.
$s(x_i)$	The true quality score of image X_i $(1 \leq i \leq N)$.
$s^t(x_i)$	The estimated quality score of image x_i after the tth round of the test.
n_{ij}^t	The number of times that image x_i is annotated better than x_j after the tth round of the test.
n_{ij}	The number of times that image x_i is annotated better than x_j.
MR^t	The miss ratio of the test after the tth round of the test
$\widehat{MR^t}$	An approximation of MR^t used in the real application.
$I(x_i, x_j)$	The informativeness of image pair (x_i, x_j).
$R(x_i, x_j, n)$	The reliability of image pair (x_i, x_j)'s preference label with n comparisons.
$\Delta R(x_i, x_j, n)$	The reliability gain of image pair (x_i, x_j)'s preference label with n comparisons, if it is to be compared one more time.
$\Delta R^t(x_i, x_j)$	The reliability gain of image pair (x_i, x_j)'s preference label if it is to be compared one more time in the tth round.
$I_r(x_i, x_j, n)$	The reliable informativeness of image pair (x_i, x_j) when it is compared n times.
$\Delta I_r(x_i, x_j, n)$	The reliable informativeness gain of image pair (x_i, x_j)'s preference label with n comparisons, if it is to be comparedone more time.
$\Delta I_r^t(x_i, x_j)$	The reliable informativeness gain of image pair (x_i, x_j)'s preference label, if it is to be compared one more timein the tth round.
$P_c(x_i, x_j)$	The probability that a subject provides the correct preference label of image pair (x_i, x_j).
$P_d(x_i, x_j)$	The probability that a subject is able to distinguish the quality difference of image pair (x_i, x_j).
JND	Just Noticeable Difference, the minimum quality difference that subjects can distinguish.
λ, k	The two coefficients of Weibull Distribution.
$I^0(x_i, x_j)$	A general measure of informativeness of image pair (x_i, x_j).
V_{ij}	A random variable indicating the correctness of a subject's preference answer of image pair (x_i, x_j).
V_{ij}^n	The percentage of correct answers among n comparisons of x_i and x_j.
M	The minimum number of a good representative of all the potential subjects.

III. THE PROPOSED METHODS

A. Formulation

Suppose that the subjective image quality assessment task is conducted on an image dataset, $X = \{x_1, x_2, \ldots, x_N\}$, where there are N images, and x_i $(1 \leq i \leq N)$ represents an image in X. The goal of the subjective image quality assessment task is to obtain the global ranking of all the images of X, using a ranking through pairwise preference algorithm. Assume that for $\forall x_i \in X$, the true quality score of x_i is $s(x_i)$, representing x_i's gold-standard quality, whose exact value is unavailable with limited tests. Higher scores mean better qualities. Quality scores are calculated by the ranking algorithm iteratively with the annotating task going on. The initial prediction of $s(x_i)$, denoted as $s^0(x_i)$, is set according to the ranking algorithm applied. In each round of the test, a pair of two images, x_i and x_j, are selected for comparison, and subjects are supposed to provide a binary answer, x_i better than x_j or x_i worse than x_j, which is regarded as a preference label. The evaluation criteria of images is set according to the aim of the annotating task. Usually in image quality assessment [40], and also in our later experiments, an image is more preferred than another when the image is better in quality, which means that it is clearer and less distorted. With this new preference label, the values of images' quality scores are updated through the ranking algorithm applied. Assume that $s^t(x_i)$ is the value of image x_i's quality score generated from a preference aggregation algorithm (e.g., Bradley-Terry model [39]) after the tth round, when t preference labels have been collected. The variables to be used in this paper are shown in Table I.

Our goal is to design an active sampling strategy to reduce the number of pairwise comparisons needed as much as possible, without decreasing the accuracy of the global ranking. The strategy's performance is evaluated by measuring two parameters: the number of pairwise comparisons queried, and the miss ratio (MR). MR is the percentage of mismatched pairs between the subjective comparison results and global rankings. The miss ratio after the tth round of the test is defined as follows where minimization is the goal of our strategy [41]:

$$MR^t = \frac{\sum_{(x_i,x_j)\in C} mis^t(x_i,x_j)}{Card(C)} \tag{1}$$

$$mis^t(x_i,x_j) = \begin{cases} \mathbf{1}_{\mathbf{s(x_i)\geq s(x_j)}}, & \text{if } s^t(x_i) < s^t(x_j) \\ \mathbf{1}_{s(x_i)\leq s(x_j)}, & \text{if } s^t(x_i) > s^t(x_j) \\ \mathbf{1}_{s(x_i)\neq s(x_j)}, & \text{if } s^t(x_i) = s^t(x_j) \end{cases} \tag{2}$$

$$\mathbf{1}_{f(\cdot)} = \begin{cases} 1, & \text{if } f(\cdot) = true \\ 0, & \text{if } f(\cdot) = false \end{cases} \tag{3}$$

where C is the set of all the image pairs that are comparable and $Card(C)$ represents the cardinality of the set C. $\mathbf{1}$ is the indicator function indicating if the judgement $f(\cdot)$ equals to $true$. Specifically, the definition of C depends on the annotating task. For example, in most image quality assessment experiments based on images' distortion such as TID2013 [40], only the images of the same content are to be compared, in which case, $C = \{(x_i,x_j)|\ 1 \leq i < j \leq N$, x_i and x_j are of the same content.$\}$. $mis^t(x_i,x_j)$ represents whether the predicted score function values $s^t(x_i)$ and $s^t(x_j)$ correctly show the relative qualities of x_i and x_j. And MR^t is a measure of s and s^t's rank correlation. It can be seen as the percentage of the image pairs whose preference labels are not correctly predicted after tth round of the test among the image pairs in C. The smaller MR^t is, the higher s and s^t are rank correlated. The ideal value of MR^t is 0, which means that the global ranking of all the images according to their predicted quality score s^t, is totally the same as their true global quality ranking according to s.

B. Reliable Informativeness

Different image pairs contain different amount of information. For image pair (x_i,x_j), the amount of information it contains can be presented as its informativeness $I(x_i,x_j)$, similar to the concept of informativeness in previous work [10]–[15]. To get the correct preference label of an image pair, the image pair is supposed to be compared several times, and then the final preference label is obtained by majority voting. However, with a limited number of comparisons, it is difficult to guarantee the correctness of the final preference label. Thus, we introduce the reliability of an image pair's preference label, which is the probability that a certain number of comparisons offer the correct preference label of an image pair by majority voting. The accurate definition of the reliability R is as

$$\begin{aligned} &R(x_i,x_j,n) \\ &\quad = P(s(x_i) > s(x_j), n_{ij} > n_{ji} | n_{ij}+n_{ji} = n) \\ &\qquad + P(s(x_i) < s(x_j), n_{ij} < n_{ji} | n_{ij}+n_{ji} = n) \\ &\qquad + P(s(x_i) = s(x_j), n_{ij} = n_{ji} | n_{ij}+n_{ji} = n) \end{aligned} \tag{4}$$

where n is the number of times that the images x_i and x_j have been compared. n_{ij} is the number of times that x_i is preferred to x_j among the n comparisons, while n_{ji} is the number of times that x_j is preferred to x_i. The first term in the right of (4) represents the probability that the quality of x_i is better than x_j and that n comparisons offer the correct preference label by majority voting. The second and the third terms in the right of (4) represent the probabilities in a similar way. Therefore, as the sum of these three terms, the reliability $R(x_i,x_j,n)$ is the probability that n comparisons can offer the correct preference label of image pair (x_i,x_j) by majority voting.

With the informativeness $I(x_i,x_j)$ and the reliability $R(x_i,x_j,n)$ defined, there comes the reliable informativeness I_r, shown as

$$I_r(x_i,x_j,n) = R(x_i,x_j,n) \times I(x_i,x_j). \tag{5}$$

When the preference label of image pair $(x_i,\ x_j)$ is correct, the informativeness it contains can be fully obtained, which is $I(x_i,x_j)$. Otherwise, we assume that the informativeness it contains cannot be obtained at all, which is 0. Therefore, the reliable informativeness $I_r(x_i,x_j,n)$ can be seen as the expectation of the available informativeness of image pair $(x_i,\ x_j)$, when x_i and x_j are compared n times. In previous work [10]–[15], subjects are usually assumed to be always able to provide the correct label if they are attentive, which means $R(x_i,x_j,n) = 1$. Under this assumption, $I_r = I$, and the informativeness used in the previous work can be seen as a special case of our reliable informativeness. Our proposed active sampling strategy is designed based on maximizing reliable informativeness I_r, rather than informativeness I. In this way, our strategy is able to adjust the effort spent on different image pairs, and to avoid having subjects compare two images that are too similar to distinguish, which will be concretely illustrated in Section III-E. The estimation of reliability R and the selection of informativeness I are to be introduced respectively in Section III-C and Section III-D.

