Online Crowdsourcing Subjective Image Quality Assessment *

Qianqian Xu
Graduate University, Chinese Academy of Sciences, Beijing 100049, China
qqxu@jdl.ac.cn

Qingming Huang
Graduate University, Chinese Academy of Sciences, Beijing 100049, China
qmhuang@jdl.ac.cn

Yuan Yao*
School of Mathematical Sciences, LMAM and LMP, Peking University, Beijing 100871, China
yuany@math.pku.edu.cn

ABSTRACT

Recently, HodgeRank on random graphs has been proposed as an effective framework for multimedia quality assessment problem based on paired comparison method. With the random design on large graphs, it is particularly suitable for large scale crowdsourcing experiments on Internet. However, to make it more practical toward this purpose, it is necessary to develop online algorithms to deal with sequential or streaming data. In this paper, we propose an online rating scheme based on HodgeRank on random graphs, to assess image quality when assessors and image pairs enter the system in a sequential way in a crowdsourceable scenario. The scheme is shown in both theory and experiments to be effective by exhibiting similar performance to batch learning under the Erdös-Rényi random graph model for sampling. It enables us to derive global rating and monitor intrinsic inconsistency in the real time. We demonstrate the effectiveness of the proposed framework on LIVE and IVC databases.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology*; C.4 [**Performance of Systems**]: Design studies; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*

General Terms

Performance, Experimentation, Human Factors

Keywords

Subjective Image Quality Assessment, Online, Crowdsourcing, Paired Comparison, HodgeRank, Random Graphs, Triangular Curl, Topology Evolution, Persistent Homology

*Area chair: Heng Tao Shen.
*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$15.00.

1. INTRODUCTION

Image Quality Assessment (IQA) fundamentally relies on subjective experiments to capture the *true* perception of human observers. Therefore, subjective tests are often used to provide the ground-truth and verification for objective models in IQA. In a typical Mean Opinion Score (MOS) test [1], individuals are asked to give a rating from Bad to Excellent (e.g. Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to grade the quality of an image. However, such a test may suffer from various problems such as ambiguity in definition of scales, dissimilar interpretations of the scale among users, etc. [7]. Therefore, paired comparison method is recently gaining rising attention, in which raters are asked to compare two images simultaneously and vote which one has the better quality; this is an easier and less demanding task for raters, whence more reliable in practice.

However, paired comparison approach leaves a heavier burden on participants with a larger number $\binom{n}{2}$ of comparisons. To address this issue, there has been a large volume of statistical literature on deterministic incomplete block design [11]. However, these designs are not suitable for crowdsourcing on Internet where the raters are distributive over network with varied backgrounds and it is hard to control with traditional experimental designs. To meet this challenge, the work in [15] proposes a randomized paired comparison method which randomly selects small subsets of pairs for each assessor to view; the work shows that randomization is effective in reducing costs of a complete design without jeopardizing the intended purpose. However, it leaves some open problems arising from randomization: (1) how to systematically deal with the resulting imbalanced and incomplete data; (2) how many samples are needed to achieve certain approximation of the complete design.

To address these two questions, a new framework called HodgeRank on Random Graphs (***HRRG***) is introduced to analyze the imbalanced and incomplete data in random design experiments [40, 39]. In this framework, paired comparison data are mapped to edge flows on a paired comparison graph which is often a random multigraph in random design, and then Hodge decomposition on graphs [19] leads to an orthogonal decomposition of such edge flows into global rating as gradient flow, local inconsistency as triangular curl flow, and global inconsistency as harmonic flow. Random graphs are shown as good models to design random sampling schemes in particular for crowdsourcing experiments. For example, Erdös-Rényi random graphs select pairs of videos or images uniformly from all possible candi-

dates, while random k-regular graphs keep a balanced sampling for each video/image which receives the same number of comparisons against others and thus important for sparse graph designs [39]. Consistent with recent developments in random graph theory, when sampling complexity is large enough one can remove the global inconsistency. Experiments show that such a scheme in random designs provides good approximations of global ratings derived from complete experimental designs. In the successful developments above for subjective multimedia assessment, it leaves open to explore the online algorithms to deal with streaming data in crowdsourcing experiments on Internet.

Crowdsourceable quality assessment on Internet collects paired comparison data in a distributive and streaming way from a large population over Internet participants [7]. The streaming data calls for online algorithms as a sequential decision process via incremental data updates to improve its prediction accuracy. Although the image quality itself is constant, in subjective IQA, preferences may vary over raters and image pairs with different criteria based on different salient features of images in attention, noise from environment, and levels of attention, etc. Thus it is a fundamental question in subjective IQA to aggregate preferences of multiple assessors into a consistent global score, reflecting the statistical consensus on image quality over population. In this paper, we fill in this gap by presenting an online rating algorithm for HodgeRank on random graphs. Our algorithm is based on the classic Robbins-Monro procedure [28] which has been widely exploited in online learning, e.g. [32, 41].

Online algorithms could offer significant computational advantages over batch algorithms and the benefits of online learning become more evident when dealing with streaming or large-scale data. Besides, to tackle the scenario that in crowdsourcing assessors and image pairs come in an unspecified way, Erdös-Rényi random graph is systematically exploited in the paper which is equivalent to the standard assumption in statistical learning that the sample sequence is independent and identically distributed (I.I.D.). Furthermore, we note that online algorithms can be applied to more general settings with edge independent sampling such as Mutli-attribute random graphs, dependent sampling such as Markov sampling, and tracking time-varying environment.

We demonstrate the effectiveness and generality of the proposed framework on LIVE [3] and IVC [2] databases, which include 15 different reference images and 15 distorted versions of each reference. Totally 186 observers have carried out the experiment via Internet, providing us 23,097 paired comparisons. Experimental results show that the proposed online rating algorithm is promising and a robust assessment method suitable for crowdsourceable subjective IQA.

Our contribution in this work is the following:

1. To the best of our knowledge, it is the first time to propose an online rating framework for exploratory quality assessment. The framework provides the possibility of making assessment procedure significantly faster without deteriorating the accuracy, while maintaining the freedom of assessors.

2. The online rating algorithm is based on Robbins-Monro procedure or stochastic gradient descent for HodgeRank on random graphs. For an independent sampling process, the online rating reaches minimax convergence rates whence asymptotically as efficient as a batch algorithm. Moreover, online tracking of ranking inconsistency is possible in this framework.

3. Crowdsourcing image assessment experiments are conducted based on Erdös-Rényi random graph designs, which further confirms the theoretical analysis by showing that the proposed online rating algorithm achieves similar convergences to batch algorithms.

The remainder of this paper is organized as follows. Section 2 contains a review of related works. Then we describe the proposed framework in Section 3, and establish the online HodgeRank models based on batch HodgeRank. The detailed experiments are demonstrated in Section 4. Section 5 presents the conclusive remarks along with discussion for future work.

