Interestingness Prediction by Robust Learning to Rank*

Yanwei Fu[1], Timothy M. Hospedales[1], Tao Xiang[1,**], Shaogang Gong[1], and Yuan Yao[2,**]

[1] School of EECS, Queen Mary University of London, UK
[2] School of Mathematical Sciences, Peking University, China
{y.fu,t.hospedales,t.xiang,s.gong}@qmul.ac.uk, yuany@math.pku.edu.cn

Abstract. The problem of predicting image or video interestingness from their low-level feature representations has received increasing interest. As a highly subjective visual attribute, annotating the interestingness value of training data for learning a prediction model is challenging. To make the annotation less subjective and more reliable, recent studies employ crowdsourcing tools to collect pairwise comparisons – relying on majority voting to prune the annotation outliers/errors. In this paper, we propose a more principled way to identify annotation outliers by formulating the interestingness prediction task as a unified robust learning to rank problem, tackling both the outlier detection and interestingness prediction tasks jointly. Extensive experiments on both image and video interestingness benchmark datasets demonstrate that our new approach significantly outperforms state-of-the-art alternatives.

1 Introduction

The problem of automatically predicting if people would find an image or video interesting has started to receive increasing attention [7,16,21]. Interestingness prediction has a number of real-world applications. In particular, since the number of images and videos uploaded to the Internet is growing explosively, people are increasingly relying on image/video search engines or recommendation tools to select which ones to view. Given a query, ranking the retrieved data with relevancy to the query based on the predicted interestingness would improve the user satisfaction. Similarly user stickiness can be increased if a media-sharing website such as YouTube can recommend videos that are both relevant and interesting. Other applications such as web advertising and video summarisation can also benefit.

Learning a computational model of how humans perceive interestingness is however extremely challenging due to the following two reasons. First, what

* The research of Yuan Yao was supported in part by National Basic Research Program of China (973 Program 2012CB825501), NSFC grant 61071157, and a joint NSFC-Royal Society grant 61211130360, IE110976 with Tao Xiang.

** Corresponding authors.

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 488–503, 2014.

defines interestingness and what cues contribute to the human perception of interestingness are still under investigation in psychophysics [39], cognitive sciences [4] and recently computer vision [7,16,21]. Therefore current research in computer vision on interestingness is primarily focused on designing relevant feature representations. Second, in order to predict interestingness from low-level features, training data with labelled interestingness values are required. This is problematic because as a highly subjective visual attribute, directly annotating an interestingness value for a data point is unreliable, e.g. on a scale of 1 to 10, 10 being the most interesting, different people will have very different ideas on what a scale 5 means for an image, especially without any common reference point.

In order to obtain more reliable interestingness annotation and thus learn better prediction models, recent studies [16,21] propose to model interestingness from human pairwise comparison data collected using crowdsourcing tools such as Amazon Mechanic Turk (AMT). The annotation task is to select between a pair of images or videos which one is more interesting. This is considered to be a much easier task, resulting in more reliable annotations. However, this brings about two new problems: (1) *sparsity* – the number of pairwise comparisons required is much bigger than for directly annotated interestingness values (there are $n^2 - n$ pairs give n data points); even with crowdsourcing tools, the annotation will be sparse, i.e. not all pairs are compared and each pair is only compared few times. (2) *Outliers* – it is well known that crowdsourced data are noisy [6,43,30]. Existing approaches [16,21] solve the outlier problem by majority voting which requires multiple comparisons for each pair of data points; but its effectiveness is severely limited by the sparsity of the data.

In this paper we propose a novel approach for predicting interestingness from sparse and noisy pairwise comparison data. Different from existing approaches which first remove outliers by majority voting, followed by regression [16] or learning to rank [21], we formulate a unified robust learning to rank framework to jointly solve both the outlier detection and interestingness prediction problems. Critically, instead of detecting outliers locally and independently at each pair by majority voting, our outlier detection method operates globally integrating all local pairwise comparisons together to minimise a cost that corresponds to global inconsistency of ranking order. This enables us to identify outliers that receive majority votes yet cause large global ranking inconsistency and thus should be removed. Furthermore, as a global method, only one comparison per pair is required, therefore significantly reducing the data sparsity problem compared to the conventional majority voting approach. Extensive experiments on benchmark image and video interestingness datasets demonstrate that our method significantly outperforms the state-of-the-art alternatives. In addition, since interestingness is a special case of relative attributes, we also validate our method on predicting more general image relative attributes for image classification tasks.

2 Related Work

Predicting Image and Video Interestingness. Early efforts on image interestingness prediction focus on different aspects than interestingness as such, including image quality [22], memorability [19], and aesthetics [7]. These properties are related to interestingness but different. For instance, it is found that memorability can have a low correlation with interestingness - people often remember things that they find uninteresting [16]. The work of Gygli et al [16] is the first systematic study of image interestingness. It shows that three cues contribute the most to interestingness: aesthetics, unusualness/novelty and general preferences, the last of which refers to the fact that people in general find certain types of scenes more interesting than others, for example outdoor-natural vs. indoor-manmade. Different features are then designed to represent these cues as input to a prediction model. In comparison, video interestingness has received much less attention, perhaps because it is even harder to understand its meaning and contributing cues. [28] focuses on key frames so essentially treats it as an image interestingness problem, whilst [21] is the first work that proposes benchmark video interestingness datasets and evaluates different features for video interestingness prediction. In a broader sense of attributes [26,11,12,27,13] interestingness can be considered as one type of relative attributes [35], although those attributes, such as how smiling a person is, are much less subjective.