C. Estimating Reliability

1) Reliability Calculating: As it makes no sense to compare two images that are totally the same, we assume that any two of the images in X are not the same. When an image pair has been compared an odd number of times, its reliability R defined as (4) can be calculated as the probability that the number of correct answers is larger than that of the wrong among n comparisons. Therefore, when n is odd, $R(x_i,x_j,n)$ can be calculated as

$$R(x_i,x_j,n) = \sum_{a>n-a}^{a\leq n} \binom{n}{a} (1-P_c(x_i,x_j))^{n-a} P_c(x_i,x_j)^a \tag{6}$$

where $P_c(x_i,x_j)$ is the probability that a subject provides the correct preference label when comparing images x_i and x_j, and a is the number of correct answers among n pairwise comparisons, which should be larger than the number of wrong answers $n-a$.

When an image pair has been compared an even number of times, which means $n = 2m$, it is vague to deal with the condition that $n_{ij} = n_{ji}$, because majority voting does not

work. Since $R(x_i, x_j, n)$ is expected to be smooth with n increasing, here we simply approximate $R(x_i, x_j, 2m)$ as the average of $R(x_i, x_j, 2m-1)$ and $R(x_i, x_j, 2m+1)$. When an image pair has never been compared, the preference label has to be randomly set, in which case, the reliability is $\frac{1}{2}$. Therefore, the entire calculation of the reliability can be formulated as

$$R(x_i, x_j, n) = \begin{cases} \frac{1}{2}, & n = 0 \\ \sum_{a>n-a}^{a\le n} \binom{n}{a}(1 - P_c(x_i, x_j))^{n-a} P_c(x_i, x_j)^a, & n \text{ is odd} \\ \big(R(x_i, x_j, n-1) + R(x_i, x_j, n+1)\big)/2, & \text{otherwise.} \end{cases} \tag{7}$$

As for the computation of $P_c(x_i, x_j)$, $P_d(x_i, x_j)$ will be used, which is the probability that a subject is able to distinguish the relative qualities of x_i and x_j. When a subject is able to distinguish two images, we assume that he/she will provide the correct answer, which means that the probability that the label is correct is 1. Otherwise, when a subject is not able to distinguish two images, he/she will randomly provide a binary answer with $\frac{1}{2}$ probability to be correct. Therefore, $P_c(x_i, x_j)$ can be calculated as

$$\begin{aligned} P_c(x_i, x_j) &= P_d(x_i, x_j) \times 1 + (1 - P_d(x_i, x_j)) \times \frac{1}{2} \\ &= (1 + P_d(x_i, x_j))/2. \end{aligned} \tag{8}$$

According to the theory of Just Noticeable Difference (JND)[18], people are able to tell the difference between two images only when the difference between them exceeds some threshold. Therefore, a subject is able to distinguish the relative qualities of x_i and x_j only when the quality difference between them, $|s(x_i) - s(x_j)|$, exceeds the threshold of JND. Here we take JND as a variant, and $P_d(x_i, x_j)$ can be calculated as

$$P_d(x_i, x_j) = P(\text{JND} < |s(x_i) - s(x_j)|). \tag{9}$$

2) JND Modeling: According to (6)–(9), the key of calculating R is to obtain the distribution of JND. The first thought of modeling JND is to use Gaussian distribution, since there are no prior assumptions about JND. Assume that JND obeys a Gaussian distribution, which means JND $\sim \mathcal{N}(\mu, \sigma^2)$. μ and σ are the mean value and standard deviation of the Gaussian distribution respectively. Thus, according to (8) and (9), $P_c(x_i, x_j) = (1 + P_d(x_i, x_j))/2 = \frac{1}{2} + \frac{1}{2}\Phi(\frac{|s(x_i) - s(x_j)| - \mu}{\sigma})$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Therefore, $P_c(x_i, x_j) \to \frac{1}{2} + \frac{1}{2}\Phi(-\frac{\mu}{\sigma}) > \frac{1}{2}$, when $|s(x_i) - s(x_j)| \to 0$. This is inconsistent with the fact because $|s(x_i) - s(x_j)| \to 0$ means that images x_i and x_j are very similar, in which case, subjects can only randomly provide a preference label, leading to $P_c(x_i, x_j) \to \frac{1}{2}$. Therefore, Gaussian distribution is not a good choice theoretically.

To better model the distribution of JND, we introduce the Weibull distribution [42] in our work, which is the most commonly used distribution for modeling reliability. The Weibull distribution can model the life time of different products. Here the scenario that the quality difference between two images becomes visible with quality difference increasing, can be simulated as the process that a product fails with time going on. And here JND can be treated as the life time of a product, if quality difference is treated as the time. Above all, the JND can be modeled as

$$\begin{aligned} P(\text{JND} < |s(x_i) - s(x_j)|) &= F_{\text{Weibull}}(|s(x_i) - s(x_j)|) \\ &= 1 - e^{-(|s(x_i) - s(x_j)|/\lambda)^k} \end{aligned} \tag{10}$$

where λ and k are coefficients of Weibull distribution.

Combining (8), (9) and (10), we arrive at the expression of P_c as

$$P_c(x_i, x_j) = 1 - \frac{1}{2} e^{-(|s(x_i) - s(x_j)|/\lambda)^k}. \tag{11}$$

Therefore, $P_c(x_i, x_j)$ can be estimated, once the two coefficients λ and k are estimated. From (11), we can see that when images x_i and x_j are very similar, which means $|s(x_i) - s(x_j)| \to 0$, subjects can only randomly provide a preference label, as they can hardly tell the relative qualities between x_i and x_j. In such cases, $P_c(x_i, x_j) \to \frac{1}{2}$. When images x_i and x_j differ a lot in quality, which means $|s(x_i) - s(x_j)| \to +\infty$, subjects can very easily tell the preference between the two images. In that case, subjects can surely provide a correct preference label, and $P_c(x_i, x_j) \to 1$.

3) Coefficients Training: As illustrated above, for an image pair (x_i, x_j), there exists a relation between $|s(x_i) - s(x_j)|$ and $P_c(x_i, x_j)$ as (11). Therefore, the coefficients λ and k of the Weibull distribution can be obtained by curve fitting over a number of $(|s(x_i) - s(x_j)|, P_c(x_i, x_j))$ pairs.

As for the calculation of the quality difference $|s(x_i) - s(x_j)|$, since the true quality score $s(x_i)$ and $s(x_j)$ are not available, we can approximate them with $s^t(x_i)$ and $s^t(x_j)$ in the $(t+1)$th round of the test, which are the estimated values of x_i and x_j's quality scores.

As for the calculation of the probability $P_c(x_i, x_j)$, it can be treated as a frequency problem. Defined as the probability that one subject provides the correct answer, $P_c(x_i, x_j)$ can be calculated as the percentage of the correct answers among all the comparison answers provided by all the potential subjects. As we cannot afford to have all the potential subjects compare an image pair, we select a set of subjects to represent all the potential subjects. Assume that a randomly selected and big enough subject set can be a good representative of all the potential subjects. In that case, $P_c(x_i, x_j)$ can be calculated as the percentage of the correct answers among all the comparison answers provided by the representative subjects. Recall that n_{ij} represents the number of subjects that prefer image x_i to x_j, while n_{ji} represents the number of subjects that prefer image x_j to x_i. As $P_c(x_i, x_j) \ge 1/2$ according to (8), the majority of the $n_{ij} + n_{ji}$ will provide the correct answers. Thus, the number of the correct answers among the totally $n_{ij} + n_{ji}$ comparisons is $\max\{n_{ij}, n_{ji}\}$. Therefore, $P_c(x_i, x_j)$ can be calculated as

$$P_c(x_i, x_j) = \frac{\max\{n_{ij}, n_{ji}\}}{n_{ij} + n_{ji}}. \tag{12}$$

To verify whether JND obeys Weibull distribution in statistics, we use it to predict P_c of all the image pairs in Database1

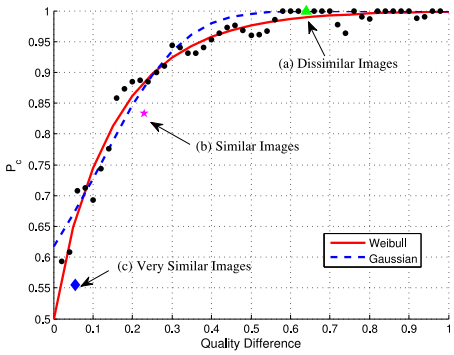

Fig. 4. Each small black point in this figure represents the average P_c of all the image pairs of a certain quality difference. The red full line and the blue dotted line respectively represent the curves fitting all these points by Weibull distribution and Gaussian distribution. The points representing the three image pairs in Fig. 3, (a) low similarity (blue diamond), (b) medium similarity (pink pentagram). (c) high similarity (green triangle), are shown in the figure as well.

introduced in Section IV-A. The RMSE (Root Mean Square Error) is 0.167. To make a comparison, we have also tried Gaussian distribution, in which case, the RMSE is 0.211. Therefore, we can see that Weibull distribution predicts P_c better than Gaussian. Fig. 4 shows the two curves that fit the quality difference and the P_c of image pairs in Database1 with Weibull distribution and Gaussian distribution respectively. The points representing the three image pairs in Fig. 3 are also shown in Fig. 4. Each of these image pairs has been compared 24 times.