2. RELATED WORK

2.1 Subjective Quality Assessment

There have been studies on the design of subjective tests to evaluate image/video quality in paired comparison method. One such example is [7], which proposes a crowdsourceable framework based on paired comparison. However, one major shortcoming of [7] lies in that it makes a strong assumption that all paired comparison data collected are complete which is impossible for a large number of videos. For example, the way to evaluate Transitivity Satisfaction Rate (TSR) depends on such complete design assumption. To address this issue, the work in [15] suggests a randomised pair comparison method in which a random subset of all pairs are chosen for different participants to reduce the number of comparisons. However, this work does not address how to deal with the imbalanced and incomplete data arisen in random sampling, and also leaves the issue open on how many samples one needs.

To solve these problems, [40, 39] present a framework based on HodgeRank [19] on random graphs, which deal with incomplete and imbalanced data distributed on random graphs and further derive the constraints on sampling complexity in crowdsourcing experiment that the random selection must adhere to.

2.2 Online Learning

Online learning is a well established subfield of machine learning concerned with estimation problems with limited access to the entire data. It is a sequential decision process $(f_t)_{t\in N}$ in the hypothesis space, where each f_t is decided by the current observation $z_t = (x_t, y_t)$ and f_{t-1} which only depends on previous examples, i.e. $f_t = T_t(f_{t-1}, z_t)$. As a contrast, batch learning refers to a decision utilizing the whole set of examples available at time t [33, 10]. Examples of online learning algorithms include Perceptrons [29] and Adaline [35], etc.

The performance of the online learning algorithms is often measured by a loss function, which is often assumed to be convex such that convex optimization technique can be used to solve the problem. Typical examples of loss functions include hinge loss and square loss. The hinge loss is used in Support Vector Machines (SVM) [9] for classifications and the square loss leads to Least Mean Square method, such as Adaline and its variations [36].

Because of the lower computational cost of online learning compared with batch leaning, it has been shown to benefit a number of computer vision applications such as object

recognition [12, 18], object detection [25, 38] and tracking [26, 17, 22]. The benefits of online learning become more evident when dealing with streaming or very large-scale data.

In this paper, we propose an online learning scheme based on stochastic gradient decent method in the setting of crowdsourceable subjective IQA.

2.3 Random Graphs

Random graph is a graph generated by some random process [5, 8]. It starts with a set of n vertices and adds edges between them at random. With such models we aim at crowdsourcing experimental designs where assessors may select image pairs at random. Different random graph models produce different probability distributions on graphs. The most commonly studied one is the Erdös-Rényi random graph [16] which is a stochastic process that starts with n vertices and no edges, and at each step adds one new edge uniformly. Besides, there are some other kinds of random models, such as random regular graph [37], preferential attachment random graph [4], small world random graph [34], and geometric random graph [24], which may also play important roles under certain circumstances.

However, as Erdös-Rényi random graph can be viewed as a random sampling process of image pairs or edges independently and identically distributed (I.I.D.), and thus is well suited to our online crowdsourcing test system. In this paper, we particularly focus on this kind of random graph, Erdös-Rényi random graph, leaving other models for future studies.

3. ONLINE HODGERANK

In this section, we propose a new online design to conduct paired comparison for subjective IQA and Erdös-Rényi random graph model is chosen to tackle the scenario that in crowdsourcing raters and pairs come in an unspecified way. Specifically, we first describe Hodge theory on general graphs, and then explain how to develop the online rating algorithms. An upper bound for convergence of such online rating algorithms is given to justify the settings that minimax parametric rate is met. Finally, we discuss how to online track triangular curls and topological changement.

3.1 Batch HodgeRank on Graphs

HodgeRank [19] is a general framework to decompose paired comparison data on graphs, possibly imbalanced (where different image pairs may receive different number of comparisons) and incomplete (where every participant may only give partial comparisons), into three orthogonal components:

$$\textit{aggregate paired ranking} =$$

$$\textit{global ranking} \bigoplus \textit{local inconsistency} \bigoplus \textit{global inconsistency}$$

To be precise, consider paired ranking data on a graph $G = (V, E)$, $Y_\alpha : E \to R$ such that $Y_{ij}^\alpha = -Y_{ji}^\alpha$ where α is the participant index. Without loss of generality, one assumes that $Y_{ij}^\alpha > 0$ if α prefers i to j and $Y_{ij}^\alpha \le 0$ otherwise, with the magnitude representing the degree of preference. In a dichotomous choice, Y_{ij}^α can be taken as $\{\pm 1\}$.

In subjective multimedia assessment, it is natural to assume

$$Y_{ij}^\alpha = s_i^* - s_j^* + \varepsilon_{ij}^\alpha \tag{1}$$

where $s^* : V \to \mathbb{R}$ is some true scaling score on V and ε_{ij}^α are independent noise of mean zero and fixed variance.

Under such assumptions, Gauss-Markov theorem tells us that the unbiased estimator is given by the following least square problem

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j,\alpha} \omega_{ij}^\alpha (s_i - s_j - Y_{ij}^\alpha)^2, \tag{2}$$

where ω_{ij}^α denotes the number of paired comparisons on $\{i, j\}$ made by rater α. It can be rewritten as the following weighted least square form

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j} \omega_{ij} (s_i - s_j - \hat{Y}_{ij})^2, \tag{3}$$

where $\hat{Y}_{ij} = (\sum_\alpha \omega_{ij}^\alpha Y_{ij}^\alpha)/(\sum_\alpha \omega_{ij}^\alpha)$ and $\omega_{ij} = \sum_\alpha \omega_{ij}^\alpha$.

To characterize the solution and residue of (3), we first define the triangle set of G as all the 3-cliques in G:

$$T = \left\{ \{i, j, k\} \epsilon \binom{V}{3} \,|\, \{i, j\}, \{j, k\}, \{k, i\} \epsilon E \right\}. \tag{4}$$

Then every $\hat{Y}$ admits an orthogonal decomposition adapted to G

$$\hat{Y} = \hat{Y}^g + \hat{Y}^h + \hat{Y}^c, \tag{5}$$

where

$$\hat{Y}_{ij}^g = \hat{s}_i - \hat{s}_j, \quad \text{for some } \hat{s} \in \mathrm{R}^V, \tag{6}$$

$$\hat{Y}_{ij}^h + \hat{Y}_{jk}^h + \hat{Y}_{ki}^h = 0, \text{ for each } \{i, j, k\} \in T, \tag{7}$$

$$\sum_{j \sim i} \omega_{ij} \hat{Y}_{ij}^h = 0, \text{ for each } i \in V. \tag{8}$$

where $\hat{Y}^g$ satisfies (6) and $\hat{Y}^h$ satisfies two conditions (7) and (8). The residue $\hat{Y}^c$ actually satisfies (8) but not (7). Residues $\hat{Y}^h$ and $\hat{Y}^c$ account for inconsistencies of the global ranking obtained which show the validity of the ranking and can be further studied in terms of its geometric scale, namely whether inconsistency in the ranking data arises locally or globally. Local inconsistency can be fully characterized by triangular cycles (e.g. $i \succ j \succ k \succ i$), while global inconsistency involves loops consisting nodes more than three (e.g. $i \succ j \succ k \succ ... \succ i$), which may arise due to data incompleteness and once presented with a large component indicates some serious conflicts in ranking data.

