This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels

Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Jiechao Xiong,
Shaogang Gong, Yizhou Wang, and Yuan Yao

Abstract—The problem of estimating subjective visual properties from image and video has attracted increasing interest. A subjective visual property is useful either on its own (e.g. image and video interestingness) or as an intermediate representation for visual recognition (e.g. a relative attribute). Due to its ambiguous nature, annotating the value of a subjective visual property for learning a prediction model is challenging. To make the annotation more reliable, recent studies employ crowdsourcing tools to collect pairwise comparison labels. However, using crowdsourced data also introduces outliers. Existing methods rely on majority voting to prune the annotation outliers/errors. They thus require a large amount of pairwise labels to be collected. More importantly as a local outlier detection method, majority voting is ineffective in identifying outliers that can cause global ranking inconsistencies. In this paper, we propose a more principled way to identify annotation outliers by formulating the subjective visual property prediction task as a unified robust learning to rank problem, tackling both the outlier detection and learning to rank jointly. This differs from existing methods in that (1) the proposed method integrates local pairwise comparison labels together to minimise a cost that corresponds to global inconsistency of ranking order, and (2) the outlier detection and learning to rank problems are solved jointly. This not only leads to better detection of annotation outliers but also enables learning with extremely sparse annotations.

Index Terms—Subjective visual properties, outlier detection, robust ranking, robust learning to rank, regularisation path

1 Introduction

The solutions to many computer vision problems involve the estimation of some visual properties of an image or video, represented as either discrete or continuous variables. For example scene classification aims to estimate the value of a discrete variable indicating which scene category an image belongs to; for object detection the task is to estimate a binary variable corresponding the presence/absence of the object of interest and a set of variables indicating its whereabouts in the image plane (e.g. four variables if the whereabouts are represented as bounding boxes).Most of these visual properties are objective; that is, there is no or little ambiguity in their true values to a human annotator.

In comparison, the problem of estimating subjective visual properties is much less studied. This class of computer vision problems nevertheless encompass a variety of important applications. For example: estimating attractiveness [1] from faces would interest social media or online dating websites; and estimating properties of consumer goods such as shininess of shoes [2] improves customer experiences on online shopping websites. Recently, the problem of automatically predicting if people would find an image or video interesting has started to receive increasing attention [3], [4], [5]. 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 recommendation tools to select which ones to view. Given a query, ranking the retrieved data with relevance to the query based on the predicted interestingness would improve 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. Subjective visual properties such as the above-mentioned ones are useful on their own. But they can also be used as an intermediate representation for other tasks such as visual recognition, e.g., different people can be recognised by how pale their skin complexions are and how chubby their faces are [6]. When used as a semantically meaningful representation, these subjective visual properties often are referred to as relative attributes [2], [6], [7].

Learning a model for subjective visual property (SVP) prediction is challenging primarily due to the difficulties in obtaining annotated training data. Specifically, since most SVPs can be represented as continuous variables (e.g. an interestingness/aesthetics/shininess score with a value range of 0 to 1 with 1 being most interesting/aesthetically appealing/shinning), SVP prediction can be cast as a regression problem – the low-level feature values are regressed to the SVP values given

- *Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, UK. Email: {y.fu,t.hospedales,t.xiang,s.gong}@qmul.ac.uk.*
- *Jiechao Xiong, and Yuan Yao are with the School of Mathematical Sciences, Peking University, China. Email:yuany@math.pku.edu.cn*
- *Tao Xiang and Yuan Yao are the corresponding authours.*
- *Yizhou Wang is with the School of Electronics Engineering and Computer Science, Peking University, China. Email: yizhou.wang@pku.edu.cn*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

a set of training data annotated with their true SVP values. However, since by definition these properties are subjective, different human annotators often struggle to give an absolute value and as a result the annotations of different people on the same instance can vary hugely. For example, on a scale of 1 to 10, different people will have very different ideas on what a scale 5 means for an image, especially without any common reference point. On the other hand, it is noted that humans can in general more accurately rank a pair of data points in terms of their visual properties [8], [9] , e.g. it is easier to judge which of two images is more interesting relatively than giving an absolute interestingness score to each of them. Most existing studies [2], [1], [9] on SVP prediction thus take a learning to rank approach [10], where annotators give comparative labels about pairs of images/videos and the learned model is a ranking function that predicts the SVP value as a ranking score.

To annotate these pairwise comparisons, crowdsourcing tools such as Amazon Mechanic Turk (AMT) are resorted to, which allow a large number of annotators to collaborate at very low cost. Data annotation based on crowdsourcing is increasingly popular [6], [2], [4], [5] recently for annotating large-scale datasets. However, this brings about two new problems: (1) **Outliers** – The crowd is not all trustworthy: it is well known that crowdsourced data are greatly affected by noise and outliers [11], [12], [13] which can be caused by a number of factors. Some workers may be lazy or malicious [14], providing random or wrong annotations either carelessly or intentionally; some other outliers are unintentional human errors caused by the ambiguous nature of the data, thus are unavoidable regardless how good the attitudes of the workers are. For example, the pairwise ranking for Figure 1(a) depends on the cultural/psychological background of the annotator – whether s/he is more familiar/prefers the story of Monkey King or Cookie Monster[1]. When we learn the model from labels collected from many people, we essentially aim to learn the consensus, i.e. what most people would agree on. Therefore, if most of the annotators growing up watching Sesame Street thus consciously or subconsciously consider the Cookie Monster to be more interesting than the Monkey King, their pairwise labels/votes would represent the consensus. In contrast, one annotator who is familiar with the stories in Journey to the West may choose the opposite; his/her label is thus an outlier under the consensus. (2) **Sparsity** – the number of pairwise comparisons required is much bigger than the number of data points because n instances define a $\mathcal{O}(n^2)$ pairwise space. Consequently, even with crowdsourcing tools, the annotation remains be sparse, i.e. not all pairs are compared and each pair is only compared a few times.