We call an image pair that is compared by a good representative of all the potential subjects a sufficiently compared image pair. The values of coefficients λ and k are iteratively updated with the number of the sufficiently compared image pairs increasing as the testing rounds go on. If there exist some sufficiently compared image pairs before the test, they can be used for pretraining the initial values of coefficients λ and k.

4) Sufficient Comparison: In real applications, we assume that an image pair is sufficiently compared if it is compared by at least a certain number (M) of randomly selected subjects. As for the value of M, we set $M = 5$ in our work. Assume that V_{ij} is a random variable indicating the correctness of a subject's preference answer, defined as

$$V_{ij} = \begin{cases} 1, & \text{if a subject provides the correct preference of } x_i \text{ and } x_j \\ 0, & \text{if a subject provides the wrong preference of } x_i \text{ and } x_j. \end{cases} \tag{13}$$

Thus, V_{ij} follows Bernoulli distribution, and the variance of V_{ij} is

$$Var(V_{ij}) = P_c(x_i, x_j) \times (1 - P_c(x_i, x_j)). \tag{14}$$

Assume that V_{ij}^n is the percentage of correct answers among n comparisons of x_i and x_j. Thus

$$V_{ij}^n = \frac{1}{n} \sum_{i=1}^{n} V_{ij} \tag{15}$$

which comes to the variance of V_{ij}^n

$$Var(V_{ij}^n) = \frac{1}{n} P_c(x_i, x_j) \times (1 - P_c(x_i, x_j)) \leq \frac{1}{4n}. \tag{16}$$

Therefore, when $n \geq 5$, $Var(V_{ij}^n) \leq 0.05$, which means that $M = 5$ subjects can be a quite good representative of all the potential subjects.

To verify the correctness of setting $M = 5$, an experiment is made on Database1. We randomly select 5 comparisons of each image pair from the 24 comparisons, with which we can calculate the P_c of each image pair according to (12), represented as $P_c^{(5)}$. And with the totally 24 comparisons, we can calculate the P_c of each image pair, represented as $P_c^{(24)}$. The MSE (Mean Square Error) of $P_c^{(5)}$ and $P_c^{(24)}$ is 0.0127, which shows that the value of $P_c^{(n)}$ does not change much with n increasing from 5 to 24. Based on the above two observations, we think $M = 5$ is a reasonable choice in our experiment.

D. Selecting Informativeness

The informativeness of an image pair x_i and x_j, $I(x_i, x_j)$, represents the image pair's worth for comparison on the assumption that subjects are to provide correct preference labels. There are various designs of informativeness measures as discussed in Section II-C. The existing informativeness measures are usually a combination of uncertainty measures and information density measures, which means that they are a trade-off between exploration and exploitation. Some of them are designed for annotating tasks like classification and recognition, instead of ranking. However, taking pairwise comparison as binary classification, these measures still work for many LTR algorithms like RankSVM [32], [33]. Above all, most of the informativeness measures designed in previous works can be an option for the informativeness measure of image pairs in the subjective IQA problem.

However, as the above informativeness measures are initially designed for the algorithms based on learning, it is difficult to apply them into IQA tests based on non-learning ranking algorithms, which are not rare, especially when there are no features or learners suitable for the quality assessment task. To make our active sampling strategies work in these scenarios, we design a general measure of informativeness, I^0, shown as

$$\begin{aligned} I^0(x_i, x_j) = & -P_c(x_i, x_j) \log P_c(x_i, x_j) \\ & - (1 - P_c(x_i, x_j)) \log(1 - P_c(x_i, x_j)). \end{aligned} \tag{17}$$

A subject will provide the correct answer of the image pair (x_i, x_j) with the probability of $P_c(x_i, x_j)$, and the wrong answer with the probability of $1 - P_c(x_i, x_j)$. Thus written as the form of entropy, $I^0(x_i, x_j)$ is the average amount of information contained in the image pair (x_i, x_j). Note that as calculated in (11), $P_c(x_i, x_j)$ increases from 0.5 to 1 when the quality difference $|s(x_i) - s(x_j)|$ increases from 0 to $+\infty$. Therefore, $I^0(x_i, x_j)$ increases with the decreasing of $|s(x_i) - s(x_j)|$, which means that the closer two images' qualities are, the more informativeness the image pair is. From another point of view, $I^0(x_i, x_j)$

Algorithm 1: The whole process

Input: test image dataset X,
the set of the image pairs that are comparable C,
sufficiently compared additional dataset for coefficients training $D(D = \emptyset$, if no data for pretraining)
Output: X's quality score $\{s(x_i)|x_i \in X\}$
initialize X's quality score $\{s^0(x_i)|x_i \in X\}$
$t = 0$, $\lambda = 1$, $k = 1$
while *the test cost is within the budget* **do**
  train coefficients λ and k on D
  for *each image pair* (x_i, x_j) *in* C **do**
    compute $\Delta I_r^t(x_i, x_j)$ based on λ and k;
  end
  $(i^*, j^*) = \arg\max_{(x_i,x_j)\in C} \Delta I_r^t(x_i, x_j)$
  query for image pair (x_{i^*}, x_{j^*})'s preference judgement from a randomly selected subject
  $t = t + 1$
  update X's quality score $s^t(x_i)$
  //update the dataset D for training coefficients λ and k
  if *image pair* (x_{i^*}, x_{j^*}) *is sufficiently compared* **then**
    $D = D \cup \{(x_{i^*}, x_{j^*})\}$
  end
end
return $s^t(x_i)$;

shows the subject's uncertainty on the image pair's relative quality, as the larger $P_c(x_i, x_j)$ is, the smaller $I^0(x_i, x_j)$ is.

E. Active Sampling

Our active sampling strategy is based on reliable informativeness I_r as (5). In each round of the test, instead of comparing the image pair of the maximum informativeness as the previous work, we choose to compare the image pair of the maximum reliable informativeness gain. Suppose images x_i and x_j have been compared n times. If they are to be compared one more time, which is the $(n+1)$th time, their expected reliable informativeness gain, $\Delta I_r(x_i, x_j, n)$ can be defined as

$$\begin{aligned}\Delta I_r(x_i, x_j, n) &= I_r(x_i, x_j, n+1) - I_r(x_i, x_j, n) \\ &= \Delta R(x_i, x_j, n) \times I(x_i, x_j) \qquad (18)\end{aligned}$$

$$\Delta R(x_i, x_j, n) = R(x_i, x_j, n+1) - R(x_i, x_j, n) \qquad (19)$$

where n is the number of times that images x_i and x_j have been compared, and $\Delta R(x_i, x_j, n)$ represents the reliability gain. Therefore, in the tth round of the test, the image pair to be selected for comparison is (x_{i^*}, x_{j^*}), as

$$(i^*, j^*) = \arg\max_{(x_i,x_j)\in C} \Delta I_r^t(x_i, x_j) \qquad (20)$$

$$\Delta I_r^t(x_i, x_j) = \Delta R^t(x_i, x_j) \times I(x_i, x_j) \qquad (21)$$

$$\begin{aligned}\Delta R^t(x_i, x_j) &= R(x_i, x_j, n_{ij}^{t-1} + n_{ji}^{t-1} + 1) \\ &\quad - R(x_i, x_j, n_{ij}^{t-1} + n_{ji}^{t-1}) \qquad (22)\end{aligned}$$

where n_{ij}^{t-1} represents the number of times that x_i is preferred to x_j in the first $t-1$ rounds of the test, while n_{ji}^{t-1} represents the number of times that x_j is preferred to x_i in the first $t-1$ rounds of the test.