Global rating score can be the minimal norm least square solution $\hat{s}$ of the following normal equation

$$\triangle_0 \, \hat{s} = \delta_0^* \hat{Y} \tag{9}$$

where $\delta_0 : \mathrm{R}^V \to \mathrm{R}^E$ is a finite difference operator (matrix) on G defined by $\delta_0((i, j), i) = -1$, $\delta_0((i, j), j) = 1$, and otherwise zero, $\delta_0^* = \delta_0^T W$ ($W = \mathrm{diag}(\omega_{ij})$), $\Delta_0 = \delta_0^* \cdot \delta_0$ is the unnormalized graph Laplacian defined by $(\Delta_0)_{ii} = \sum_{j \sim i} \omega_{ij}$ and $(\Delta_0)_{ij} = -\omega_{ij}$, and $(\cdot)^\dagger$ is the Moore-Penrose (pseudo) inverse.

An interesting variation of this l_2-norm scheme (3) is an analogous l_1-projection onto the space of gradient flows,

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j} \omega_{ij} |s_i - s_j - \hat{Y}_{ij}|. \tag{10}$$

This optimization problem is applied to the case that the noise is sparse but can be large, often regarded as outliers. It is more robust to outliers when compared with the l_2-norm, and thus can be regarded as robust ranking. For more details, readers may refer to [19, 23].

As the input of this HodgeRank framework is a paired comparison multigraph (the whole set of paired comparison data in one batch) provided by participants, we may call this type of work as batch HodgeRank. For details of the theoretical development, readers may refer to [19]. The work in [40] adopts such batch HodgeRank to obtain quality scores of videos. However, for crowdsourcing test on Internet, participants and image pairs enter the system one by one in a dynamic and random way. Therefore, batch HodgeRank is not an efficient tool for crowdsourcing. To meet this challenge, we propose an online HodgeRank as Robins-Monro procedure or stochastic approximation of (9).

3.2 Online Rating Algorithms

The online rating algorithm considered in this paper is constructed from Robbins-Monro procedure [28] to solve linear operator equation $\bar{A}x = \bar{b}$,

$$x_{t+1} = x_t - \gamma_t(A_t x_t - b_t),\ E(A_t) = \bar{A},\ E(b_t) = b. \quad (11)$$

Now consider the normal equation (9) for the least square problem (2), $\Delta_0 s = \delta_0^* \hat{Y}$. In this case, at time t when a new rating $Y_t(i_t, j_t) = -Y_t(j_t, i_t)$ entered on pair (i_t, j_t), we have

- A_t is a $|V|\times|V|$ matrix defined by $A_t(i_t, i_t) = A_t(j_t, j_t) = -A_t(i_t, j_t) = -A_t(j_t, i_t) = 1$ and otherwise zero;
- b_t is a $|V|$-dimensional vector defined by $b_t(i_t) = -b_t(j_t) = Y_t(i_t, j_t)$ and otherwise zero.

Let $s_t = x_t$. The Robbins-Monro procedure becomes

$$\begin{aligned} s_{t+1}(i_t) &= s_t(i_t) - \gamma_t[s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)] \\ s_{t+1}(j_t) &= s_t(j_t) + \gamma_t[s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)] \end{aligned} \quad (12)$$

where the initial choice is $s_0 = 0$ or any vector such that $\sum_i s_0(i) = 0$, and the step size γ_t is a nonnegative sequence whose choice is often taken in the following form

$$\gamma_t = \frac{a}{(t + t_0)^\theta}, \quad \theta \in [0, 1].$$

The choice of step size will be discussed in more detail in the next subsection with a convergence analysis which shows minimax rates with independent and identically distributed sampling. Algorithm 1 below shows the procedure of this online rating method.

For the sake of comparison, we also present a stochastic subgradient method for online rating with l_1-norm in (10), which is given by:

$$\begin{aligned} s_{t+1}(i_t) &= s_t(i_t) - \gamma_t\ \mathrm{sign}(s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)) \\ s_{t+1}(j_t) &= s_t(j_t) + \gamma_t\ \mathrm{sign}(s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)) \end{aligned} \quad (13)$$

with similar choices on initial score and steps.

Note that updates here only occur locally on the nodes associated with edge $\{i_{t+1}, j_{t+1}\}$, which is suitable for asynchronized parallel implementation.

Note that for l_1-based online algorithm it suffices to change $g_{ij} = \mathrm{sign}(s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t))$.

Algorithm 1: Online Rating Procedure.

Initialization:
$s_0 = 0$ or any vector such that $\sum_i s_0(i) = 0$; // Initialize the quality scores of each images.
With a new rating $Y_t(i_t, j_t)$; // A new paired comparison (i_t, j_t) occurs at time t.
Compute $g_{ij} = s_t(i_t) - s_t(j_t) - Y_t(i_t, j_t)$;
Then
$s_{t+1}(i_t) = s_t(i_t) - \gamma_t * g_{ij}$;
$s_{t+1}(j_t) = s_t(j_t) + \gamma_t * g_{ij}$. // Quality scores at time t+1.

3.3 Convergence Analysis

There have been studies on convergence analysis of subgradient methods, e.g. [31]. Typical convergence results require the conditions that step sizes $\sum_t \gamma_t^2 < \infty$ while $\sum_t \gamma_t = \infty$, and boundedness of subgradients, which are in particular $s(i) - s(j) - Y(i, j)$ and $\mathrm{sign}(s(i) - s(j) - Y(i, j))$ here. When general convex loss functions are assumed, the analysis is typically formulated as regret bounds [27].

In particular, when the square loss is adopted, one may achieve the following probabilistic upper bound, whose proof is given by [21] for general edge independent random graphs, which in fact reaches the minimax optimal rates for parametric regression up to a logarithmic factor.

In the following theorem, assume that $Y_t(i_t, j_t)$ is an independent and identically distributed (I.I.D.) sequence. For example, each rater follows a sampling on Erdös-Rényi random graph. Another example is Multiplicative-Attribute Graph Models [20] once node attributes are given which is however not pursued in this paper. The convergence analysis can be based on general Robbins-Monro procedure (11) with independent sampling sequence.

Define a random matrix

$$\Pi_k^t = \begin{cases} (I - \gamma_t A_t) \dots (I - \gamma_k A_k), & k \le t; \\ I, & k > t. \end{cases} \quad (14)$$

If we replace A_i by $\bar{A}$, we obtain a deterministic positive definite matrix, say $\bar{\Pi}_k^t$.

The following lemma leads to a martingale decomposition for error $x_t - x^*$, given in [32, 41], which is crucial to lead to the error bounds.