To deal with the outlier problem in crowdsourced data, existing studies take a majority voting strategy [6], [2], [4], [15], [16], [17], [18]. That is, a large budget of $5-10$ times the number of actual annotated pairs required is allocated to obtain multiple annotations for each pair. These annotations are then averaged over so as to eliminate label noise. However, the effectiveness of the majority voting strategy is often limited by the sparsity problem – it is typically infeasible to have many annotators for each pair. Furthermore, there is no guarantee that outliers, particularly those caused by unintentional human errors can be dealt with effectively. This is because majority voting is a local consistency detection based strategy – when there are contradictory/inconsistent pairwise rankings for a given pair, the pairwise rankings receiving minority votes are eliminated as outliers. However, it has been found that when pairwise local rankings are integrated into a global ranking, it is possible to detect outliers that can cause global inconsistency and yet are locally consistent, i.e. supported by majority votes [19]. Critically, outliers that cause global inconsistency have more significant detrimental effects on learning a ranking function for SVP prediction and thus should be the main focus of an outlier detection method.

Figure 1. Examples of pairwise comparisons of subjective visual properties.

In this paper we propose a novel approach to subjective visual property prediction from sparse and noisy pairwise comparison labels collected using crowdsourcing tools. Different from existing approaches which first remove outliers by majority voting, followed by regression [4] or learning to rank [5], we formulate a unified robust learning to rank (URLR) framework to solve jointly both the outlier detection and learning to rank 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 those outliers that receive majority votes but cause large global ranking inconsistency and thus should be removed. Furthermore, as a global method that aggregates comparisons across different pairs, our method can operate with as few as one comparison per pair, making our method much more robust against the data sparsity problem compared to the conventional majority voting approach that aggregates comparisons for each pair in isolation. More specifically, the proposed model generalises a partially penalised LASSO optimisation or Huber-LASSO formulation [20],

1. This is also known as Halo Effect in Psychology.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

[21], [22] from a robust statistical ranking formulation to a robust learning to rank model, making it suitable for SVP prediction given unseen images/videos. We also formulate a regularisation path based solution to solve this new formulation efficiently. Extensive experiments are carried out on benchmark datasets including two image and video interestingness datasets [4], [5] and two relative attribute datasets [2]. The results demonstrate that our method significantly outperforms the state-of-the-art alternatives.

2 Related work

Subjective visual properties Subjective visual property prediction covers a large variety of computer vision problems; it is thus beyond the scope of this paper to present an exhaustive review here. Instead we focus mainly on the image/video interestingness prediction problem which share many characteristics with other SVP prediction problem such as image quality [23], memorability [24], and aesthetics [3] prediction.

Predicting image and video interestingness Early efforts on image interestingness prediction focus on different aspects than interestingness as such, including memorability [24] and aesthetics [3]. These SVPs 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 [4]. The work of Gygli et al [4] 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. Liu et al. [25] focus on key frames so essentially treats it as an image interestingness problem, whilst [5] is the first work that proposes benchmark video interestingness datasets and evaluates different features for video interestingness prediction.

Most earlier works cast the aesthetics or interestingness prediction problem as a regression problem [23], [3], [24], [25]. 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 [4] and video [5] interestingness all collect pairwise comparison data by crowdsourcing. Both use majority voting to remove outliers first. After that the prediction models differ – [4] converts pairwise comparisons into an absolute interestingness values and use a regression model, whilst [5] employs rankSVM [10] 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 outliers more effectively – even if they correspond to comparisons receiving majority votes, thanks to its global formulation.

Relative attributes In a broader sense interestingness can be considered as one type of relative attribute [6]. Attribute-based modelling [26], [27] has gained popularity recently as a way to *describe* instances and classes at an intermediate level of representation. Attributes are then used for various tasks including N-shot and zero-shot transfer learning. Most previous studies consider binary attributes [26], [27]. Relative attributes [6] were recently proposed to learn a ranking function to predict relative semantic strength of visual attributes. Instead of the original class-level attribute comparisons in [6], this paper focuses on instance-level comparisons due to the huge intra-class variations in real-world problems. With instance-level pairwise comparisons, relative attributes have been used for interactive image search [2], and semi-supervised [28] or active learning [29], [30] of visual categories. However, no previous work addresses the problem of annotation outliers except [2], which adopts the heuristic majority voting strategy.

Learning from noisy paired crowdsourced data 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 [14], [31], [32], [13] highlight the necessity of validating the random or malicious labels/workers and give filtering heuristics for data cleaning. However, these are primarily based on majority voting which requires a costly volume of redundant annotations, and has no theoretical guarantee of solving the outlier and sparsity problems. As a local (per-pair) filtering method, majority voting does not respect global ordering and even risks introducing additional inconsistency due to the well-known Condorcet's paradox in social choice and voting theory [33]. Active learning [34], [29], [30] is an another way to circumvent the $\mathcal{O}(n^2)$ pairwise labelling space. It actively poses specific requests to annotators and learns from their feedback, rather than the 'general' pairwise comparisons discussed in this work. Besides paired crowdsourced data, majority voting is more widely used in crowdsourcing where multiple annotators directly label instances, which attracted lots of attention in the machine learning community [16], [17], [18], [15]. In contrast, our work focuses on pairwise comparisons which are relatively easier for annotators in evaluating the subjective visual properties [8] .