The whole process of the subjective test with the active sampling strategy applied can be described as Algorithm 1.

IV. Experiments

We conduct IQA experiments based on three ranking through preference algorithms, which are Bradley-Terry model [39] HodgeRank [23], [24] and RankSVM [32], [33]. In each experiment based on one ranking algorithm, we compare the performances of the conventional active sampling strategies and the proposed strategy where preference labels' reliability is considered. The results of the experiments show that taking preference labels' reliability into consideration makes a notable contribution to reducing sampling scale without decreasing the accuracy.

A. Database

To conduct the evaluation experiment, we collect four databases with preference judgements for IQA problems. 1) Database1 contains 28 images from LIVE database [43], which are all from the same reference image "womanhat", and of 5 distortion types (fastfading, Gaussian blur, JPEG compression, JPEG2000 compression and white noise) and 5 to 6 distortion levels for each distortion type. We make an online pairwise test on the database, and each pair of the images is compared 24 times by different subjects. 2) Database2 contains 104 images from TID2013 database [40]. The 104 images are divided into 4 groups. Each group is from the same reference image. And the four reference images of the four groups are respectively marked as "I05", "I10", "I18" and "I19" in TID2013. Each group contains the reference image and 25 distorted images of 5 distortion types (additive Gaussian noise, Gaussian blur, JPEG compression, JPEG2000 compression and contrast change) and 5 distortion levels. We make an online pairwise test on the database, and each pair of the images in the same group is compared 9 times. 3) Database3 is from [23]. It contains 240 images from LIVE database [43] and IVC database [44], and is divided into 15 groups. Each group is from the same reference image, but of different distortion types and distortion levels. And each pair of images in the same group are compared at least 4 times. 4) Database4 is from [4]. It contains 120 images, including 20 undistorted reference images and 100 distorted images derived from the 20 reference images. Every pair of the 120 images are compared 5 times.

Part of the reference images of the databases are as shown in Fig. 5. We can see that all the databases together contain various kinds of images.

Subjective Test: The subjective test results of Database3 and Database4 are kindly offered by the authors of previous work [4], [23], who made the subjective IQA tests on crowdsourcing platforms. Database3 collected 23097 preference judgements from 186 subjects, while Database4 collected 35700 preference judgements from 195 subjects. The subjective tests on Database1 and Database2 are conducted by ourselves on an online crowdsourcing platform we establish. The testing process is as follows, which is similar to that in [23]. We establish a website for online subjective tests, whose main web interface for the test is shown in Fig. 6. At the beginning of the test, the

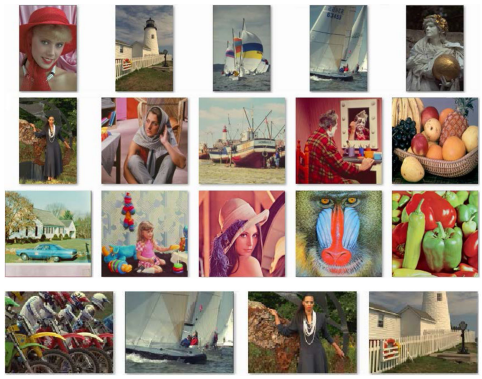

Fig. 5. Part of the reference images of the databases, showing the variety of the reference images.

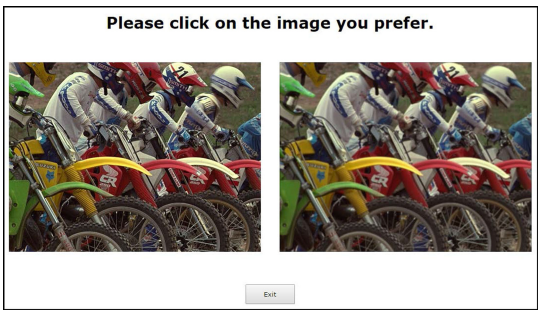

Fig. 6. Web interface of the subjective test. Subjects provide their preference judgements by clicking on the image that they prefer to the other. And they can terminate the test whenever they like by clicking on the button "Exit".

guidance of what image is better in quality is shown to the subject. And in our experiments, an image in higher level of quality means that it is clearer and less distorted. In each round of the test, two images derived from the same reference are randomly selected and shown side by side on the screen. Subjects finish the annotating task on their own personal computers (no mobile devices) through the Internet. It means that the testing environment is not strictly controlled, and varies with the subject. The subjects provide their preference judgements by clicking on the image that they prefer to the other. After their clicking, another two images are shown on the screen, and the test process goes on. Subjects are paid to do the test, and they can stop and restart the test whenever they would like. 33 subjects are involved in the test, 6 of which are researchers or students in the multimedia field. There are 18 males and 15 females among them, and their ages range from 18 to 48. Finally we collect 9072 judgements on Database1 and 11700 judgements on Database2. Each subject judges about 450 image pairs on average.

B. Evaluation Measure

The performance of IQA algorithms is usually evaluated by the correlation coefficients between the estimated scores and the gold-standard scores, for example Spearman rank correlation coefficient (SRCC) and Kendall rank correlation coefficient (KRCC) [40]. However, as our experiments are based on pairwise comparisons, and no directly available quality scores can be used as gold-standard scores. In that case, we transform the complete pairwise comparisons into global scores with Bradley-Terry model [39]. The generated global scores can be regarded as the "gold-standard scores".

There is one concern on taking the "gold-standard scores" above as the ground-truth, for they are estimated from limited pairwise comparisons. A new evaluation measure, miss ratio (MR), is proposed to avoid this concern. Defined as (1), it is the percentage of the image pairs whose preference labels are not correctly predicted after the tth round of the test among all the image pairs in C, similar to that in [23].

As the true quality score $s(x_i)$ is unknown to us, MR^t in (1) cannot be calculated directly. Therefore, in real application, we use an approximation of it, $\widehat{MR^t}$, as

$$\widehat{MR^t} = \frac{\sum_{(x_i,x_j)\in C} \widehat{mis^t}(x_i, x_j)}{Card(C)} \tag{23}$$

$$\widehat{mis^t}(x_i, x_j) = \begin{cases} \dfrac{n_{ij}}{n_{ij}+n_{ji}}, & \text{if } s^t(x_i) < s^t(x_j) \\ \dfrac{n_{ji}}{n_{ij}+n_{ji}}, & \text{if } s^t(x_i) > s^t(x_j) \\ \mathbf{1}_{|n_{ij}-n_{ji}|>\epsilon}, & \text{if } s^t(x_i) = s^t(x_j) \end{cases} \tag{24}$$

where $C = \{(x_i, x_j) | \, 1 \le i < j \le N, x_i \text{ and } x_j \text{ are of the same content.}\}$, and n_{ij} is the total number of times that x_i is preferred to x_j in the databases we conduct experiments on, while n_{ji} is the total number of times that x_j is preferred to x_i. $\mathbf{1}_{f(\cdot)}$ is the indicator function as defined in (3). ϵ is a threshold, which we set as 1 here. It means that when $s^t(x_i) = s^t(x_j)$, n_{ij} and n_{ji} should be very close. Defined as the percentage of preference labels that disagree with the quality score s_t, $\widehat{mis^t}(x_i, x_j)$ is an approximation of $mis^t(x_i, x_j)$ in (2).

C. Evaluation Experiments

To test the performance of the proposed active sampling strategy, we simulate the subjective IQA experiment by repeatedly and randomly sampling from real judgements collected in the above four databases. To be more specific, for an image pair x_i and x_j in the databases, the preference label provided by one subject in one query is x_i better than x_j with a probability of $\frac{n_{ij}}{n_{ij}+n_{ji}}$, and x_j better than x_i with a probability of $\frac{n_{ji}}{n_{ij}+n_{ji}}$.

According to the scenario whether the feature space of the test images is known, different rank aggregation algorithms are applied in the subjective test. We compared our proposed active sampling strategy with other sampling strategies under these two different scenarios.

1) Experiments Based on Bradley-Terry and HodgeRank: In the scenario where the feature space of the test images is unknown, only the ranking algorithm that is not based on learning can be applied, such as Bradley-Terry [39] and HodgeRank [23]. To simulate this scenario, we conduct two IQA tests based on Bradley-Terry model and online HodgeRank respectively.