Lemma. For all $t \in \mathbb{N}$,

$$x_t = \Pi_1^{t-1} x_0 + \sum_{k=1}^{t-1} \gamma_k \Pi_{k+1}^{t-1} b_t \quad (15)$$

and

$$x_t - x^* = \bar{\Pi}_1^{t-1}(x_0 - x^*) - \sum_{k=1}^{t-1} \xi_k, \quad (16)$$

where

$$\xi_k = \begin{cases} \gamma_k \bar{\Pi}_{k+1}^{t-1}((A_k - \bar{A})x_k - (b_k - \bar{b})), & 1 \le k < t; \\ 0, & k \ge t. \end{cases}$$

is a martingale difference sequence such that $E[\xi_t : \mathcal{F}_{t-1}] = 0$ for a filtration $\mathcal{F}_{t-1}$ up to time $t-1$.

The first part in error, $\bar{\Pi}_1^{t-1}(x_0 - x^*)$, is called the *initial error* and the martingale difference tail, $\sum \xi_k$, is called the *sample error*. Initial error can be bounded deterministically, while the sample error can be bounded via a Pinelis-Bernstein probabilistic inequality. Combining these bounds

will lead to the following theorem, whose derivation follows closely [41].

Theorem 3.3. *Let E consists of the edge set of the expected graph, and $0 = \lambda_0 < \lambda_1 \leq \ldots \leq \lambda_{n-1}$ are eigenvalues of expected graph Laplacian $\Delta_0 = E(A_t)$. Assume that $A = 2 \vee \lambda_{n-1}$ and $|Y_t(i,j)| \leq B$. Then there exists a choice of step size $\gamma_t = a/(t+t_0)$ (e.g. $a = 1/\lambda_1$ and $t_0 \geq B/\lambda_1$) such that the following holds for all $t \in \mathbb{N}$ with probability at least $1-\delta$ ($\delta > 0$),*

$$\|s_t - s^*\|_2 \leq \frac{7\sqrt{A}B|E|}{\lambda_1^{3/2}} t^{-1/2} \log(t+t_0) \cdot \log \frac{2}{\delta}$$

where s_t is defined by (12).

The theorem says that the online rating algorithm converges to the underlying true score s^* under independent sampling process. The convergence rate is minimax optimal at $O(t^{-1/2})$. The choice of step size $\gamma_t \sim t^{-1}$ is crucial, with large enough t_0. Although the choice of a and t_0 does not affect the asymptotic rate in theory, in practice they influence the speed of convergence when t is small. We shall see this in experimental section. Moreover, in our applications, we find the performance distinctions are ignorable between two types of online algorithms, least square (12) and least absolute value (13), whence we do not pursue a thorough convergence analysis here for l_1-based online rating (13).

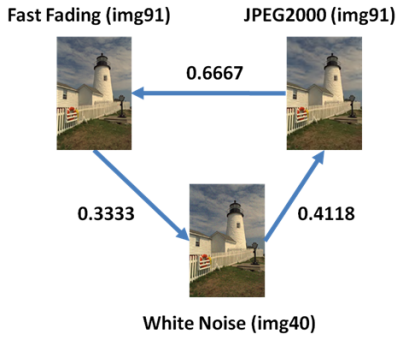

Figure 1: Large curl due to multicriteria in paired comparisons among users. The image is undistinguishable due to its small size, so image IDs in LIVE database are printed here.

3.4 Online Tracking of Triangular Curls

Hodge decomposition (5) has a component $\hat{Y}^c$ which satisfies $\hat{Y}^c_{ij} + \hat{Y}^c_{jk} + \hat{Y}^c_{ki} \neq 0$ for each triangle $(i,j,k) \in T$. This encodes the information about triangular or local inconsistency. For a graph $G = (V,E)$ whose 3-clique complex $\chi_G = (V,E,T)$ does not contain a "loop" (i.e. the first Betti number $\beta_1 = 0$), global inconsistency vanishes and such triangular inconsistency explains all sorts of inconsistency. It happens when Erdös-Rényi random graphs and k-regular random graphs are sufficiently dense [40, 39]. Due to such an importance, it is desired to track triangular curls:

$$curl_{ijk} = \hat{Y}^c_{ij} + \hat{Y}^c_{jk} + \hat{Y}^c_{ki} = \hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki}.$$

which is nothing but triangular trace of $\hat{Y}$ [19]. Curl is easy for online and parallel realizations. In [40], another relative curl is introduced as extensions of combinatorial intransitive triangles,

$$rel\text{-}curl_{ijk} = \frac{|\hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} \in [0,1].$$

Relative curl on a triangle $(i,j,k) \in T$ is one if and only if (i,j,k) is intransitive.

The existence of large curls or intransitive triangles may be either due to noise or suggesting the existence of multicriteria in paired comparisons. If the latter case happens on a triangule (i,j,k), on each edge say $(i,j) \in E$, it will have a $\hat{Y}_{ij}$ consistently away from zero, and incur a large curl. In Figure 1, we exhibit one example of such intransitive triangle existing in the data we collected so far, which indicates a stable cyclic preference on a natural scene picture in LIVE dataset such that JPEG2000 (img91) is better than Fast Fading (img91), Fast Fading (img91) is better than White Noise (img40), and White Noise (img40) is better than JPEG2000 (img91). This is due to the fact when different pairs of images are presented to raters, different salient features are adopted by raters implicitly. Triangular curls due to noise will vanish when the sample size goes to infinity while curls due to multicriteria will persist with the increase of sample complexity. Therefore, online tracking of curls will be useful to identify such a kind of inconsistency.

Algorithm 2 outlined below shows how to track the triangular curl in an online way.

Algorithm 2: Online Tracking of Curls.

With a new rating $Y^{(t)}_{ij}$;

$n^{(t+1)}_{ij} = n^{(t)}_{ij} + 1$; // $n^{(t)}_{ij}$ is the number of paired comparisons up to time t.

$\hat{Y}^{(t+1)}_{ij} = (1 - 1/n^{(t+1)}_{ij})\hat{Y}^{(t)}_{ij} + Y^{(t)}_{ij}/n^{(t+1)}_{ij}$; // $\hat{Y}^{(t)}_{ij}$ follows the same definition in Section 3.1.

for *each* k *s.t. (i,j,k) is a triangle* **do**

$curl^{(t+1)}_{ijk} = \hat{Y}^{(t+1)}_{ij} + \hat{Y}^{(t+1)}_{jk} + \hat{Y}^{(t+1)}_{ki}$

$rel\text{-}curl^{(t+1)}_{ijk} = \frac{|curl^{(t+1)}_{ijk}|}{|\hat{Y}^{(t+1)}_{ij}| + |\hat{Y}^{(t+1)}_{jk}| + |\hat{Y}^{(t+1)}_{ki}|}$

end

3.5 Online Tracking of Topology Evolution

The work in [40] shows that when the resultant graph provided by assessors is connected, we can derive global scores for all the images in comparison from batch HodgeRank. Besides, when its clique complex is loop-free, there is no global inconsistency whence tracking local inconsistency (triangular curls) presented above will be enough. Motivated by these two observations, [40] adopts persistent homology [14, 42, 6, 13] to check if a given graph instance satisfies the two conditions.