Statistical ranking and learning to rank Statistical ranking has been widely studied in statistics and computer science [35], [36], [8], [37]. 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 ranking functions for applications such as interestingness prediction, a feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

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 [38], [39], [40]. However, existing learning to rank works do 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 prediction 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 subjective visual property prediction using noisy and sparse pairwise comparison/ranking labels as training data. (2) For the first time, the problems of detecting outliers and estimating linear ranking models are solved jointly in a 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. An earlier and preliminary version of this work is presented in [41] which focused only on the image/video interestingness prediction problem.

3 Unified Robust Learning to Rank

3.1 Problem definition

We aim to learn a subjective visual property (SVP) prediction model from a set of sparse and noisy pairwise comparison labels, each comparison corresponding to a local ranking between a pair of images or videos. Suppose our training set has N data points/instances represented by a feature matrix $\Phi = \left[\boldsymbol{\phi}_i^T\right]_{i=1}^N \in \mathbb{R}^{N\times d}$, where $\boldsymbol{\phi}_i$ is a d-dimensional column low-level feature vector representing instance i. The pairwise comparison labels (annotations collected using crowdsourcing tools) can be naturally represented as a directed comparison graph $G = (V, E)$ with a node set $V = \{i\}_{i=1}^N$ corresponding to the N instances and an edge set $E = \{e_{ij}\}$ corresponding to the pairwise comparisons.

The pairwise comparison labels can be provided by multiple annotators. They are dichotomously saved: Suppose annotator α gives a pairwise comparison for instance i and j ($i, j \in V$). If α considers that the SVP of instance i is stronger/more than that of j, we save $(i, j, y^\alpha_{e_{ij}})$ and set $y^\alpha_{e_{ij}} = 1$. If the opposite is the case, we save $(j, i, y^\alpha_{e_{ji}})$ and set $y^\alpha_{e_{ji}} = 1$. All the pairwise comparisons between instances i and j are then aggregated over all annotators who have cast a vote on this pair; the results are represented as $w_{e_{ij}} = \sum_\alpha [\![y^\alpha_{e_{ij}} = 1]\!]$ which is the total number of votes on i over j for a specific SVP, where $[\![]\!]$ indicates the Iverson's bracket notation, and $w_{e_{ji}}$ which is defined similarly. This gives an edge weight vector $\boldsymbol{w} = \left[w_{e_{ij}}\right] \in \mathbb{R}^{|E|}$ where $|E|$ is the number of edges. Now the edge set can be represented as $E = \left\{e_{ij} | w_{e_{ij}} > 0\right\}$ and $w_{e_{ij}} \in \mathbb{R}$ is the weight for the edge e_{ij}. In other words, an edge e_{ij}: $i \to j$ exists if $w_{e_{ij}} > 0$. The topology of the graph is denoted by a flag indicator vector $\boldsymbol{y} = \left[y_{e_{ij}}\right] \in \mathbb{R}^{|E|}$ where each indicator $y_{e_{ij}} = 1$ indicates that there is an edge between instances i to j regardless how many votes it carries. Note that all the elements in $\boldsymbol{y}$ have the value 1, and their index e_{ij} gives the corresponding nodes in the graph.

Given the training data consisting of the feature matrix Φ and the annotation graph G, there are two tasks:

1) Detecting and removing the outliers in the edge set E of G. To this end, we introduce a set of unknown variables $\boldsymbol{\gamma} = \left[\gamma_{e_{ij}}\right] \in \mathbb{R}^{|E|}$ where each variable $\gamma_{e_{ij}}$ indicates whether the edge e_{ij} is an outlier. The outlier detection problem thus becomes the problem of estimating $\boldsymbol{\gamma}$.
2) Estimating a prediction function for SVP. In this work a linear model is considered due to its low computational complexity, that is, given the low-level feature $\boldsymbol{\phi}_x$ of a test instance x we use a linear function $f(x) = \boldsymbol{\beta}^T\boldsymbol{\phi}_x$ to predict its SVP, where $\boldsymbol{\beta}$ is the coefficient weight vector of the low-level feature $\boldsymbol{\phi}_x$. Note that all formulations can be easily updated to use a non-linear function.

So far in the introduced notations three vectors share indices: the flag indicator vector $\boldsymbol{y}$, the outlier variable vector $\boldsymbol{\gamma}$ and the edge weight vector $\boldsymbol{w}$. For notation convenience, from now on we use y_{ij}, γ_{ij} and w_{ij} to replace $y_{e_{ij}}$, $\gamma_{e_{ij}}$ and $w_{e_{ij}}$ respectively. As in most graph based model formulations, we define $C \in \mathbb{R}^{|E|\times N}$ as the incident matrix of the directed graph G, where $C_{e_{ij}i} = -1/1$ if the edge e_{ij} enters/leaves vertex i.