Computational Models of Interestingness. Most earlier work casts the aesthetics or interestingness prediction problem as a regression problem [22,7,19,28]. However, as discussed before, obtaining an absolute value of interestingness for each data point is too subjective and affected too much by unknown personal preference/social background to be reliable. Therefore the most recent two studies on image [16] and video [21] interestingness all collect pairwise comparison data by crowdsourcing. Both use majority voting to remove outliers first. After that the prediction models differ – [16] converts pairwise comparisons into an absolute interestingness values and use a regression model, whilst [21] employs rankSVM [3] to learn a ranking function, with the estimated ranking score of an unseen video used as the interestingness prediction. We compare with both approaches in our experiments and demonstrate that our unified robust learning to rank approach is superior as we can remove better outliers – even if they correspond to comparisons receiving majority votes – thanks to its global formulation.

Learning from Noisy Crowdsourced Data. Beyond interesting prediction, many large-scale computer vision problems rely on human intelligence tasks (HIT) using crowdsourcing services, e.g. AMT (Amazon Mechanical Turk) to collect annotations. Many studies [23,40,36,30] highlight the necessity of validating the random or malicious labels/workers and give some filtering heuristics for data collection. However, existing approaches to annotation noise are primarily based on majority voting which requires a costly volume of redundant annotations. Moreover, as a local (per-pair) inconsistency filtering method, it has no effect on global inconsistency and even risks introducing additional inconsistency due to the well-known Condorcet's paradox [15].

Robust Learning to Rank. Statistical ranking has been widely studied in statistics [20,10] and computer science [44,45]. By aggregating pairwise local rankings into a global ranking, methods such as Huber-LASSO [46,18] have the potential to be robust against local ranking noise [5,31]. However, statistical ranking only concerns the ranking of the observed/training data, but not learning to predict unseen data by learning ranking functions. To learn such ranking functions for applications such as interestingness prediction, a feature representation of the data points must be used as model input in addition to the local ranking orders. This is addressed in learning to rank which is widely studied in machine learning [1,29,41,2]. However, existing learning to rank work does not explicitly model and remove outliers for robust learning: a critical issue for learning from crowdsourced data in practice. In this work, for the first time, we study the problem of robust learning to rank given extremely noisy and sparse crowdsourced pairwise labels. We show both theoretically and experimentally that by solving both the outlier detection and ranking estimation problems jointly, we achieve better outlier detection than existing statistical ranking methods and better ranking prediction than existing learning to rank method such as rankSVM without outlier detection.

Our Contributions. are threefold: (1) We propose a novel robust learning to rank method for interestingness prediction from noisy and sparse pairwise comparison data. (2) For the first time, the problems of detecting outliers and estimating ranking score are solved jointly in our unified framework. (3) We demonstrate both theoretically and experimentally that our method is superior to existing majority voting based methods as well as statistical ranking based methods.

3 A Unified Robust Learning to Rank (URLR) Framework

3.1 Problem Statement

We aim to learn an interestingness prediction model from a set of sparse and noisy pairwise comparisons, each comparison corresponding to a local ranking between a pair of images or videos. Suppose our training set has I data points/instances represented by a low-level feature matrix $\Phi = \left[\phi_i^T\right]_{i=1}^I \in R^{I\times d}$, where ϕ_i is a d-dimensional column feature vector for representing instance i. The annotations or data labels are represented as an annotation matrix Y. In particular, assume each pair of instances on average receive K votes by annotators. We will have $Y_{ij}^k = 1$ if the k-th vote indicates that instance i is more interesting than instance j, and $Y_{ji}^k = 1$ otherwise. The annotation matrix is then constructed as $Y_{ij} = \sum_k Y_{ij}^k$. These pairwise comparisons can be naturally represented by a directed graph $G = (V, E)$ with node set $V = \{i\}_{i=1}^I$ and edge set $E = \{i \to j | Y_{ij} > 0\}$. That is, an edge $i \to j$ exists if $Y_{ij} > 0$.

Given the training data Φ and Y, there are two tasks: (1) removing the outliers in Y and (2) estimating an interestingness prediction function. In this work a linear function is considered due to its low computational complexity, that is, given the low-level feature ϕ_x of a test instance x we use a linear function $f(x) = \beta^T\phi_x$ to predict its interestingness, where β is the coefficient weight vector of the low-level feature ϕ_x. All formulations can be easily updated to use a non-linear function.

Note that the annotation matrix Y is not symmetric – in an ideal case, one hopes that the votes received on each pair are unanimous, e.g. $Y_{ij} > 0$ and $Y_{ji} = 0$; but often there are disagreements, i.e. both $Y_{ij} > 0$ and $Y_{ji} > 0$. Assuming both cannot be true simultaneously, one of them will be an outlier. In this case, one is the majority and the other minority which will be pruned by the majority voting method. This is why majority voting is a local outlier detection method and requires as many votes per pair as possible to be effective (the wisdom of a crowd).