In fact, persistent homology is an online algorithm to check topology evolution when nodes, edges and triangles enter in a sequential way. Here we just discuss in brief the application of persistent homology to monitor the number of connected components (β_0) and loops (β_1) in our online settings. In random graph designs for image comparisons, we can assume that the images (nodes) are created at the same time, after that pairs of images (edges) are presented to assessors independently one by one. A triangle $\{i,j,k\}$ is created immediately when all the three associated edges

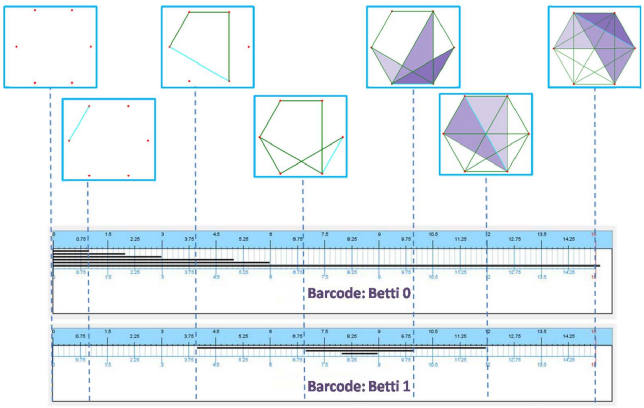

Figure 2: An example of persistence Barcodes of Betti numbers .

appeared. In practice with sampling of multigraph data, one may consider certain thresholds on edges and triangles for their presence, which can be dealt with in a similar way. With such a streaming data, persistent homology may return the number of β_0 and β_1 at each time when a new node/edge/triangle is born.

Figure 2 illustrates an example of this birth process and its associated Betti numbers (β_0 and β_1) that are computed and plotted by JPlex [30]. At the first frame (say $t = 0$), 6 images as nodes are collected, which corresponds to $\beta_0 = 6$ at $t = 0$ in Barcode: Betti 0. On the second frame ($t = 1$), an edge connecting a pair of nodes is created which drops the number of connected components from 6 to 5, i.e. $\beta_0 = 5$ at $t = 1$ in Barcode: Betti 0. The same procedure follows and particularly at the fifth frame $t = 4$, it creates a loop and there are 3 connected components in the graph, which can be read from $\beta_0 = 3$ at $t = 4$ and $\beta_1 = 1$ at $t = 4$, respectively. Note that after the thirteenth frame $t = 12$, there is only one connected component $\beta_0 = 1$ left and no loop exists $\beta_1 = 0$ as indicated by the Barcodes.

4. EXPERIMENTS

In this section, we systematically evaluate the performance of the proposed online HodgeRank algorithm against batch HodgeRank. First, the datasets used for the experiments are briefly explained. Then we present the experimental design of obtaining online paired comparison data, followed by the results with online and batch methods. Finally, we show how to track the curls and topological evolution online with persistent homology.

4.1 Datasets

Two publicly available datasets, LIVE [3] and IVC [2], are used in this work. The LIVE dataset contains 29 reference images and 779 distorted images. The distorted images are obtained using five different distortion processes–JPEG2000, JPEG, White Noise, Gaussian Blur, and Fast Fading Rayleigh. Considering the resolution limit of most test computers, we only choose 6 different reference images (480×720) and 15 distorted versions of each reference, for a total of 96 images. The second dataset, IVC, which is also

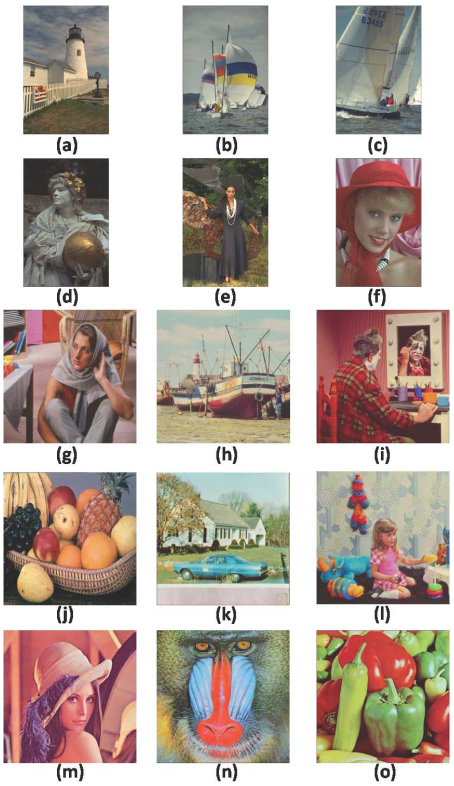

Figure 3: Images in LIVE and IVC databases (The first six are images from LIVE and the remaining images are from IVC).

a broadly adopted dataset in the community of IQA, includes 10 reference images and 185 distorted images derived from four distortion types–JPEG2000, JPEG, LAR Coding, and Blurring. Following the collection strategy in LIVE, we further select 9 different reference images (512×512) and 15 distorted images of each reference. Eventually, we obtain a medium-sized image set that contains a total of 240 images from 15 references, as illustrated in Figure 3. Note that we do not use the subjective scores in LIVE and IVC, but only borrow the image sources they provide. Different from them, we propose to assess image quality with paired comparison method. There are two aspects about the size of dataset: (1) number of distortion types; (2) number of reference images. The first is the number of nodes in our paired comparison graphs, which is n=16 here. Even on such a scale, it is almost impossible for a single person to perform all $\binom{n}{2}$ paired comparisons. So it suffices to illustrate the performance of online algorithm against batch algorithm. The second does not affect the computational complexity of algorithms, whence a random choice 15 from LIVE and IVC database is to show performance consistency over these examples.

4.2 Online Paired Data Collection

We now present our experiment design for collecting the set of online paired data. Different from traditional complete design in paired comparison, a session in our test can have an arbitrary duration (down to a single pair) and par-

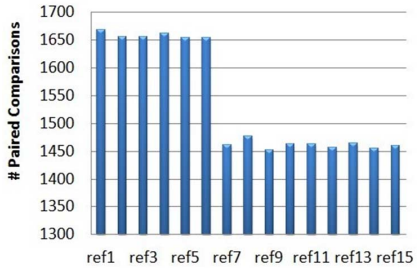

Figure 4: Number of paired comparisons each reference received in LIVE and IVC databases.

ticipants are free to decide when to quit. In other words, the number of pairs ($\#pairs$) shown to participants can be adjusted according to their time constraint and preference. That is, when a participant's time is adequate, $\#pairs$ can be a bigger value. But if he/she is under the pressure of time or prefers not to spend more time with the experiment , $\#pairs$ will be smaller.