Note that in an ideal case, one hopes that the votes received on each pair are unanimous, e.g. $w_{ij} > 0$ and $w_{ji} = 0$; but often there are disagreements, i.e. we have both $w_{ij} > 0$ and $w_{ji} > 0$. 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

In contrast to majority voting, we propose to prune outliers globally and jointly with learning the SVP prediction function. To this end, the outlier variables γ_{ij} for outlier detection and the coefficient weight vector $\boldsymbol{\beta}$ for SVP prediction are estimated in a unified framework. Specifically, for each edge $e_{ij} \in E$, its corresponding flag indicator y_{ij} is modelled as

$$y_{ij} = \boldsymbol{\beta}^T\boldsymbol{\phi}_i - \boldsymbol{\beta}^T\boldsymbol{\phi}_j + \gamma_{ij} + \varepsilon_{ij}, \tag{1}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise with zero mean and a variance σ, and the outlier variable $\gamma_{ij} \in \mathbb{R}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

is assumed to have a higher magnitude than σ. For an edge e_{ij}, if y_{ij} is not an outlier, we expect $\boldsymbol{\beta}^T\boldsymbol{\phi}_i - \boldsymbol{\beta}^T\boldsymbol{\phi}_j$ should be approximately equal to y_{ij}, therefore we have $\gamma_{ij} = 0$. On the contrary, when the prediction of $\boldsymbol{\beta}^T\boldsymbol{\phi}_i - \boldsymbol{\beta}^T\boldsymbol{\phi}_j$ 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) can be re-written in its matrix form

$$\boldsymbol{y} = C\Phi\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{y} = [y_{ij}] \in \mathbb{R}^{|E|}$, $\boldsymbol{\gamma} = [\gamma_{ij}] \in \mathbb{R}^{|E|}$, $\boldsymbol{\epsilon} = [\varepsilon_{ij}] \in \mathbb{R}^{|E|}$ and $C \in \mathbb{R}^{|E|\times N}$ is the incident matrix of the annotation graph G.

In order to estimate the $|E| + d$ unknown parameters ($|E|$ for $\boldsymbol{\gamma}$ and d for $\boldsymbol{\beta}$), we aim to minimise the discrepancy between the annotation $\boldsymbol{y}$ and our prediction $C\Phi\boldsymbol{\beta} + \boldsymbol{\gamma}$, as well as keeping the outlier estimation $\boldsymbol{\gamma}$ sparse. Note that $\boldsymbol{y}$ only contains information about which pairs of instances have received votes, but not how many. The discrepancy thus needs to weighted by the number of votes received, represented by the edge weight vector $\boldsymbol{w} = [w_{ij}] \in \mathbb{R}^{|E|}$. To that end, we put a weighted l_2-loss on the discrepancy and a sparsity enhancing penalty on the outlier variables. This gives us the following cost function:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\|\boldsymbol{y} - C\Phi\boldsymbol{\beta} - \boldsymbol{\gamma}\|^2_{2,\boldsymbol{w}} + p_\lambda(\boldsymbol{\gamma}) \tag{3}$$

where

$$\|\boldsymbol{y} - C\Phi\boldsymbol{\beta} - \boldsymbol{\gamma}\|^2_{2,\boldsymbol{w}} = \sum_{e_{ij}\in E} w_{ij}(y_{ij} - \gamma_{ij} - \boldsymbol{\beta}^T\boldsymbol{\phi}_i + \boldsymbol{\beta}^T\boldsymbol{\phi}_j)^2,$$

and $p_\lambda(\boldsymbol{\gamma})$ is the sparsity constraint on $\boldsymbol{\gamma}$. With this cost function, our Unified Robust Learning to Rank (URLR) framework identifies outliers globally by integrating all local pairwise comparisons together. Note that in Eq (3), the noise term $\boldsymbol{\epsilon}$ has been removed because the discrepancy is mainly caused by outliers due to their larger magnitude.

Ideally the sparsity enhancing penalty term $p_\lambda(\boldsymbol{\gamma})$ should be a l_0 regularisation term. However, for a tractable solution, a l_1 regularisation term is used: $p_\lambda(\boldsymbol{\gamma}) = \lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} = \lambda\sum_{e_{ij}} w_{ij}|\gamma_{ij}|$, where λ is a free parameter corresponding to the weight for the regularisation term. With this l_1 penalty term, the cost function becomes convex:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\|\sqrt{W}(\boldsymbol{y} - \boldsymbol{\gamma}) - X\boldsymbol{\beta}\|^2_2 + \lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}}, \tag{4}$$

where $X = \sqrt{W}C\Phi$, $W = \mathrm{diag}(\boldsymbol{w})$ is the diagonal matrix of $\boldsymbol{w}$ and $\sqrt{W} = \mathrm{diag}(\sqrt{\boldsymbol{w}})$.

Setting $\frac{\partial L}{\partial \boldsymbol{\beta}} = 0$, the problem of minimisation of the cost function in (4) can be decomposed into the following two subproblems:

1) Estimating the parameters $\boldsymbol{\beta}$ of the prediction function $f(x)$:

$$\hat{\boldsymbol{\beta}} = (X^TX)^\dagger X^T\sqrt{W}(\boldsymbol{y} - \boldsymbol{\gamma}), \tag{5}$$

Mathematically, the Moore-Penrose pseudo-inverse of X^TX is defined as $(X^TX)^\dagger = \lim_{\mu\to 0}((X^TX)^T(X^TX) + \mu I)^{-1}(X^TX)^T$, where I is the identity matrix. The scalar variable μ is introduced to avoid numerical instability [42], and typically assumes a small value[2]. With the the introduction of μ, Eq (5) becomes:

$$\hat{\boldsymbol{\beta}} = (X^TX + \mu I)^{-1}X^T\sqrt{W}(\boldsymbol{y} - \boldsymbol{\gamma}). \tag{6}$$

A standard solver for Eq (6) has a $O(|E|d^2)$ computational complexity, which is almost linear with respect to the size of the graph $|E|$ if $d \ll n$. Faster algorithms based on the Krylov iterative and algebraic multi-grid methods [43] can also be used.