3.2 Framework Formulation

We propose to prune outliers globally. To this end, we introduce an unknown variable γ_{ij} for each element of Y which indicates whether Y_{ij} is an outlier. We thus aim to estimate both γ_{ij} for outlier detection and β for interestingness prediction in a unified framework. Specifically, for each edge $i \to j \in E$, Y_{ij} is modelled as,

$$Y_{ij} = \beta^T\phi_i - \beta^T\phi_j + \gamma_{ij} \tag{1}$$

where $\gamma_{ij} \in R$ is a variable that indicates annotation outliers. For an edge $i \to j$, if Y_{ij} is not an outlier, we expect $\beta^T\phi_i - \beta^T\phi_j$ should be approximately equal to Y_{ij}, therefore we have $\gamma_{ij} = 0$. On the contrary, when the prediction of $\beta^T\phi_i - \beta^T\phi_j$ differs greatly from Y_{ij}, we can explain Y_{ij} as an outlier and compensate for the discrepancy between the prediction and the annotation with a nonzero value of γ_{ij}. The only prior knowledge we have on γ_{ij} is that it is a sparse variable, i.e. in most cases $\gamma_{ij} = 0$.

For the whole training set, Eq (1) is written in its matrix form

$$Y = C\Phi\beta + \Gamma \tag{2}$$

where $Y = [Y_{ij}]$, $\Gamma = [\gamma_{ij}]$, and C is the incident matrix of the directed graph G, where $C_{ie} = -1/1$ if the edge e leaves/enters vertex i.

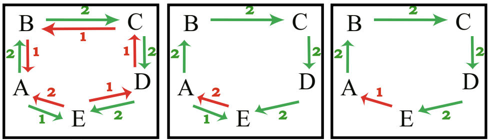
In order to estimate the $I^2 + d$ unknown parameters (I^2 for Γ and d for β), we aim to minimise the discrepancy between the annotation Y and our prediction $C\Phi\beta + \Gamma$, as well as keeping the outlier estimation Γ sparse. To that end, we put a l_2-loss on the discrepancy and a l_1- penalty on the outlier variables as a regularisation measure. This gives us the following cost function:

$$\min_{\beta,\Gamma} \frac{1}{2}\|Y - C\Phi\beta - \Gamma\|_2^2 + \lambda\|\Gamma\|_1 \tag{3}$$

$$:= \sum_{i\to j\in E}\left[\frac{1}{2}(Y_{ij} - \gamma_{ij} - \beta^T\phi_i + \beta^T\phi_j)^2 + \lambda|\gamma_{ij}|\right] \tag{4}$$

where λ is a free parameter corresponding to the weight for the regularisation term. With this cost function, our Unified Robust Learning to Rank (URLR) framework identifies outliers globally by integrating all local pairwise comparison together.

Figure 1(a) illustrates why our URLR framework is advantageous over the local majority voting method for outlier detection. Assume there are five images $A - E$ with five pairs compared three time each, and the correct ranking order of these 5 images in terms of interestingness is $A < B < C < D < E$. Figure 1(a) shows that among the five compared pairs, majority voting can successfully identify four outlier cases: $A > B$, $B > C$, $C > D$, and $D > E$, but not the fifth one $E < A$. However when considered globally, it is clear that $E < A$ is an outlier because if we have $A < B < C < D < E$, we can deduce $A < E$. Our formulation can detect this tricky outlier. More specifically, if the estimated β makes $\beta^T \phi_A - \beta^T \phi_E > 0$, it has a small local inconsistency cost for that minority vote edge $A \to E$. However, such β value will be 'propagated' to other images by using the voting edges $B \to A$, $C \to B$, $D \to C$, and $E \to D$, which are accumulated into much bigger global inconsistency with the annotation. This makes our model detect $E \to A$ as an outlier, contrary to the majority voting decision. In particular, the majority voting will introduce a loop comparison $A < B < C < D < E < A$ which is the well-known Condorcet's paradox [15]. We further give two more extreme cases in Fig. 1(b) and (c). Due to such Condorcet's paradox, in Fig. 1(b) the estimated β from majority voting is even worse than that from all annotation pairs which at least save the right annotation $A \to E$. Furthermore, Fig. 1(c) shows that when each pair only receives votes along one direction, majority voting will cease to work altogether, but our URLR can still detect outliers by examining the global cost.



(a) (b) (c)

Fig. 1. Better outlier detection can be achieved using our URLR framework than majority voting. Green arrows indicate correct annotations, while red arrows are outliers.

3.3 Problem Decomposition

To solve Eq (3), we rewrite the cost function as,

$$L(\beta, \mathit{\Gamma}) = \frac{1}{2} \|Y - X\beta - \mathit{\Gamma}\|_2^2 + \lambda \|\mathit{\Gamma}\|_1. \tag{5}$$

where $X = C\mathit{\Phi}$. With $\frac{\partial L}{\partial \beta} = 0$, we have

$$\hat{\beta} = (X^T X)^\dagger X^T (Y - \mathit{\Gamma}). \tag{6}$$

The Moore-Penrose pseudo-inverse of X^TX is equivalent to the limit of ridge regression solution: $(X^TX)^\dagger = \lim_{\mu \to 0}((X^TX)^T \cdot (X^TX) + \mu \mathbf{1})^{-1}(X^TX)^T$, where $\mathbf{1}$ is the eye matrix. To avoid numerical instability in many practical applications, we can replace the pseudo-inverse with ridge regression by setting $\mu > 0$. The standard solvers for Eq (6) will require $O(I^3)$ computational complexity. To reduce the complexity, the Krylov iterative and algebraic multi-grid methods [17] can be used.