Before starting the experiment, each participant is briefed about the goal of the experiment and given a short training session to familiarize themselves with the testing procedure. In the testing process, images are displayed side by side at their native resolutions to prevent any distortions due to scaling operations performed by software or hardware. Besides, to make it impossible for participants to cheat our system by inputting "smart" answers, the order of each pair and the order within each pair are totally random for each participant. Each assessor is allowed to take as much time as needed to enter their choice. However, the assessors could not change their choice once entered or view the image again. Once the choice is entered, the next image pair is displayed.

Moreover, we hope to avoid the situation with successive pairs of test images from the same reference, to avoid contextual and memory effects in their judgments of quality. For this purpose, after the playlist for one participant is constructed, our program would go over the entire playlist to determine if adjacent pairs correspond to the same reference. If such a case is detected, one of the pairs would be swapped with another randomly chosen pair in the playlist which does not suffer from the same problem.

Finally, 186 observers of different cultural level (students, tutors, and researchers), each of whom performs a varied number of comparisons via Internet, provide 23,097 paired comparisons in total. The number of responses each reference image receives is different, as illustrated in Figure 4. It should be noted that, for ref1-6 from LIVE database, we start our test from October 10th, 2011. Later, 9 references images (ref7-15) from IVC database are added to our test system from November 2nd, 2011. Our collecting task is still on-going now for further larger-scale studies.

4.3 Comparison with "Batch" HodgeRank

The experimental evaluation involves evaluating the online HodgeRank algorithms (12) on various data sets against the performance of batch HodgeRank. We also compare these algorithms against the l_1-norm online algorithm in (13) on the same data sets.

The metric that we used in the evaluation of the performance of various algorithms is the Mismatch Ratio (***MR***),

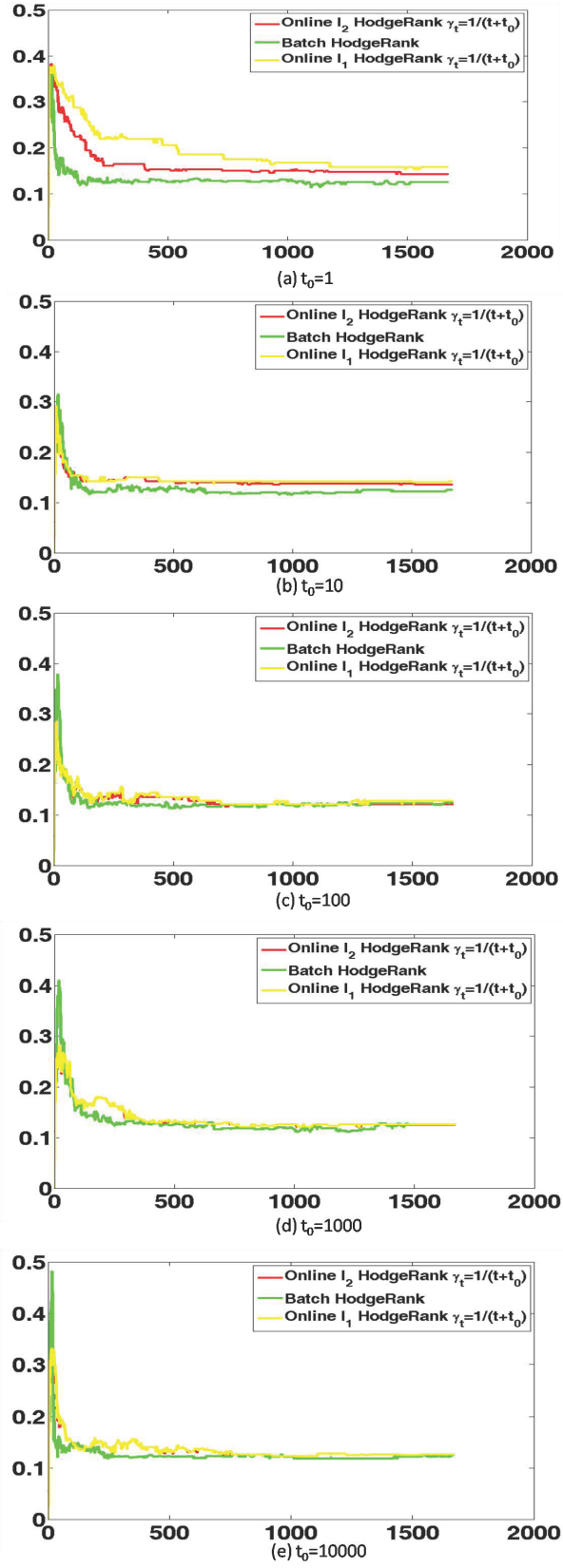

Figure 5: The impact of the t_0 parameter on the performances. *MR* (y-axis) versus the number of samples (x-axis) on reference1.

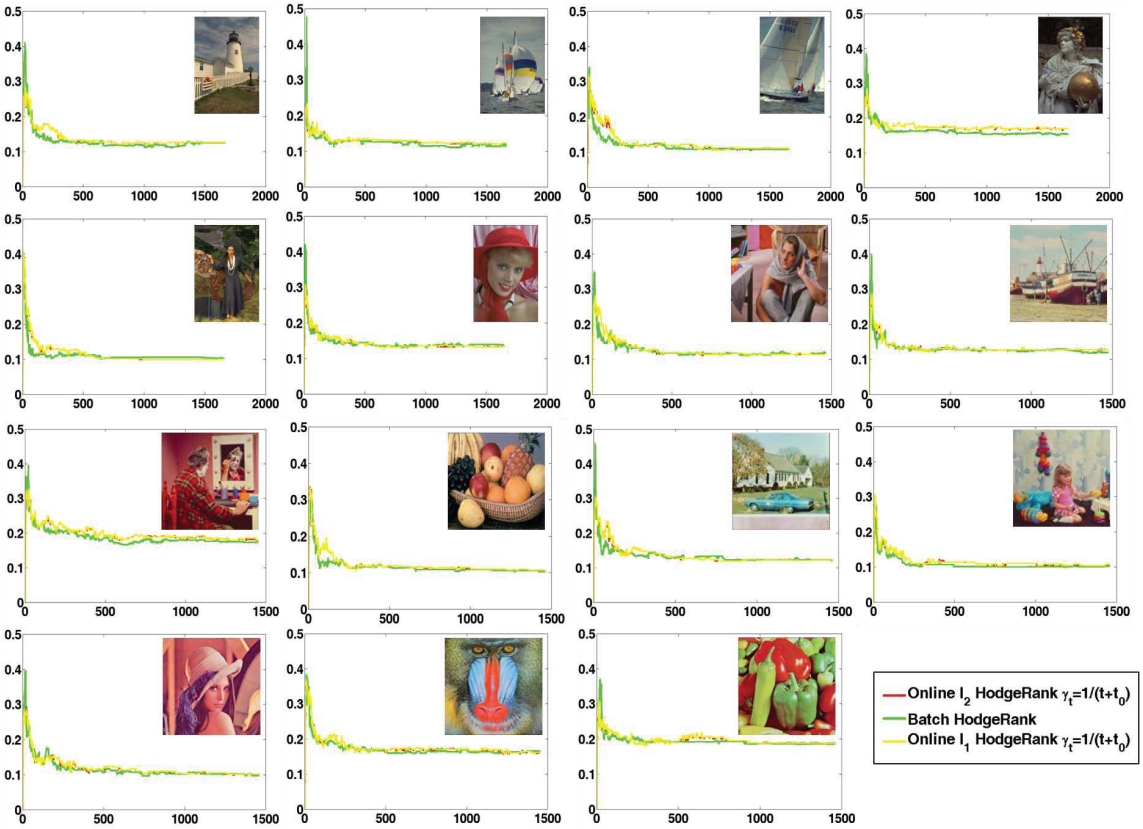

Figure 6: Experimental results of online HodgeRank vs. batch HodgeRank. *MR* (y-axis) versus the number of samples (x-axis) on 15 reference datasets.