2) Outlier detection:

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \tfrac{1}{2}\|(I - H)\sqrt{W}(\boldsymbol{y} - \boldsymbol{\gamma})\|^2_2 + \lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} \tag{7}$$

$$= \arg\min_{\boldsymbol{\gamma}} \tfrac{1}{2}\|\tilde{\boldsymbol{y}} - \widetilde{X}\boldsymbol{\gamma})\|^2_2 + \lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}} \tag{8}$$

where $H = X(X^TX)^\dagger X^T$ is the hat matrix, $\widetilde{X} = (I - H)\sqrt{W}$ and $\tilde{\boldsymbol{y}} = \widetilde{X}\boldsymbol{y}$. Eq (7) is obtained by plugging the solution $\hat{\boldsymbol{\beta}}$ back into Eq (4).

3.3 Outlier detection by regularisation path

From the formulations described above, it is clear that outlier detection by solving Eq (8) is the key – once the outliers are identified, the estimated $\hat{\boldsymbol{\gamma}}$ can be used to substitute $\boldsymbol{\gamma}$ in Eq (5) and the estimation of the prediction function parameter $\boldsymbol{\beta}$ becomes straightforward. Now let us focus on solving Eq (8) for outlier detection.

Note that solving Eq (8) is essentially a LASSO (Least Absolute Shrinkage and Selection Operator) [20] problem. For a LASSO problem, tuning the regularisation parameter λ is notoriously difficult [44], [45], [46], [47]. In particular, in our URLR framework, the λ value directly decides the ratio of outliers in the training set which is unknown. A number of methods for determining λ exist, but none is suitable for our formulation:

1) Some heuristics rules on setting the value of λ such as $\lambda = 2.5\hat{\sigma}$ are popular in existing robust ranking models such as the M-estimator [44], where $\hat{\sigma}$ is a Gaussian variance set manually based on human prior knowledge. However setting a constant λ value independent of dataset is far from optimal because the ratio of outliers may vary for different crowdsourced datasets.
2) Cross validation is also not applicable here because each edge e_{ij} is associated with a γ_{ij} variable and any held-out edge e_{ij} also has an associated unknown variable γ_{ij}. As a result, cross validation can only optimise part of the sparse variables while

2. In this work, μ is set to 0.001.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

leaving those for the held-out validation set undetermined.

3) Data adaptive techniques such as Scaled LASSO [45] and Square-Root LASSO [46] typically generate over-estimates on the support set of outliers. Moreover, they rely on the homogeneous Gaussian noise assumption which is often not valid in practice.
4) The other alternatives e.g. Akaike information criterion (AIC) and Bayesian information criterion (BIC) are often unstable in outlier detection LASSO problems [47][3].

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, a regularisation path can be efficiently computed by the R-package "glmnet" [48][4]. 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 cost function in Eq (8). When λ is changed from ∞ to 0, LASSO[5] will first select the variable subset accounting for the highest deviations to the observations $\widetilde{X}$ in Eq (8). These high deviations should be assigned higher priority to represent the nonzero elements[6] of $\boldsymbol{\gamma}$ of Eq (2), because $\boldsymbol{\gamma}$ compensates the discrepancy between annotation and prediction. Based on this idea, we can order the edge set E according to which nonzero γ_{ij} appears first when λ is decreased from ∞ to 0. In other words, if an edge e_{ij} whose associated outlier variable $\boldsymbol{\gamma}_{ij}$ becomes nonzero at a larger λ 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 $\boldsymbol{\gamma}$ using Eq (8) is a binary outlier indicator vector $\boldsymbol{f} = \left[f_{e_{ij}}\right]$:

$$f_{e_{ij}} = \begin{cases} 1 & e_{ij} \in \Lambda_{1-p} \\ 0 & e_{ij} \in \Lambda_p \end{cases}$$

where each element $f_{e_{ij}}$ indicates whether the corresponding edge e_{ij} is an outlier or not.

Now with the outlier indicator vector $\boldsymbol{f}$ estimated using regularisation path, instead of estimating $\boldsymbol{\beta}$ by substituting $\boldsymbol{\gamma}$ in Eq (5) with an estimated $\hat{\boldsymbol{\gamma}}$, $\boldsymbol{\beta}$ can be computed as

$$\hat{\boldsymbol{\beta}} = (X^T F X + \mu I)^{-1} X^T \sqrt{W} F \boldsymbol{y} \tag{9}$$

3. We found empirically that the model automatically selected by BIC or AIC failed to detect any meaningful outliers in our experiments. For details of the experiments and a discussion on the issue of determining the outlier ratio, please visit the project webpage at http://www.eecs.qmul.ac.uk/~yf300/ranking/index.html

4. http://cran.r-project.org/web/packages/glmnet/glmnet.pdf

5. For a thorough discussion from a statistical perspective, please see [49], [50], [51], [47].

6. This is related with LASSO for covariate selection in a graph. Please see [52] for more details.

Algorithm 1 Learning a unified robust learning to rank (URLR) model for SVP prediction

Input: A training dataset consisting of the feature matrix Φ and the pairwise annotation graph G, and an outlier pruning rate $p\%$.
Output: Detected outliers $\boldsymbol{f}$ and prediction model parameter $\boldsymbol{\beta}$.

1) Solve Eq (8) using Regularisation Path;
2) Take the top $p\%$ pairs as outliers to obtain the outlier indicator vector $\boldsymbol{f}$;
3) Compute $\boldsymbol{\beta}$ using Eq (9).

where $F = \mathrm{diag}(\boldsymbol{f})$, that is, we use $\boldsymbol{f}$ to 'clean up' $\boldsymbol{y}$ before estimating $\boldsymbol{\beta}$.

The pseudo-code of learning our URLR model is summarised in Algorithm 1.