Bradley-Terry model [39] is a widely used probability model that can predict the outcome of a comparison, which can be used to generate quality scores through pairwise comparisons. For images x_i and x_j, the probability that x_i is better than x_j is

modeled as

$$P(x_i \text{ is better than } x_j) = \frac{e^{s(x_i)}}{e^{s(x_i)} + e^{s(x_j)}}. \quad (25)$$

The values of the quality scores $s(\cdot)$ can be calculated with maximum likelihood estimation.

HodgeRank is a general framework to decompose pairwise comparison data on graphs into three orthogonal components: global ranking, local inconsistency and global inconsistency. When it is used for subjective multimedia quality assessment, the global quality score s is estimated by the following least square problem

$$\min_{s} \sum_{(x_i,x_j)\in C} (n_{ij} + n_{ji})(s(x_i) - s(x_j) - \widehat{Y}_{ij})^2 \quad (26)$$

where

$$\widehat{Y}_{ij} = \frac{n_{ij} - n_{ji}}{n_{ij} + n_{ji}}. \quad (27)$$

And the online version of HodgeRank is illustrated in [23].

The IQA experiment based on Bradley-Terry model or HodgeRank aims to obtain the subjective qualities of all the images in Database1, Database2 and Database3, containing totally 374 images. We apply IQA tests on every group of images from the same reference image in Database1, Database2 and Database3 respectively. Therefore, there are 20 groups of images tested in total. And Database4 is used to pre-train the coefficients λ and k of (11). As Bradley-Terry and HodgeRank are not based on learning, most active learning strategies can not be applied. Thus, we use the margin of an image pair, which is the difference between two images' quality scores, to measure an image pair's informativeness. The lower the margin between two images' quality scores is, the more informative the image pair is. We call this measure of informativeness as "LM". Our general measure of informativeness, I^0, as described in Section III-D, is also another form of "LM". It is because I^0 is monotonically decreasing with the margin between two images increasing. I^0 and "LM" measure image pairs' informativeness in the same way. And the result is the same if we replace "LM" with I^0, except the computational complexity. Therefore, by comparing the performance of a sampling strategy purely based on "LM", and the performance of our proposed sampling strategy using I^0 as the informativeness measure, the value of taking reliability into consideration can be proved.

We have compared the following 8 sampling strategies.

1) Ours1, I^0(LM), without-pretraining: The sampling strategy is based on the proposed method of reliable informativeness. The informativeness measure is set to the default option, I^0, as described in Section III-D, which can be regarded as a new form of "LM". "without-pretraining" means that there are no additional existing data to pre-train the coefficients λ and k in the model of reliability.
2) Ours2, I^0(LM), with-pretraining: The same as sampling strategy A, only with the exception that the coefficients λ and k are pre-trained with the data of Database4.
3) Ours3, only Reliability: Selecting the image pair of the maximum reliability gain each time, regardless of informativeness measure. As the experiments show that whether the coefficients λ and k are pre-trained has little influence on the performance of this method, we only show the performance without pretraining.
4) Random: In each round of the test, two images are randomly selected to be compared one time.
5) LM, n = 1: In each round of the test, the two images of the lowest margin are selected to be compared one time, and once the two images have been compared, they will never be selected again.
6) LM, n = 3, without-replacement: In each round of the test, the two images of the lowest margin are selected to be compared by three subjects. And majority voting is used to decide its final preference label. "without-replacement" means that once the two images are selected in one round of the test, it will never be selected in other rounds.
7) LM, n = 3, with-replacement: In each round of the test, the two images of the lowest margin are selected to be compared one time. And any image pair will be selected at most 3 times.
8) Ye [4], This method is from the work of Ye *et al.* [4]. The original method is a combination of CR test and PC test. Here we simplify the original method and only PC test is applied. Note that in this method the PC results are transformed into quality scores with a Bayesian approach of the original method itself, rather than Bradley-Terry model or HodgeRank.

Figs. 7 and 8 respectively present the performance of the above sampling strategies with Bradley-Terry model and HodgeRank. Both experiments are repeated 500 times. As the tests are conducted on 20 groups of images, the performances in Figs. 7 and 8 are respectively an average of the performances on the 20 groups. Fig. 7(a) and Fig. 8(a) show the SRCC of generated global quality scores increasing with the accumulation of preference labels. Fig. 7(b) and Fig. 8(b) evaluate the performances with KRCC. Fig. 7(c) and Fig. 8(c) show the miss ratio (MR) of global quality scores decreasing with the accumulation of preference labels. Please note that the yellow star curves in Figs. 7 and 8 are the same, for "H: Ye [4]" uses its own approach for preference aggregation, rather than Bradley-Terry or HodgeRank. The figures of SRCC, KRCC, and MR show similar relative performances of different sampling strategies. From both of the three evaluating measures, we can tell that:

1) *Most active sampling strategies outperform random sampling.* It shows that the appropriate application of active sampling is able to improve the performance compared with random sampling.
2) *Having each selected image pair compared by an equivalent and fixed number of times is not wise.* Sampling strategies E and F represent the sampling strategies where each selected image pair is compared by an equivalent number of times. Here, the number is set as 1 in E, and 3 in F. As E outperforms F in the beginning of the test, while F obtains a lower MR than E does in the end, it shows that having each image pair compared multiple times makes a contribution to the accuracy in the long term, while comparing each image pair fewer times can

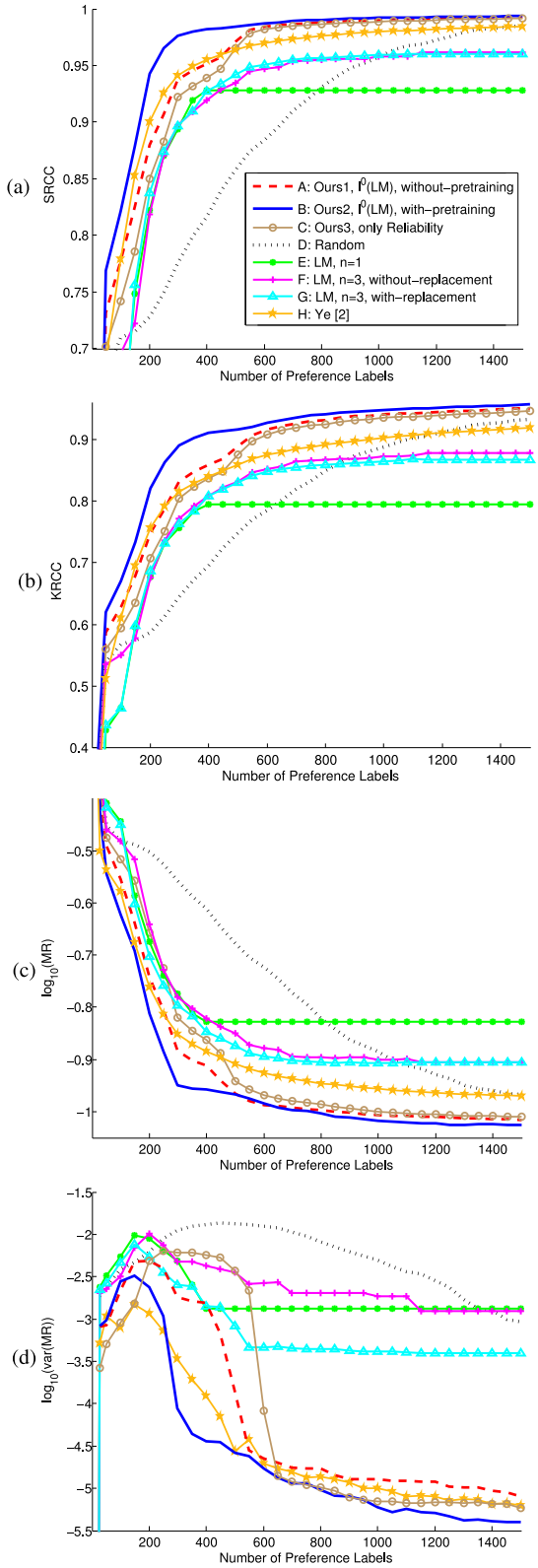

Fig. 7. These four figures show the performances of the eight different sampling strategies. The first seven sampling strategies (A-G) work with Bradley-Terry model as the preference aggregation algorithm, while the sampling strategy of H [4] is an independent method using a Bayesian approach of itself for preference aggregation. The plot in (a) evaluates the performances with SRCC. The plot in (b) evaluates the performances with KRCC. The plots in (c) and (d) evaluate the performances with miss ratio (MR). (a) SRCC. (b) KRCC. (c) Miss Ratio. (d) Variance of MR.