i.e., at time t, the percentage of mismatch pairs of a global rating s_t made on all previous examples,

$$\epsilon_t = \frac{1}{2t}\sum_{\tau=1}^{t} |\text{sign}(s_t(i_\tau) - s_t(j_\tau)) - Y_\tau(i_\tau, j_\tau)|, \quad Y_\tau(i,j) \in \{\pm 1\}$$

Since in this paper we use the simplest binary choice for $Y_t(i,j)$, one natural performance measure for different algorithms is simply to count the ratio of mismatched paired comparisons up to time t. Minimization of such an objective function is known to be NP-hard. However, in our experiments we find the batch HodgeRank, online l_2 HodgeRank and online l_1 HodgeRank exhibit similar effects on minimizing such an objective function, whose distinctions thus can be ignored in practice like this.

Although in theory, the choice of t_0 won't affect asymptotic convergence rates (Theorem 3.3); one has to be careful on choosing it in practice. First, we study the impact of the t_0 parameter in (12) and (13) on the performances of online rating algorithms. For simplicity, we randomly take reference1 as an illustrative example while other reference images exhibit similar results. Figure 5 shows the influences of t_0 on mismatch ratios, in a comparison with batch HodgeRank. It can be seen from this figure that with the increase of collected data, the ***MR*** curve generally decreases. When t_0 is chosen to be small, l_2 online rating (12) shows better performance than l_1 online (13). However, when t_0 increases (e.g. from 1 to 10000), such a performance difference diminishes and both methods approach performance of batch HodgeRank. We note that when t_0 is chosen to be too large, initial tracking performance will drop which shows a lag behind of batch learning curve. Although this does not hurt the long term behavior eventually, those who care the initial iterations should be careful on this. In the following experiments, we will choose $t_0 = 1000$ as a balance of these effects for further studies.

Figure 6 shows the performance comparisons of online HodgeRank against batch HodgeRank with $t_0 = 1000$ for other 14 reference datasets. It is interesting to see that on all of these large scale data collections, both of these two online algorithms are able to maintain competitive performances with the batch case. From these results, we may conclude that large amplitude outliers are not a significant issue in our crowdsourcing data collection as $|Y_t(i,j)|$ is bounded by 1. Besides, Table 1 shows the computation complexity achieved by l_2 online HodgeRank, l_1 online HodgeRank and batch HodgeRank. It is easy to see that on our dataset, online HodgeRank can achieve up to nearly 370 times faster than batch HodgeRank, with similar prediction errors.

4.4 Online Tracking of Topology and Curls

In our online settings, due to the multiple comparisons between a pair of images, a natural question is raised that how many samples are needed to satisfy the connected & loop-free conditions? As each reference is similar in sampling scheme, we compute the online mean Betti numbers

Table 1: **Computation complexity (s) comparison of online and batch HodgeRank.**

	ref1	ref2	ref3	ref4	ref5	ref6	ref7	ref8	ref9	ref10	ref11	ref12	ref13	ref14	ref15	mean
l_2	0.166	0.158	0.159	0.162	0.164	0.163	0.125	0.128	0.131	0.125	0.130	0.128	0.133	0.135	0.133	0.143
l_1	0.159	0.162	0.165	0.166	0.162	0.163	0.132	0.127	0.125	0.124	0.125	0.123	0.124	0.129	0.132	0.141
Batch	59.28	60.78	58.25	58.65	60.09	58.22	53.15	49.58	47.45	47.81	47.84	48.01	50.29	47.40	47.43	52.95

over 15 references, as illustrated in Figure 7 (a). As we can see, after about 70 samples on this multigraph, with high probability the resultant graph is connected & loop-free. In other words, it is easy to meet these two requirements and thus can avoid the possible issue of harmonic inconsistency in global ranking.

In addition, we can further set a threshold for each edge which can be treated as a confidence level. That is to say, only edges on which the number of paired comparisons are larger than this threshold will be added in our resultant graph. The bigger the threshold is set, the more robust the topological structure of the graph is. Figure 7 (b) shows the online tracking of the first two Betti numbers by persistent homology when threshold is set to be 3. One can see more examples (250) are needed to reach the connected and loop-free condition.

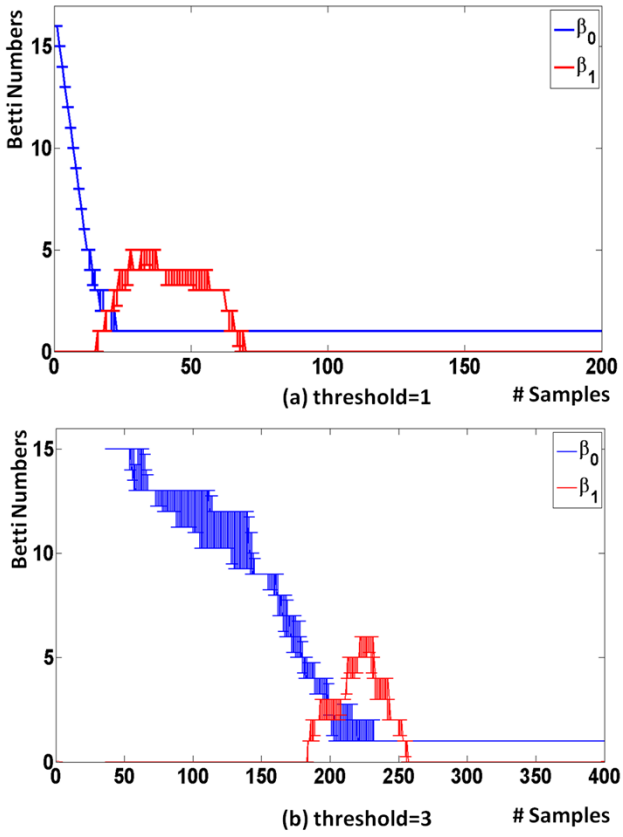

Figure 7: Number of samples versus number of online Betti numbers. For each sample number level, the median number of Betti numbers over 15 references with [0.25, 0.75] confidence interval are plotted in the figure.