3.4 Discussions

3.4.1 Advantage over majority voting

The proposed URLR framework identifies outliers globally by integrating all local pairwise comparisons together, in contrast to the local aggregation based majority voting. Figure 2(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 of them compared three times each, and the correct global ranking order of these 5 images in terms of a specific SVP is $A < B < C < D < E$. Figure 2(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 $\boldsymbol{\beta}$ makes $\boldsymbol{\beta}^T\boldsymbol{\phi}_A - \boldsymbol{\beta}^T\boldsymbol{\phi}_E > 0$, it has a small local inconsistency cost for that minority vote edge $A \to E$. However, such $\boldsymbol{\beta}$ 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 a much bigger global inconsistency with the annotation. This enables our model to 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 [33], [19].

We further give two more extreme cases in Figures 2(b) and (c). Due to the Condorcet's paradox, in Figure 2(b) the estimated $\boldsymbol{\beta}$ from majority voting, which removes $A \to E$, is even worse than that from all annotation pairs which at least save the correct annotation $A \to E$. Furthermore, Figure 2(c) shows that when each pair only receives votes in one direction, majority voting will cease to work altogether, but our URLR can still detect outliers by examining the global cost. This example thus highlights the capability of URLR in coping with extremely

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

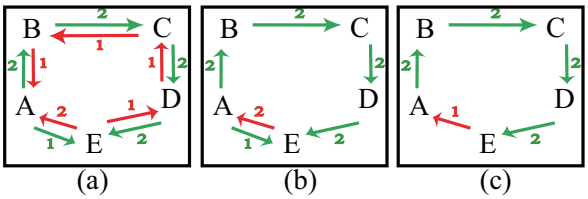


Figure 2. Better outlier detection can be achieved using our URLR framework than majority voting. Green arrows/edges indicate correct annotations, while red arrows are outliers. The numbers indicate the number of votes received by each edge.

sparse pairwise comparison labels. In our experiments (see Section 4), the advantage of URLR over majority is validated on various SVP prediction problems.

3.4.2 Advantage over robust statistical ranking

Our framework is closely related to Huber's theory of robust regression [44], which has been used for robust statistical ranking [53]. In contrast to learning to rank, robust statistical ranking is only concerned with ranking a set of training instances by integrating their (noisy) pairwise rankings. No low-level feature representation of the instances is used as robust ranking does not aim to learn a ranking prediction function that can be applied to unseen test data. To see the connection between URLR with robust ranking, consider the Huber M-estimator [44] which aims to estimate the optimal global ranking for a set of training instances by minimising the following cost function:

$$\min_{\boldsymbol{\theta}} \sum_{i,j} w_{ij} \rho_\lambda((\theta_i - \theta_j) - y_{ij}) \tag{10}$$

where $\boldsymbol{\theta} = [\theta_i] \in \mathbb{R}^{|E|}$ is the ranking score vector storing the global ranking score of each training instance i. The Huber's loss function $\rho_\lambda(x)$ is defined as

$$\rho_\lambda(x) = \begin{cases} x^2/2, & \text{if } |x| \leq \lambda \\ \lambda|x| - \lambda^2/2, & \text{if } |x| > \lambda. \end{cases} \tag{11}$$

Using this loss function, when $|(\theta_i - \theta_j) - y_{ij}| < \lambda$, the comparison is taken as a "good" one and penalised by an l_2-loss for Gaussian noise. Otherwise, it is regarded as a sparse outlier and penalised by an l_1-loss. It can be shown [53] that robust ranking with Huber's loss is equivalent to a LASSO problem, which can been applied to joint robust ranking and outlier detection [47]. Specifically, the global ranking of the training instances and the outliers in the pairwise rankings can be estimated as

$$\left\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}\right\} = \min_{\boldsymbol{\theta},\boldsymbol{\gamma}} \frac{1}{2}\|y - C\boldsymbol{\theta} - \boldsymbol{\gamma}\|^2_{2,\boldsymbol{w}} + \lambda \parallel \boldsymbol{\gamma} \parallel_{1,\boldsymbol{w}} \tag{12}$$

$$= \min_{\boldsymbol{\theta},\boldsymbol{\gamma}} \sum_{e_{ij} \in E} w_{ij} \left[\parallel \frac{1}{2}(y_{ij} - \gamma_{ij} - (\theta_i - \theta_j) \parallel^2 + \lambda|\gamma_{ij}| \right] \tag{13}$$

The optimisation problem (12) is designed for solving the robust ranking problem with Huber's loss function, hence called Huber-LASSO [53].

Our URLR can be considered as a generalisation of the Huber-LASSO based robust ranking problem above. Comparing Eq (12) with Eq (3), it can be seen that the main difference between URLR and conventional robust ranking is that in URLR the cost function has the low-level feature matrix Φ computed from the training instances, and the prediction function parameter $\boldsymbol{\beta}$, such that $\theta = \Phi\boldsymbol{\beta}$. This is because the objective of URLR is to predict SVP for unseen test data. However, URLR and robust ranking do share one thing in common – the ability to detect outliers in the training data based on a Huber-LASSO formulation. This means that, as opposed to our unified framework with feature Φ, one could design a two-step approach for learning to rank by first identifying and removing outliers using Eq (12), followed by introducing the low-level feature matrix Φ and prediction model parameter $\boldsymbol{\beta}$ and estimating $\boldsymbol{\beta}$ using Eq (9). We call this approach Huber-LASSO-FL based learning to rank which differs from URLR mainly in the way outliers are detected without considering low level features.