Now plugging the solution of $\hat{\beta}$ back into Eq (5) and defining the hat matrix $H = H(X) = X(X^TX)^{-1}X^T$, we have

$$\hat{\varGamma} = \arg\min_{\varGamma} \|Y - \varGamma - H(Y - \varGamma)\|_2^2 + \lambda \parallel \varGamma \parallel_1 \tag{7}$$

The first term in Eq (7) is L_2- loss of the residuals of the observations $Y - \varGamma$ without the outliers $\varGamma$ which is: $r = Y - \varGamma - H(Y - \varGamma) = (I - H)(Y - \varGamma)$. Interestingly, Eq (7) does not rely on the estimation of $\hat{\beta}$. We therefore can now decompose the optimisation problem (5) into two intervening sub-problems: outlier detection in (7) and estimation of β using (6).

3.4 Outlier Detection by Regularisation Path

For outlier detection, we can further simplify Eq (7) by Singular Value Decomposition (SVD),

$$X = U\varSigma A^T \tag{8}$$

where $U = [U_1, U_2]$ with U_1 an orthogonal basis of the column space of X and A is the conjugate transpose of U. Therefore, due to the orthogonality $U^TU = I$ and $U_2X = 0$, Eq (7) is now a standard Least Absolute Shrinkage and Selection Operator (LASSO) estimator [9],

$$\hat{\varGamma} = \arg\min_{\varGamma} \|U_2^TY - U_2^T\varGamma\|_2^2 + \lambda\|\varGamma\|_1 \tag{9}$$

Nevertheless, tuning the regularisation parameter λ is a notoriously difficult problem. Especially in our URLR framework, the λ value directly decides the ratio of outliers detected and the ratio is unknown. A number of methods for determining λ exist, but none is suitable for our formulation: (1) some heuristics rules like $\lambda = 2.5\hat{\sigma}$ [1] are popular in existing robust ranking models such as the M-estimator [18]. However setting a constant λ value independent of dataset is far from optimal because the ratio of outliers may vary for different crowdsourcing experiments. (2) Cross validation is also not applicable here because each edge $i \to j$ is associated with a γ_{ij} variable and any held-out edge $i \to j$ also corresponds to an unknown variable γ_{ij}. As a result, cross validation can only optimise part of the sparse variables while leaving those for the held-out validation set undetermined. (3) The other alternatives e.g. Akaike information criterion (AIC) and Bayesian information criterion (BIC) employ the relative

[1] $\hat{\sigma}$ is a Gaussian variance and is manually set by human prior knowledge.

quality and likelihood functions of the statistical models as the criterion for parameter selections. These statistical criteria however have no direct connection to the outliers pruned. Ideally λ should be a data-dependent parameter which selects a cut-off value and corresponds to the pruning rate p as the portion of the outliers among all comparisons.

This inspires us to sequentially consider all available solutions for all sparse variables along the Regularisation Path (RP) by gradually decreasing the value of the regularisation parameter λ from ∞ to 0. Specifically, based on the piecewise-linearity property of LASSO [9], RP can be efficiently computed by Least Angle Regression (LARS [8]). When $\lambda = \infty$, the regularisation parameter will strongly penalise outlier detection: if any annotation is taken as an outlier, it will greatly increase the value of the object function in Eq (9). When λ is changed from ∞ to 0, LASSO[2]will first select the variable subset accounting for the highest variances to the observations $U_2^T Y$ in Eq (9). These high variances should be assigned higher priority to represent the nonzero elements[3] of Γ of Eq (2), because Γ compensates the discrepancy between annotation and prediction. Based on this idea, we can order the edge set E by the λ values according to which nonzero γ_{ij} appears first when λ is decreased from ∞ to 0. In other words, if an edge γ_{ij} becomes nonzero at a larger λ_{ij} value, it has a higher probability to be an outlier. Following this order, we identify the top $p\%$ edge set Λ_p as the annotation outliers. And its complementary set $\Lambda_{1-p} = E \setminus \Lambda_p$ are the inliers. Therefore, the outcome of estimating Γ using Eq (9) is a binary outlier indication matrix $F_\Gamma = \left[F_{\gamma_{ij}}\right]$:

$$F_{\gamma_{ij}} = \begin{cases} 1 \; i \to j \in \Lambda_{1-p} \\ 0 \quad i \to j \in \Lambda_p \end{cases}$$

where each element $F_{\gamma_{ij}}$ indicates whether the corresponding edge $i \to j$ is an outlier or not. With this matrix, β can be solved by

$$\beta = (X^T X)^\dagger X^T (Y \odot F_\Gamma) \tag{10}$$

where $\odot$ is the Hardmard product and $F_\Gamma = \left[F_{\gamma_{ij}}\right]$. The pseudo-code of learning our URLR model is shown in Alg. 1. Note that it is very efficient to solve the entire regularisation path by LARS: “roughly the same computational cost as a single least square fit” (P438[33]).

3.5 Theretoial Advantage over Huber-LASSO

Our URLR framework is related to a widely used statistical ranking method – Huber-LASSO [46,14]. Huber-LASSO addresses estimating the robust ranking

[2] For a thorough discussion from a statistical perspective, please read [9,10,8,38].