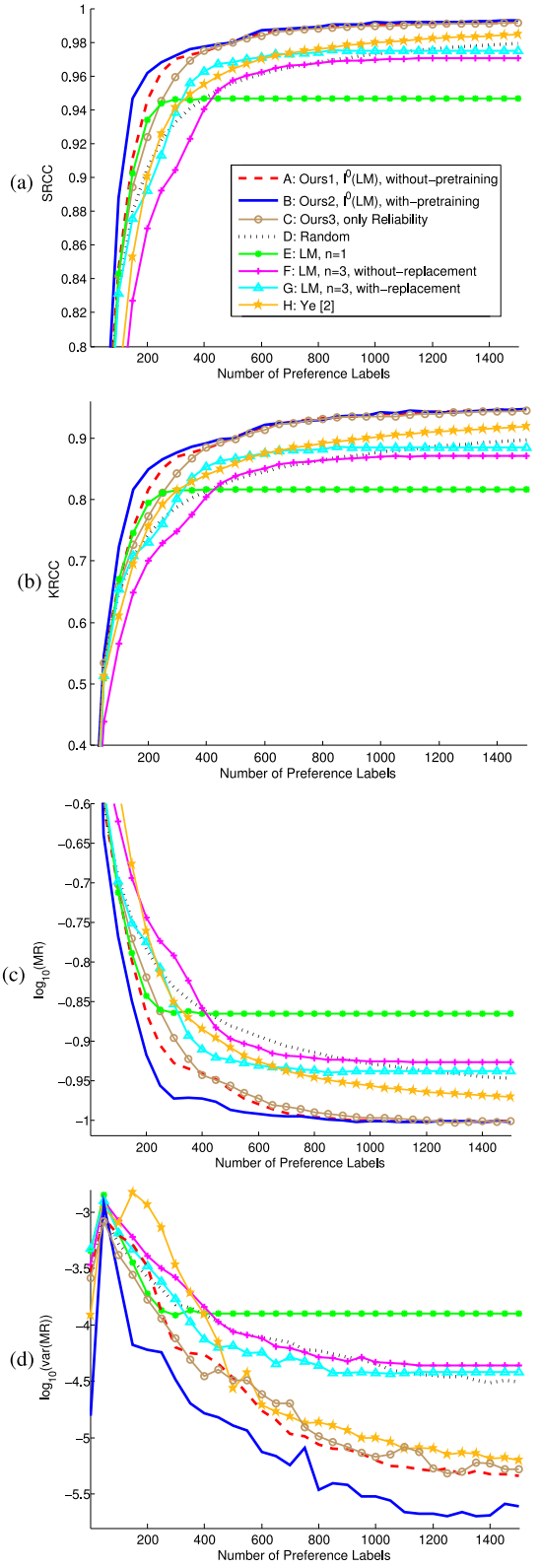

Fig. 8. These four figures show the performances of the eight different sampling strategies. The first seven sampling strategies (A-G) work with HodgeRank as the preference aggregation algorithm, while the sampling strategy of H [4] is an independent method using a Bayesian approach of itself for preference aggregation. The plot in (a) evaluates the performances with SRCC. The plot in (b) evaluates the performances with KRCC. The plots in (c) and (d) evaluate the performances with miss ratio (MR). (a) SRCC. (b) KRCC. (c) Miss Ratio. (d) Variance of MR.

decrease the MR fast in the early steps of the test. This means that adaptively adjusting the number of times each image pair compared can be an effective way to improve the performance. Sampling strategy G, which only sets the upper bound of the times that each image pairs are compared, can be seen as a simple way of the adjustment. As G outperforms E and F on the whole, it proves that the adjustment is effective for an active sampling strategy.

3) *Our idea of adjusting subjects' effort on different image pairs by introducing reliability makes sense.* With the reliability of preference labels considered, we can see that our proposed active sampling strategies A and B significantly outperform the strategies only based on informativeness. Besides, with only reliability considered, method C performs quite well. It proves that reliability is an important issue in IQA problem.
4) *In our proposed strategy, the pretraining of the coefficients λ and k can speed up the ranking in the early steps.* We can tell it from the fact that sampling strategy B outperforms A. This is because the pretraining of the coefficients helps to estimate the labels' reliability when there are not enough labels available. It proves the effect of introducing reliability in another way.

As for the stability, our methods are at least not worse than the others. Fig. 7(d) and Fig. 8(d) show the variance of MR across 500 trials for all the sampling strategies with Bradley-Terry model and HodgeRank. We can see that our strategy with pretrained coefficients, B, is of the lowest variance. The variance of our strategy without pre-trained coefficients, A, is similar with that of Ye *et al.*'s method H, and lower than the left D, E, F and G. The variance of B is low, because the pretraining of the coefficients leads to an accurate estimation of reliability in the early steps of the test, which makes a contribution to adjusting the effort spent on each image pair. As for the sampling strategy A, the inaccurate estimation in the early steps makes its performance diverse. The comparison of variance proves the stability of our strategy, especially when there are additional existing data for pretraining.

2) Experiments Based on RankSVM: In the scenario that the feature space of the test images is known, ranking algorithms based on learning can be used. To simulate this scenario, we make an IQA experiment based on linear RankSVM [33], [45]. We adopt the L2 regularization and L2 loss linear RankSVM formulation to learn the quality score function $s(x;\boldsymbol{\omega}) = \boldsymbol{\omega}^T \boldsymbol{x}$ shown in (28). Given $P = \{(x_i, x_j)|1 \le i, j \le N, s(x_i) > s(x_j)\}$ as the constraints set according to the preference labels

$$\min_{\boldsymbol{\omega}} \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + \gamma \sum_{(x_i,x_j)\in P} \max(0, 1 - \boldsymbol{\omega}^T(\boldsymbol{x_i} - \boldsymbol{x_j}))^2 \quad (28)$$

where $\boldsymbol{\omega}$ is the linear model parameter and $\gamma > 0$ is a regularization parameter, and $\boldsymbol{x_i}$ represents the feature vector of image x_i. The optimal model parameters are learned through Newton's method [45].

As the experiments on Bradley-Terry model and HodgeRank, we apply the algorithm on every group of images from the same reference image in Database1, Database2 and Database3 respectively. Therefore, there are 20 groups of images tested in total. And Database4 is used to pretrain the values of the coefficients λ and k in (11). For the representation of each image, we adopt the 12-dimension feature of Spatial-Spectral Entropy-based Quality (SSEQ) index [46].

In the IQA test based on RankSVM, we adopt two measures of informativeness, "LM" and "DLM". "LM" is defined the same as the experiments based on Bradley-Terry model and HodgeRank in Section IV-C1. The performances of the sampling strategies with "LM" as the informativeness measure are shown in Fig. 9. "DLM" [10] is another measure of informativeness. Compared to "LM", it introduces diversity constraints by requiring that the two images for comparison should be from different clusters. The cluster here is a result of conducting K-means [47] on the images features. The performances of the sampling strategies with "DLM" as the informativeness measure are shown in Fig. 10. The concrete definition of the 8 sampling strategies compared in Fig. 10 is as follows.

1) Ours1, I^0(DLM), without-pretraining: The same as the strategy used in the experiments in Figs. 7, 8 and 9, with the exception that the two selected images are required to be from two different clusters. "without-pretraining" means that there are no additional existing data to pretrain the coefficients λ and k in the model of reliability.
2) Ours2, I^0(DLM), with-pretraining: The same as A, only with the exception that the coefficients λ and k are pretrained with the data of Database4.
3) Ours3, Diversity & Reliability: The same as the sampling strategy of C in Figs. 7, 8 and 9, with the exception that the two selected images are required to be from two different clusters. As the experiments show that whether the coefficients λ and k are pre-trained has little influence on the performance of this method, we only show the performance without pretraining.
4) Random: The same as that in the experiments shown in Figs. 7 and 8.
5) DLM, n = 1: In each round of the test, the two images from two different clusters, the margin between whose quality scores is the lowest, are selected to be compared one time, and once the two images have been compared, they will never be selected again.
6) DLM, n = 3, without-replacement: In each round of the test, the two images from two different clusters, the margin between whose quality scores is the lowest, are selected to be compared by three subjects. And majority voting is used to decide its final preference label. "without-replacement" means that once the two images are selected in one round of the test, it will never be selected in other rounds.
7) DLM, n = 3, with-replacement: In each round of the test, the two images from two clusters, the margin between whose quality scores is the lowest, are selected to be compared one time. And any two images will be selected at most 3 times.
8) Ye [4]: The same as that in the experiments shown in Figs. 7, 8 and 9.