Next we show that there is a critical theoretical advantage of URLR over conventional Huber-LASSO in detecting outliers from the training instances. This is due to the difference in the projection space for estimating γ which is denoted as Γ. To explain this point, we decompose X in Eq (8) by Singular Value Decomposition (SVD),

$$X = \mathcal{U}\Sigma\mathcal{V}^T \tag{14}$$

where $\mathcal{U} = [\mathcal{U}_1, \mathcal{U}_2]$ with $\mathcal{U}_1$ being an orthogonal basis of the column space of X and $\mathcal{U}_2$ an orthogonal basis of its complement. Therefore, due to the orthogonality $\mathcal{U}^T\mathcal{U} = I$ and $\mathcal{U}_2^T X = 0$, we can simplify Eq (8) into

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \|\mathcal{U}_2^T\boldsymbol{y} - \mathcal{U}_2^T\boldsymbol{\gamma}\|^2_{2,\boldsymbol{w}} + \lambda\|\boldsymbol{\gamma}\|_{1,\boldsymbol{w}}. \tag{15}$$

The SVD orthogonally projects $\boldsymbol{y}$ onto the column space of X and its complement, while $\mathcal{U}_1$ is an orthogonal basis of the column space X and $\mathcal{U}_2$ is the orthogonal basis of its complement Γ (i.e. the kernel space of X^T). With the SVD, we can now compute the outliers $\hat{\boldsymbol{\gamma}}$ by solving Eq (15) which again is a LASSO problem [42], where outliers provide sparse approximations of projection $\mathcal{U}_2^T\boldsymbol{y}$. We can thus compare dimensions of the projection spaces of URLR and Huber-LASSO-FL:

- Robust ranking based on the featureless Huber-LASSO-FL[7]: to see the dimension of the projection space Γ, i.e. the space of cyclic rankings [19], [53], we can perform a similar SVD operation and rewrite Eq (12) in the same form as Eq (15), but this time we have $X = \sqrt{W}C$, $\mathcal{U}_1 \in \mathbb{R}^{|E|\times(|V|-1)}$ and $\mathcal{U}_2 \in \mathbb{R}^{|E|\times(|E|-|V|+1)}$. So the dimension of Γ for Huber-LASSO-FL is $\dim(\Gamma) = |E| - |V| + 1$.
- URLR: in contrast we have $X = \sqrt{W}C\Phi$, $\mathcal{U}_1 \in \mathbb{R}^{|E|\times d}$ and $\mathcal{U}_2 \in \mathbb{R}^{|E|\times(|E|-d)}$. So the dimension of Γ for URLR is $\dim(\Gamma) = |E| - d$.

7. We assume that the graph is connected, that is, $|E| \geq |V| - 1$; we thus have $\text{rank}(C) = |V| - 1$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

From the above analysis we can see that given a very sparse graph with $|E| \sim |V|$, the projection space Γ for Huber-LASSO-FL will have a dimension $(|E| - |V| + 1)$ too small to be effective for detecting outliers. In contrast, by exploiting a low dimensional ($d \ll |V|$) feature representation of the original node space, URLR can enlarge the projection space to that of dimension $|E| - d$. Our URLR is thus able to enlarges its outlier detection projection space Γ. As a result our URLR can better identify outliers, especially for sparse pairwise annotation graphs. In general, this advantage exists when the feature dimension d is smaller than the number of training instance $|V| = N$, and the smaller the value of d, the bigger the advantage over Huber-LASSO. In practice, given a large training set we typically have $d \ll |V|$. On the other hand, when the number of instances is small, and each instance is represented by a high-dimensional feature vector, we can always reduce the feature dimension using techniques such as PCA to make sure that $d \ll |V|$. This theoretical advantage of URLR over conventional Huber-LASSO in outlier detection is validated experimentally in Section 4.

3.4.3 Regularisation on β

It is worth mentioning that in the cost function of URLR (Eq (3)), there are two sets of variables to be estimated, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, but only one l_1 regularisation term on $\boldsymbol{\gamma}$ to enforce sparsity. When the dimensionality of $\boldsymbol{\beta}$ (i.e. d) is high, one would expect to see a l_2 regularisation term on $\boldsymbol{\beta}$ (e.g. ridge regression) due to the fact that the coefficients of highly correlated low-level features can be poorly estimated and exhibit high variance without imposing a proper size constraint on the coefficients [42]. The reason we do not include such a regularisation term is because, as mentioned above, using URLR we need to make sure the low-level feature space dimensionality d is low, which means that the dimensionality of $\boldsymbol{\beta}$ is also low, making the regularisation term $\boldsymbol{\beta}$ redundant. This leads to the applicability of much simpler solvers and we show empirically in the next section that satisfactory results can be obtained with this simplification.

4 Experiments

Experiments were carried out on five benchmark datasets (see Table 1) which fall into three categories: (1) experiments on estimating subjective visual properties (SVPs) that are useful on their own including image (Section 4.1) and video interestingness (Section 4.2), (2) experiments on estimating SVPs as relative attributes for visual recognition (Section 4.3), and (3) experiments on human age estimation from face images (Section 4.4). The third set of experiments can be considered as synthetic experiments – human age is not a subjective visual property although it is ambiguous and poses a problem even for humans [56]. However, as ground truth is available, this set of experiments are designed to gain insights into how different SVP prediction models work.

4.1 Image interestingness prediction

Datasets The image interestingness dataset was first introduced in [24] for studying memorability. It was later re-annotated as an image interestingness dataset by [4]. It consists of 2222 images. Each was represented as a 915 dimensional attribute[8] feature vector [24], [4] such as central object, unusual scene and so on. 16000 pairwise comparisons were collected by [4] using AMT and used as annotation. On average, each image is viewed and compared with 11.9 other images, resulting a total of 16000 pairwise labels[9].