[3] This is related with LASSO for covariate selection in a graph. Please read [32] for more details.

Algorithm 1. Learning a unified robust learning to rank model

```
Input: A training dataset $\Phi$ with pairwise annotation $Y$ and an outlier pruning rate $p\%$.
Output: Detection of outliers $F_\Gamma$ and prediction model parameter $\beta$.

1. Perform SVD on $X$ using Eq (8);
2. Solve Eq (9) using Regularisation Path;
3. Take the top $p\%$ pairs as outliers and estimate the outlier indicator matrix $F_\Gamma$;
4. Compute $\beta$ using Eq (10).
```

of the training data rather than learning to predict the ranking of test data; therefore only the annotation part of the training data Y is required, instead of both Y and Φ in URLR. Specifically, given the annotation Y of the training data, Huber-LASSO estimates the global ranking order θ by

$$\begin{aligned} \hat{\theta} = & \quad \min\nolimits_\theta \quad \frac{1}{2}\|Y - C\theta - \Gamma\|_2^2 + \lambda \parallel \Gamma \parallel_1 \\ := & \sum\nolimits_{(i,j)\in E} \left[\frac{1}{2}(Y_{ij} - \gamma_{ij} - \theta_i + \theta_j)^2 + \lambda|\gamma_{ij}|\right] \end{aligned} \tag{11}$$

where θ_i is the ranking score for instance i. Eq (11) is studied in [46,14] and is proved to be equivalent to the robust regression problem with Huber's loss function [18]. This is why it is called Huber-LASSO[4].

Our URLR model can be seen as an extension of Huber-LASSO for the ability to predict interestingness. It introduces the prediction model parameter β estimated as $\hat{\beta} = \hat{\theta}\Phi$. But this is not the most critical difference – one could still use Huber-LASSO to remove outliers and then use the same Eq (10) to estimate β. The more important difference is that URLR can better identify outliers, especially for sparse graphs. More specifically, to solve Eq (11), a similar formulation as Eq (9) can be used, solved by the same regularisation path method as in URLR. However, instead of SVD decomposing X in Eq (8), for Huber-LASSO, the matrix C is decomposed. This means the solution space of Eq (11) is $\dim(\Gamma) = |E| - I + 1$ where $|E|$ is the number of pairs compared and I is the number of graph nodes, i.e. training images or videos. Given a sparse dataset, this space is very small. In contrast, URLR enlarges $\dim(\Gamma)$ by including the subspace of original node space orthogonal to the feature space (Eq (9)). This means the solution space of Eq (9) is $\dim(\Gamma) \approx |E| - d$. When the feature dimension d is smaller than the number of images/videos I, the dimension of the solution space of Γ for URLR is higher than that of Huber-LASSO, leading to better outlier detection capability. Typically, we have $d < I$ in a large dataset; however if not, it can be made so by reducing the feature dimension.

[4] Note that broadly speaking, our method is still a special case of Huber-LASSO.

4 Experiments

Datasets. We conduct experiments on two image and video interestingness datasets and two relative image attribute datasets[5]. These datasets are summarised in Table 1. The image interestingness dataset was first introduced in [19] for studying memorability. It was later re-annotated as an image interestingness dataset by [16]. It consists of 2222 images, each represented as a 932 dimensional feature vector as in [16]. 16000 pairwise comparisons were collected by [16] using AMT and are used as annotation.

The video interestingness dataset is the YouTube interestingness dataset introduced in [21], which contains 14 different categories, each of which has 30 YouTube videos. 10 $\sim$ 15 annotators were asked to give complete interesting comparisons for all the videos in each category. So the original annotation is noisy but not sparse. We use bag-of-words of Scale Invariant Feature Transform (SIFT) and Mel-Frequency Cepstral Coefficient (MFCC) as the feature representation which are shown to be effective in [21] for predicting video interestingness.

We also carry out experiments on two relative attributes datasets –PubFig [25] and Scene [34] to test our URLR model's ability to predict other more general relative visual attributes. PubFig and Scene considered 11 ('smiling', 'round face', etc.) and 6 ('openness', 'natural' etc.) relative attributes respectively. Pairwise attribute annotation was collected by AMT [24]. Each pair was annotated by 5 crowdsourced workers. Gist and colour histograms features are used for PubFig, and Gist alone for Scene. Each image also belongs to a class (celebrity or scene type). These two datasets were designed for classification, with attribute scores as the representation, so the classification accuracy is determined by the attribute prediction accuracy.

Table 1. Dataset summary. We use the original features to learn the ranking model in Eq (10) and reduce the feature dimension (values in brackets) using KPCA to improve outlier detection in Eq (9) by enlarging the solution space (see Sec. 3.5).

Dataset	No. of pairs	No. img/video	Feature Dim.	No. class
Image Int.[19]	16000	2222	932(150)	1
Video Int. [21]	60000	420	1000(60)	14
PubFig [25,24]	2616	772	557(100)	8
Scene [34,24]	1378	2688	512(100)	8

Evaluation Metrics. For the image and video interestingness dataset, Kendall tau rank distance is employed to measure the rank correlation between the predicted ranking order and the ground truth ranking of unseen test data provided by [16] and [21] respectively[6]. Higher Kendall tau rank distance means lower

[5] All code and features are downloadable from Yanwei's website: `http://www.eecs.qmul.ac.uk/~yf300/ranking/`

[6] Recent statistical theories [37,20] show that the dense human annotations collected in [16] and [21] can give a reasonable approximation of ground truth for interestingness.

quality of the ranking order predicted. For the scene and pubfig image dataset, the relative attributes are very sparsely collected and their prediction performance can only be evaluated indirectly by image classification accuracy with the predicted relative attributes as image representation.