Figs. 9 and 10 respectively present the performances of the sampling strategies using "LM" and "DLM" as the measure

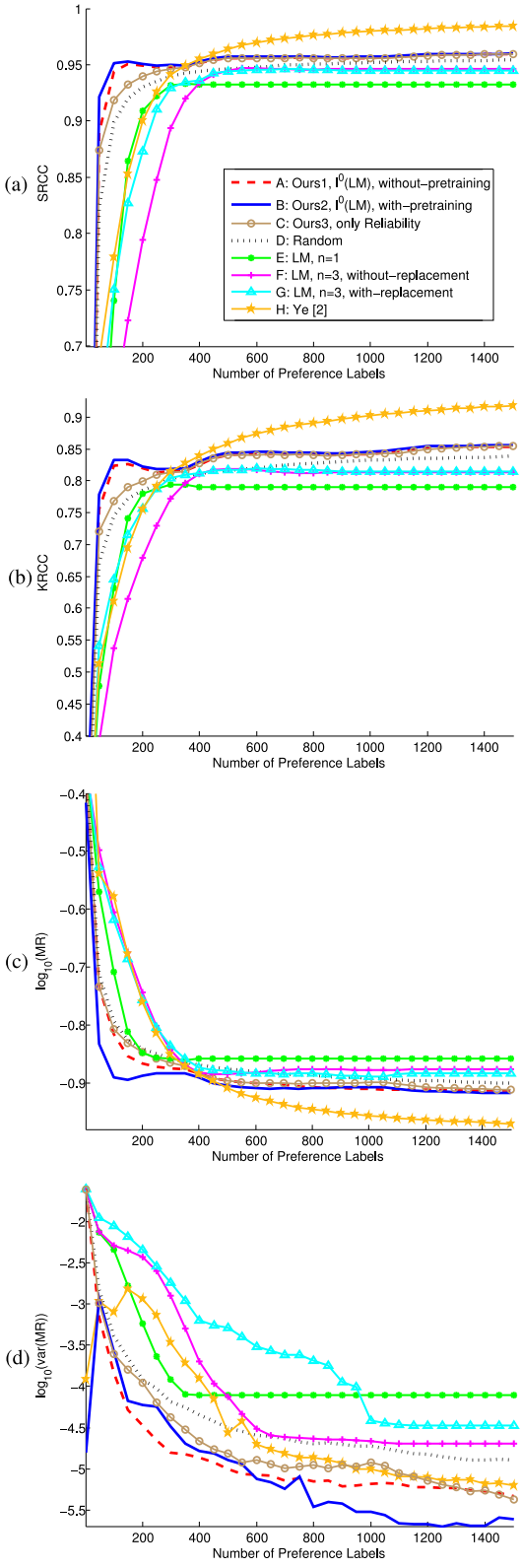

Fig. 9. These four figures show the performances of the eight different sampling strategies. The first seven sampling strategies (A-G) work with RankSVM as the preference aggregation algorithm, and "LM" as the informativeness measure. The sampling strategy of H [4] is an independent method using a Bayesian approach of itself for preference aggregation. The plot in (a) evaluates the performances with SRCC. The plot in (b) evaluates the performances with KRCC. The plot in (c) and (d) evaluate the performances with miss ratio (MR). (a) SRCC. (b) KRCC. (c) Miss Ratio. (d) Variance of MR.

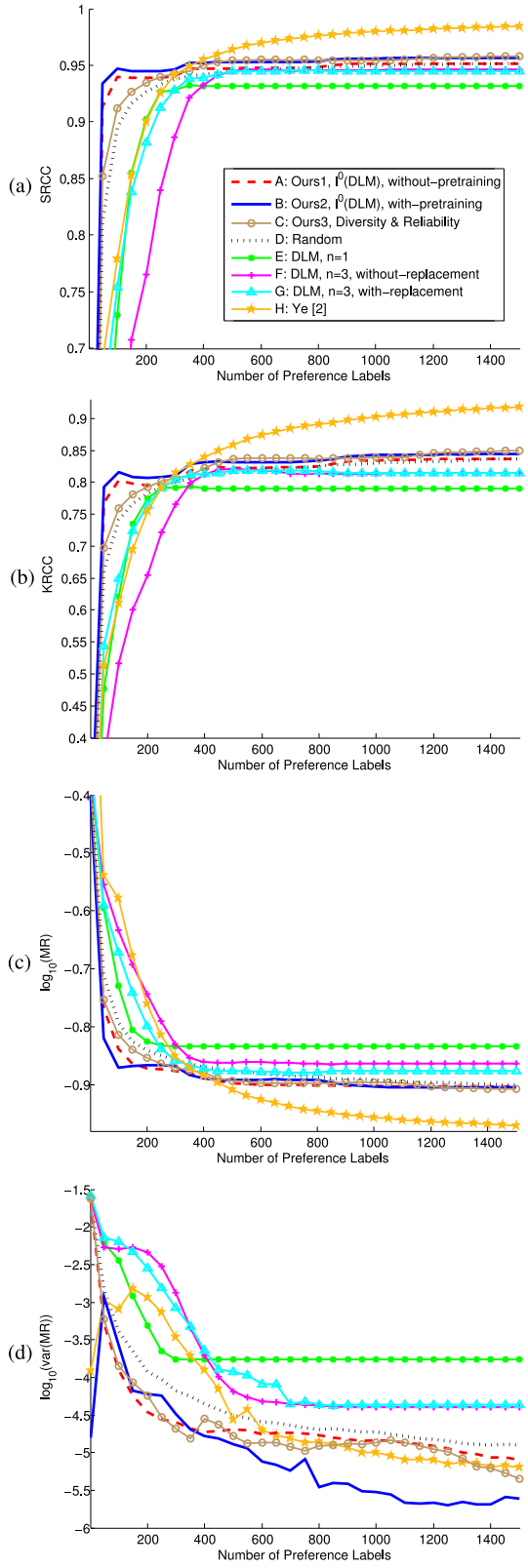

Fig. 10. These four figures show the performances of the eight different sampling strategies. The first seven sampling strategies (A-G) work with RankSVM as the preference aggregation algorithm, and "DLM" as the informativeness measure. The sampling strategy of H [4] is an independent method using a Bayesian approach of itself for preference aggregation. The plot in (a) evaluates the performances with SRCC. The plot in (b) evaluates the performances with KRCC. The plot in (c) and (d) evaluate the performances with miss ratio (MR).

of informativeness. Both experiments are repeated 500 times. As the tests are conducted on 20 groups of images, the performances in Figs. 9 and 10 are respectively an average of the performances on the 20 groups. Fig. 9(a) and Fig. 10(a) show that the SRCC of global quality scores increases with the accumulation of preference labels. Fig. 9(b) and Fig. 10(b) evaluate the performances with KRCC. Fig. 9(c) and Fig. 10(c) show that the MR of global quality scores decreases with the accumulation of preference labels. The figures of SRCC, KRCC and MR show similar relative performances of different sampling strategies.

On the whole, Figs. 9 and 10 show a similar result as Figs. 7 and 8. We can see that our strategies outperform the other strategies except "H: Ye [4]". Our strategies outperform "H: Ye [4]" in the beginning of the test, but are worse than it in the end of the test. Note that H is an independent method, and that the yellow star curves of H in Figs. 7, 8, 9 and 10 are exactly the same. Considering the performance comparison in Figs. 7 and 8, where our strategies are better than H, we think it is the preference aggregation algorithm, rather than the sampling strategy that makes our strategies worse than H in the end of the test in Figs. 9 and 10. Furthermore, RankSVM is based on learning, while the preference aggregation of H is not. Therefore, we think it is the learning based method that causes the degeneration of our strategies' performance in the end of the test. The reason might be that the performances of learning-based methods can be limited by the representation ability of images' features and the power of the learning model applied.

Fig. 9(d) and Fig. 10(d) show the variance of MR across 500 trials. We can get a similar conclusion as that in IQA tests based on Bradley-Terry and HodgeRank. It is that our proposed strategies are no worse than others in stability.

Comparing the performances of the experiments in the two scenarios, whether the feature space is known or not, we can find that the difference between our strategies A and B in the experiments based on RankSVM, is smaller than that in the experiments based on Bradley-Terry or HodgeRank. This is because with a ranking algorithm based on learning, thanks to its ability of generalization, the global quality score can converge to a value that is relatively accurate in the early step of the test, which contributes to the estimation of the reliability.