Triangular curls and relative curls defined in the last section are helpful to identify possible inconsistency or the existence of multicriteria adopted by raters in different paired comparisons. By online tracking of relative curls in Figure 8, we find the intransitive triangle shown in Figure 1 that JPEG2000 (img91) is better than Fast Fading (img91), Fast Fading (img91) is better than White Noise (img40), and White Noise (img40) is better than JPEG2000 (img91). The phenomenon suggests that one should explore the hidden multicriteria behind the paired comparisons among these images which will be left for future studies.

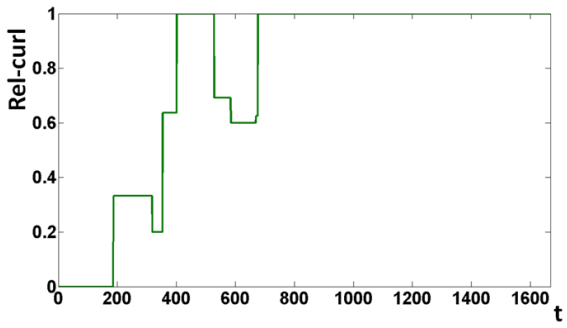

Figure 8: Online tracking of relative curl on triangle (JPEG2000 (img91), Fast Fading (img91), White Noise (img40)). One can see the intransitive triangle constantly appears over time which suggests possible different criteria adopted by users in paired comparisons made among them.

5. CONCLUSIONS

In this paper, online algorithms are proposed for crowdsourcing subjective image quality assessment where the data are collected in a streaming way. The algorithms are based on Robbins-Monro procedure or stochastic approximation to solve a HodgeRank problem on random graphs. Two variations are studied against the batch HodgeRank: one based on classical l_2-minimization or least square problem to deal with independent noise of zero mean and bounded variance, and the other based on l_1-minimization problem to deal with outliers. Experiments with the images available in LIVE and IVC databases are conducted, including 15 different reference images and 15 distorted versions of each reference in total. It is shown that in our applications, the l_2-based online HodgeRank can achieve as nearly good performance as batch HodgeRank, in both theory and experiments. Moreover, l_1-based online rating exhibits similar performance to l_2 online algorithm and thus batch HodgeRank in our experiments which indicates the binary choice in crowdsourcing data collection won't suffer much the outlier issue. Furthermore, we investigate the online tracking of

triangular curls and topology evolution of the paired ranking complex. In particular, we show that online tracking of triangular curls provides us important information about inconsistency, which may suggest the existence of multicriteria in rater's judgement of different object pairs.

Our studies show that online HodgeRank provides us an efficient approach to study large scale crowdsourcing subjective IQA on Internet. It enables us to derive global rating as well as monitor the inconsistency occurring in the data in the real time.

Additionally, we would like to point out here that the theory developed in this paper takes the standard I.I.D. sampling assumption as the main stream of statistical machine learning. It is a largely unexplored field for online learning with dependent sampling, such as Markov sampling and active sampling, which will be our future directions.

6. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, 2012CB825501, 2011CB809105, in part by National Natural Science Foundation of China: 61025011, 61071157, and 60833006.

7. REFERENCES

[1] *ITU-R Recommendation P.800. Methods for subjective determination of transmission quality*, 1996.

[2] Subjective quality assessment irccyn/ivc database. `http://www2.irccyn.ec-nantes.fr/ivcdb/`, 2005.

[3] LIVE image and video quality assessment database. `http://live.ece.utexas.edu/research/quality/`, 2008.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

[6] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[7] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. pages 491–500. ACM Multimedia, 2009.

[8] F. Chung and L. Lu. *Complex Graphs and Networks*. CBMS Regional Conference Series in Mathematics, American Mathematical Society, 2006.

[9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge Unversity Press, 2000.

[10] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. of the Amer. Math. Soc.*, 29(1):1–49, 2002.

[11] H. David. *The method of paired comparisons*. 2nd Ed., Griffin's Statistical Monographs and Courses, 41. Oxford University Press, New York, NY, 1988.

[12] Y. S. E. Granger and P. Lavoie. A pattern reordering approach based on ambiguity detection for online category learning. *PAMI*, 25:525–529, 2003.

[13] H. Edelsbrunner and J. Harer. Computational topology : an introduction. 2010.

[14] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

[15] A. Eichhorn, P. Ni, and R. Eg. Randomised pair comparison: an economic and robust method for audiovisual quality assessment. pages 63–68. NOSSDAV, 2010.

[16] P. Erdos and A. Renyi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[17] H. Grabner and H. Bischof. On-line boosting and vision. CVPR, 2006.

[18] G. H. H. Bekel, I. Bax and H. Ritter. Adaptive computer vision: Online learning for object recognition. DAGM, 2004.

[19] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 2010.

[20] M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160, 2012.

[21] L.-H. Lim and Y. Yao. Online ranking on random graphs. *preprint*, 2012.

[22] S. A. O. Javed and M. Shah. Online detection and classification of moving objects using progressively improving detectors. CVPR, 2005.

[23] B. Osting, J. Darbon, and S. Osher. Statistical ranking using the l_1-norm on graphs. *AIMS'*, 2012.

[24] M. Penrose. *Random Geometric Graphs (Oxford Studies in Probability)*. Oxford University Press, 2003.

[25] M. T. Pham and T. J. Cham. Online asymetric boosted classifiers for object detection. CVPR, 2007.

[26] Y. L. R. T. Collins and M. Leordeanu. Online selection of discriminative tracking features. *PAMI*, 27:1631–1643, 2005.

[27] A. Rakhlin. *Statistical Learning Theory and Sequential Prediction*. Lecture Notes in University of Pennsyvania, 2012.

[28] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[29] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[30] H. Sexton and M. Johansson. JPlex: a java software package for computing the persistent homology of filtered simplicial complexes. `http://comptop.stanford.edu/programs/jplex/`, 2009.

[31] N. Shor. *Minimization Methods for Non-Differenliable Functions*. Springer-Verlag, 1985.

[32] S. Smale and Y. Yao. Online learning algorithms. *Foundation of Computational Mathematics*, 6(2):145–170, 2006.

[33] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[34] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.

[35] B. Widrow and M. Hoff. Adaptive switching circuits. *IRE WESCON Convention Record*, (4):96–104, 1960.

[36] B. Widrow and M. A. Lehr. 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. pages 1415–1442. Proceedings of the IEEE, 1990.

[37] N. Wormald. Models of random regular graphs. pages 239–298. In Surveys in Combinatorics, 1999.

[38] B. Wu and R. Nevatia. Improving part based object detection by unsupervised, online boosting. CVPR, 2007.

[39] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao. HodgeRank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857, 2012.

[40] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin. Random partial paired comparison for subjective video quality assessment via HodgeRank. pages 393–402. ACM Multimedia, 2011.

[41] Y. Yao. On complexity issue of online learning algorithms. *IEEE Transactions on Information Theory*, 56(12):6470–6481, 2010.

[42] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.