Competitors. We compare our method (URLR) with four competitors. (1) Jiang *et al.* [21]: this method uses majority voting for outlier pruning and rankSVM for learning to rank. (2) Gygli *et al.* [16]: this method also first removes outliers by majority voting. After that, the fraction of selections by the pairwise comparisons for each data point is used as an absolute interestingness score and a regression model is then learned for prediction. (3) *Huber-LASSO*: this is a statistical ranking method that performs outlier detection as described in Sec. 3.5, followed by estimating β using Eq (10). (4) *Raw*: This is our URLR model without outlier detection, that is, all annotations are used to estimate β.

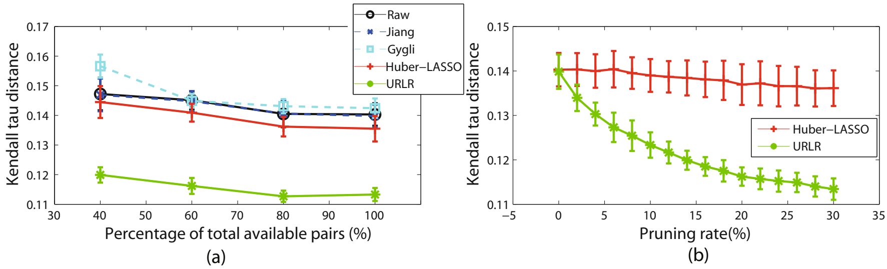


Fig. 2. Image interestingness prediction performance. Lower is better.

4.1 Image Interestingness Prediction

Experimental Settings. For this experiment, we randomly select 1000 images for training and the remaining 1222 are used for testing. All the experiments are repeated 10 times to reduce variance. The pruning rate p is set to 20%. We also vary the number of annotated pairs used to test how well each compared method copes with increasing annotation sparsity.

Comparative Results. The results are shown in Fig. 2 (a). It shows clearly that our URLR significantly outperforms the four alternatives for a wide range of annotation density. This validates the effectiveness of our method. In particular, the improvement over Jiang *et al.* [21] and Gygli *et al.* [16] demonstrates the superior outlier detection ability of URLR. URLR is superior to Huber-LASSO because the joint outlier detection and ranking estimation framework of URLR enables the enlargement of the solution space of Eq (9), resulting in better outlier detection performance. The performance of Gygli *et al.* [16] is the worst among all methods compared, particularly so given sparser annotation. This is not surprising – in order to get an reliable absolute interestingness value, dozens or even hundreds of comparisons per image are required, a condition not met by this dataset. The estimated value becomes less reliable given sparser annotations, explaining the worse relative performance. The performance of Huber-LASSO is also better than Jiang *et al.* [21] and Gygli *et al.* suggesting even a weaker global outlier detection approach is better then the majority voting based local one. Interestingly even the baseline method Raw gives a comparable result to Jiang *et al.* [21] and Gygli *et al.* [16] which suggests that just using all annotations without discrimination in a global cost function Eq (5) is as effective as majority voting.

Fig. 2 (b) evaluates how the performances of URLR and Huber-LASSO are affected by the pruning rate p. It can be seen that the performance of URLR is improving with an increasing pruning rate. This means that our URLR can keep on detecting true positive outliers. The gap between URLR and Huber-LASSO gets bigger when more comparisons are pruned showing Huber-LASSO stops detecting outliers much earlier on.

4.2 Video Interestingness Prediction

Experimental Settings. Because comparing videos across different categories is not very meaningful, we follow the same settings as in [21] and only compare the interestingness of videos within the same category. Specifically, we use 20 videos and their paired comparison for training and the remaining 10 videos for testing. The experiments are repeated for 10 rounds and the averaged results are reported. We use rankSVM with χ^2 kernel which is approximated by additive kernel of explicit feature mapping [42]. Kendall tau rank distance is used, and we find that the same results are obtained if the prediction accuracy used in [21] is used instead. The pruning rate is again set to 20%.

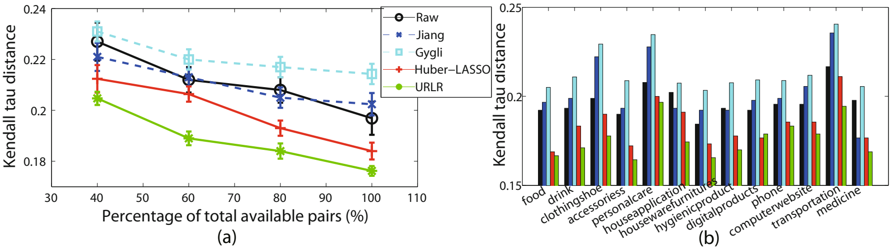


Fig. 3. Video interestingness prediction results

Comparative Results. The results of video interestingness prediction are shown in Fig 3. Fig. 3(a) compares different methods given varying amounts of annotations, and Fig. 3(b) shows the per category performance. The results show that all the observations we had for the image interestingness prediction experiment still hold here, and across all categories. However in general the gaps between our URLR and the alternatives are smaller as this dataset is densely annotated. In particular the performance of Huber-LASSO is much closer to our URLR now. This is because, as explained in Sec. 3.5, the advantage of URLR over Huber-LASSO is stronger when $|E|$ is close to I. Given a dense pairwise annotation $|E|$ is much greater than I and the effect of enlarging the solution space diminishes.