When the preference aggregation algorithm is based on learning, more active sampling strategies (or active learning strategies) can be used. We introduce an active learning strategy proposed by Zhu *et al.* [12] for the IQA problem. It is a typical active learning strategy combining uncertainty measure and information density measure. When it is applied to the IQA problem, for each image pair, we concatenate the feature vectors of the two images to represent the image pair. The uncertainty of an image pair is represented based on the difference between the two images' current quality scores. We compare Zhu *et al.* [12] with our methods as shown in Fig. 11. The methods of A, B and C in Fig. 11 are the same as that in Fig. 9. The method of D is based on the density measure proposed in [12]. The method of E is a hybrid approach based on the whole work in [12]. Fig. 11 shows that our methods obviously outperform Zhu *et al.* 's.

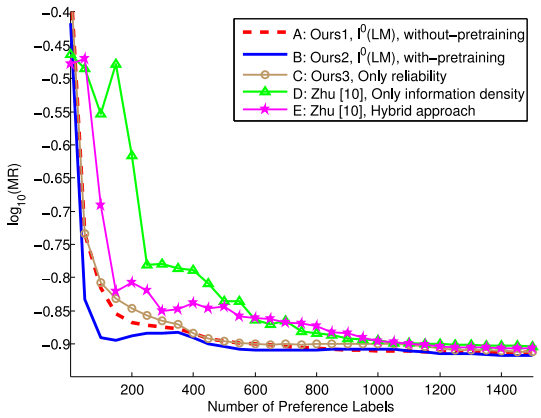

Fig. 11. Comparisons among our methods and other typical active learning methods. All the methods are conducted with RankSVM for preference aggregation.

V. Conclusion

To reduce the large testing scale of subjective IQA tests based on pairwise comparison, and to obtain a more accurate global ranking of the images, we present an active sampling strategy to select image pairs for comparison. By introducing the reliability of image pairs' preference labels into image pairs' informativeness, our active sampling strategy is based on reliable informativeness. It is able to not only select informative image pairs as previous works, but also adjust efforts spent on different image pairs based on the image pairs' ambiguity. As verified by our experiments, our method can effectively improve the performance of the active sampling strategies only based on informativeness. Our future work will focus on introducing the estimation of each subject's expertise into the sampling strategy, which aims to make our strategy more robust in crowdsourcing environment. Besides, our strategy theoretically can adapt to the evaluation of higher-level properties of more multimedia contexts and applications. For example, the evaluation of aesthetic value can be conducted by having subjects judge the relative aesthetic value. Therefore, in the future, we will extend the application of the proposed active sampling strategy to more general evaluation problems, such as the evaluation of aesthetic value and relative attributes, and verify its generality.

References

[1] *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T recommendation p. 910, Apr. 2008.
[2] M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, and K. Brunnström, "Subjective multimedia quality assessment," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. 89-A, no. 11, pp. 2920–2932, 2006.
[3] J. S. Lee, "Paired comparison for subjective multimedia quality assessment: Theory and practice," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2013, pp. 1099–1102.
[4] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," in *Proc. 27th IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 4249–4256.
[5] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 491–500.
[6] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec. 2015.

[7] J. S. Lee, F. D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 882–893, Oct. 2011.
[8] J. S. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 564–571, Feb. 2014.
[9] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag: Adding prior constraints into active learning," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1058–1065.
[10] L. Liang and K. Grauman, "Beyond comparing image pairs: Setwise active learning for relative attributes," in *Proc. 27th IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 208–215.
[11] T. Osugi, D. Kim, and S. Scott, "Balancing exploration and exploitation: A new algorithm for active machine learning," in *Proc. 5th IEEE Int. Conf. Data Min.*, 2005.
[12] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1323–1331, Aug. 2010.
[13] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *Proc. 22th IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 762–769.
[14] A. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. 22th IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 2372–2379.
[15] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proc. 6th ACM Int. Conf. Web Search Data Min.*, 2013, pp. 193–202.
[16] G. Hua, C. Long, M. Yang, and Y. Gao, "Collaborative active learning of a kernel machine ensemble for recognition," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1209–1216.
[17] C. Long, G. Hua, and A. Kapoor, "Active visual recognition with expertise estimation in crowdsourcing," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3000–3007.
[18] M. K. Stern and J. H. Johnson, "Just noticeable difference," in *Corsini Encyclopedia of Psychology*. New York, NY, USA: Wiley, 2010.
[19] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1338–1350, Jul. 2016.
[20] T. Hossfeld *et al.*, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.
[21] F. Ribeiro, D. Florncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. 36th IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 2416–2419.
[22] K. T. Chen, C. J. Chang, C. C. Wu, Y. C. Chang, and C. L. Lei, "Quadrant of euphoria: A crowdsourcing platform for QoE assessment," *IEEE Netw.*, vol. 24, no. 2, pp. 28–35, Mar./Apr. 2010.
[23] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 359–368.
[24] Q. Xu *et al.*, "HodgeRank on random graphs for subjective video quality assessment," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 844–857, Jun. 2012.
[25] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, Mar. 2009.
[26] S.-T. Park and W. Chu, "Pairwise preference regression for cold-start recommendation," in *Proc. 3th ACM Conf. Recommender Syst.*, 2009, pp. 21–28.
[27] B. Reilly, "Social choice in the south seas: Electoral innovation and the borda count in the pacific island countries," *Int. Political Sci. Rev.*, vol. 23, no. 4, pp. 355–372, 2002.
[28] N. Alon, "Ranking tournaments," *SIAM J. Discrete Math.*, vol. 20, no. 1, pp. 137–142, 2006.
[29] Y. Zhang *et al.*, "Consensus-based ranking of multivalued objects: A generalized borda count approach," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 83–96, Jan. 2014.
[30] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial Hodge theory," *Math. Program.*, vol. 127, no. 1, pp. 203–244, 2010.
[31] S. Negahban, S. Oh, and D. Shah, "Iterative ranking from pair-wise comparisons," in *Proc. 25th Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2483–2491.
[32] T. Joachims, "Training linear SVMs in linear time," in *Proc. 12th ACM Int. Conf. Knowl. Discovery Data Min.*, 2006, pp. 217–226.
[33] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMs," *Inf. Retr.*, vol. 13, no. 3, pp. 201–215, 2009.
[34] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proc. 22th ACM Int. Conf. Mach. Learn.*, 2005, pp. 89–96.
[35] N. Ailon, "An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 137–164, Jan. 2012.
[36] K. G. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," in *Proc. 24th Conf. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2240–2248.
[37] R. Yan, J. Yang, and A. Hauptmann, "Automatically labeling video data using multi-class active learning," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 516–523.
[38] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," in *Proc. 24th IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1449–1456.
[39] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
[40] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process. Image Commun.*, vol. 30, pp. 57–77, 2015.
[41] Y. Shen and T. Jiang, "Ranking consistent rate: New evaluation criterion on pairwise subjective experiments," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2077–2081.
[42] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 1994, vol. 1.
[43] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment database release 2," 2005. [Online]. Available: http://live.ece.utexas.edu/research/quality
[44] P. Le Callet and F. Autrusseau, "Subjective quality assessment irccyn/ivc database," 2005. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/
[45] C. Olivier, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, 2007.
[46] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process. Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.
[47] J. Hartigan and M. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

Zhiwei Fan received the B.S. degree in computer science from Peking University, Beijing, China, in 2010, where she is currently working toward the M.S. degree in computer applied technology.

Her research interests include image quality assessment and active learning.

Tingting Jiang (M'10) received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2001, and the Ph.D. degree in computer science from Duke University, Durham, NC, USA, in 2007.

She is now an Associate Professor of computer science with Peking University, Beijing, China. Her research interests include computer vision, image, and video quality assessment.

Tiejun Huang (M'01–SM'12) received the B.S. and M.S. degrees in computer science from the Wuhan University of Technology, Wuhan, China, in 1992 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong (Central China) University of Science and Technology, Wuhan, China, in 1998.

He is currently a Professor and the Chair of Department of Computer Science, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. His research interests include video coding, image understanding, and neuromorphic computing.

Prof. Huang was the recipient of the National Science Fund for Distinguished Young Scholars of China in 2014, and was awarded the Distinguished Professor of the Chang Jiang Scholars Program by the Ministry of Education in 2015. He is a member of the Board of the Chinese Institute of Electronics and the Advisory Board of IEEE Computing Now.