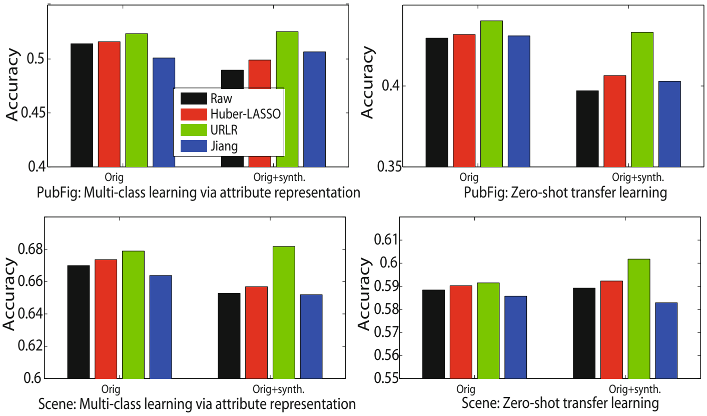
4.3 Relative Attributes Prediction for Image Classification

Experimental Settings. We evaluate image classification with relative attributes as representation on the PubFig and Scene datasets under two settings: multi-class classification where samples from all classes are available for training and zero-shot transfer learning where one class is held out during training (a different class is used in each trial with the result averaged). Our experiment setting is similar to that in [35], except that image-level, rather than class-level pairwise comparisons are used. Two variations of the setting are used:

- *Orig:* The original setting with the pairwise annotations is used as they are.
- *Orig+synth*: By visual inspection, there are limited annotation outliers in these datasets, perhaps because the relative attributes are less subjective compared to interestingness. To simulate more challenging situations, we randomly add 150 random comparison for each attribute, many of which would correspond to outliers. This will lead to around 20% extra outliers.

The pruning rate is set to 7% for original dataset (*Orig*) and 27% for dataset with additional outliers inserted for all attributes of both datasets (*Orig+synth*).

Comparative Results. Without the ground truth of relative attribute values, different models are evaluated indirectly via image classification accuracy in Fig. 4. Note that the method of Gygli *et al.* [16] is not compared here as the



PubFig: Multi-class learning via attribute representation

PubFig: Zero-shot transfer learning

Scene: Multi-class learning via attribute representation

Scene: Zero-shot transfer learning

Fig. 4. Relative attribute performance evaluated indirectly as image classification rate (chance = 0.125)

annotation is too sparse for it to learn a meaningful model. The following observations can be made: (1) Our URLR always outperforms *Huber-LASSO*, *maj-voting* (Jiang) and *Raw* for all experiment settings. The improvement is more significant when the data contain more errors (*Orig+synth*). (2) The performance of other methods is in general consistent to what we observed in the image and video interestingness experiments: Huber-LASSO is better than majority voting (Jiang *et al.* [21]) and Raw often gives better results than majority voting too. (3) It is noted that for PubFig, Jiang *et al.* [21] is better than Raw given more outliers, but it is not the case for Scene. This is probably because the annotators are more familiar with the celebrity faces in PubFig hence their attributes than those in Scene. Consequently there should be more subjective/intentional errors for Scene, causing majority voting to choose wrong local ranking orders (e.g. not many people are sure how to compare the relative values of the 'diagonal plane' attribute for two images). These majority voting + outlier cases can only be rectified by using a global approach such as our URLR, even the Huber-LASSO method to a certain extent.

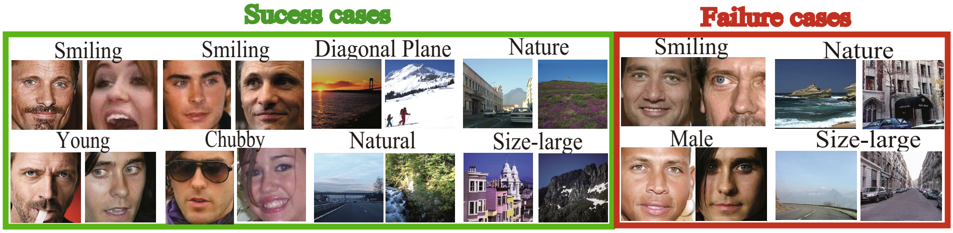


Fig. 5. Qualitative results on image relative attribute prediction

Qualitative Results. Figure 5 gives some examples of the pruned pairs for both datasets using URLR. In the success cases, the left images were (incorrectly) annotated to have more of the attribute than the right ones. However, they are either wrong or too ambiguous to give consistent answers, and as such are detrimental to learning to rank. A number of failure cases (false positive pairs identified by URLR) are also shown. Some of them are caused by unique view point (e.g. Hugh Laurie's mouth is not visible, so it is hard to tell who smiles more; the building and the street scene are too zoomed in compared to most other samples); others are caused by the weak feature representation, e.g. in the 'male' attribute example, the colour and Gist features are not discriminative enough for judging which of the two men has more 'male' attribute.

5 Conclusions

We have proposed a novel unified robust learning to rank (URLR) framework for predicting image and video interestingness. The key advantage of our method over the existing majority voting based approaches is that we can detect outliers globally by minimising a global ranking inconsistency cost. The joint outlier detection and ranking estimation formulation also provides our model with an advantage over the conventional statistical ranking methods such as Huber-LASSO for outlier detection. The effectiveness of our model in comparison with state-of-the-art alternatives has been validated using image and video interestingness datasets. Further, it is generally applicable to other relative attribute prediction tasks as demonstrated by our relative attribute based image classification experiments.

References

1. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: From pairwise approach to listwise approach. In: ICML (2007)
2. Carvalho, V.R., Elsas, J.L., Cohen, W.W., Carbonell, J.G.: A meta-learning approach for robust rank learning. In: SIGIR 2008 LR4IR - Workshop on Learning to Rank for Information Retrieval (2008)
3. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with svms. Inf. Retr. (2010)
4. Chen, A.: adn R. Pangrazi, P.D.: An examination of situational interest adn its sources. Brit. J. of Edu. Psychology (20001)
5. Chen, K., Wu, C., Chang, Y., Lei, C.: Crowdsourceable QoE evalutation framework for multimedia content. In: ACM MM (2009)
6. Chen, X., Bennett, P.N.: Pairwise ranking aggregation in a crowdsourced setting. In: ACM International Conference on Web Search and Data Mining (2013)
7. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: CVPR (2011)
8. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics (2004)
9. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. JASA (2001)
10. Fan, J., Tang, R., Shi, X.: Partial consistency with sparse incidental parameters. arXiv:1210.6950 (2012)
11. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Attribute learning for understanding unstructured social activity. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 530–543. Springer, Heidelberg (2012)
12. Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z., Gong, S.: Transductive multi-view embedding for zero-shot recognition and annotation. In: ECCV (2014)
13. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Learning multi-modal latent attributes. TPAMI (2013)
14. Gannaz, I.: Robust estimation and wavelet thresholding in partial linear models. Stat. Comput. 17, 293–310 (2007)
15. Gehrlein, W.V.: Condorcet's paradox. Theory and Decision (1983)
16. Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., Gool, L.V.: The interestingness of images. In: ICCV (2013)
17. Hirani, A.N., Kalyanaraman, K., Watts, S.: Least squares ranking on graphs. arXiv:1011.1716 (2010)
18. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
19. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable? In: CVPR (2011)
20. Jiang, X., Lim, L.H., Yao, Y., Ye, Y.: Statistical ranking and combinatorial hodge theory. Math. Program. (2011)

21. Jiang, Y.G., YanranWang, F.R., Xue, X., Zheng, Y., Yang, H.: Understanding and predicting interestingness of videos. In: AAAI (2013)
22. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR (2006)
23. Kittur, A., Chi, E.H., Suh., B.: Crowdsourcing user studies with mechanical turk. In: ACM CHI (2008)
24. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: Image search with relative attribute feedback. In: CVPR (2012)
25. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
26. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
27. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE TPAMI (2013)
28. Liu, F., Niu, Y., Gleicher, M.: Using web photos for measuring video frame interestingness. In: IJCAI (2009)
29. Liu, Y., Gao, B., Liu, T.Y., Zhang, Y., Ma, Z., He, S., Li, H.: Browserank: letting web users vote for page importance. In: ACM SIGIR (2008)
30. Long, C., Hua, G., Kapoor, A.: Active visual recognition with expertise estimation in crowdsourcing. In: ICCV (2013)
31. Maire, M., Yu, S.X., Perona, P.: Object detection and segmentation from joint embedding of parts and pixels. In: ICCV (2011)
32. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. Ann. Statist. (2006)
33. Murphy, K.P.: Machine learning: a probabilistic perspective. The MIT Press (2012)
34. Oliva, A., Torralba., A.: Modeling the shape of the scene: Aholistic representation of the spatial envelope. IJCV 42 (2001)
35. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (2011)
36. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proc. CVPR (2012)
37. Rajkumar, A., Agarwal, S.: A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In: Proceedings of the 31st International Conference on Machine Learning (2014)
38. She, Y., Owen, A.B.: Outlier detection using nonconvex penalized regression. Journal of American Statistical Association (2011)
39. Silvia, P.: Interest - the curious emotion. In: CDPS (2008)
40. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: CVPR Workshops (2008)
41. Sun, Z., Qin, T., Tao, Q., Wang, J.: Robust sparse rank learning for non-smooth ranking measures. In: ACM SIGIR (2009)
42. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. In: IEEE TPAMI (2011)
43. Wu, O., Hu, W., Gao, J.: Learning to rank under multiple annotators. In: IJCAI (2011)
44. Xu, Q., Huang, Q., Jiang, T., Yan, B., Lin, W., Yao, Y.: Hodgerank on random graphs for subjective video quality assessment. IEEE TMM (2012)
45. Xu, Q., Huang, Q., Yao, Y.: Online crowdsourcing subjective image quality assessment. In: ACM MM (2012)
46. Xu, Q., Xiong, J., Huang, Q., Yao, Y.: Robust evaluation for quality of experience in crowdsourcing. In: ACM MM (2013)