Settings 1000 images were randomly selected for training and the remaining 1222 for testing. All the experiments were repeated 10 times with different random training/test splits to reduce variance. The pruning rate p was set to 20%. We also varied the number of annotated pairs used to test how well each compared method copes with increasing annotation sparsity.

Evaluation metrics For both image and video interestingness prediction, Kendall tau rank distance was employed to measure the percentage of pairwise mismatches between the predicted ranking order for each pair of test data using their prediction/ranking function scores, and the ground truth ranking provided by [4] and [5] respectively. Larger Kendall tau rank distance means lower quality of the ranking order predicted.

Competitors We compare our method (URLR) with four competitors.

1) ***Maj-Vot-1*** [5]: this method uses majority voting for outlier pruning and rankSVM for learning to rank.
2) ***Maj-Vot-2*** [4]: 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. Note that *Maj-Vot-2* was only compared in the experiments on image and video interestingness prediction, since only these two datasets have enough dense annotations for *Maj-Vot-2*.
3) ***Huber-LASSO-FL*:** robust statistical ranking that performs outlier detection using the conventional featureless Huber-LASSO as described in Section 3.4.2, followed by estimating $\boldsymbol{\beta}$ using Eq (9).
4) ***Raw*:** our URLR model without outlier detection, that is, all annotations are used to estimate $\boldsymbol{\beta}$.

Comparative results The interestingness prediction performance of the various models are evaluated while varying the amount of pairwise annotation used. The results are shown in Figure 3 (left). 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, it can

8. We delete 8 attribute features from the original feature vector in [24], [4] such as "attractive" because they are highly correlated with image interestingness.

9. On average, for each labelled pair, around 80% of the annotations agree with one ranking order and 20% the other.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

Dataset	No. pairs	No. img/video	Feature Dim.	No. classes
Image Interestingness [24]	16000	2222	932 (150)	1
Video Interestingness [5]	60000	420	1000 (60)	14
PubFig [54], [2]	2616	772	557 (100)	8
Scene [55], [2]	1378	2688	512 (100)	8
FG-Net Face Age Dataset [56]	–	1002	55	–

Table 1
Dataset summary. We use the original features to learn the ranking model (Eq (9)) and reduce the feature dimension (values in brackets) using Kernel PCA [57] to improve outlier detection (Eq (8)) by enlarging the projection space of γ.

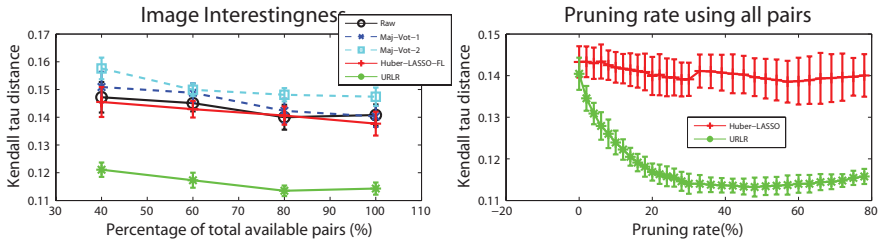


Figure 3. Image interestingness prediction comparative evaluation. Smaller Kendall tau distance means better performance. The mean and standard deviation of each method over 10 trials are shown in the plots.

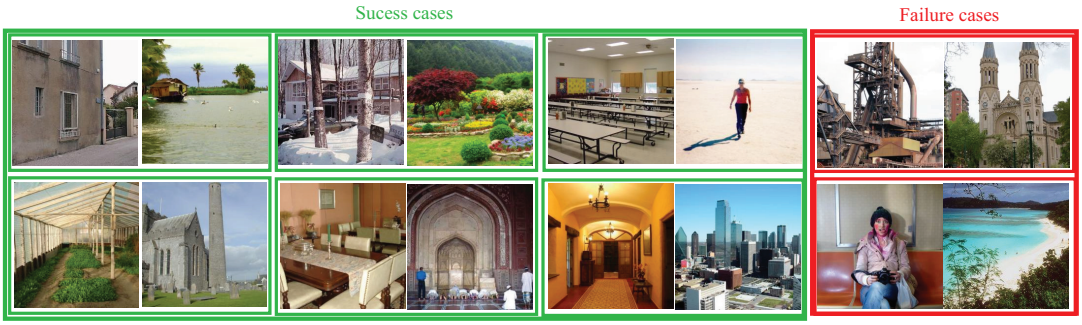


Figure 4. Qualitative examples of outliers detected by URLR. In each box, there are two images. The left image was annotated as more interesting than the right. Success cases (green boxes) show true positive outliers detected by URLR (i.e. right images are more interesting according to the ground truth). Two failure cases are shown in red boxes (URLR thinks the images on the right are more interesting but the ground truth agrees with the annotation).

be observed that: (1) The improvement over *Maj-Vot-1* [5] and *Maj-Vot-2* [4] demonstrates the superior outlier detection ability of *URLR* due to global rather than local outlier detection. (2) *URLR* is superior to *Huber-LASSO-FL* because the joint outlier detection and ranking prediction framework of *URLR* enables the enlargement of the projection space Γ for γ (see Section 3.4.2), resulting in better outlier detection performance. (3) The performance of *Maj-Vot-2* [4] 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. (4) The performance of *Huber-LASSO-FL* is also better than *Maj-Vot-1* and *Maj-Vot-2* suggesting even a weaker global outlier detection approach is better then the majority voting based local one. (5) Interestingly even the baseline method *Raw* gives a comparable result to *Maj-Vot-1* and *Maj-Vot-2* which suggests that just using all annotations without discrimination in a global cost function (Eq (4)) is as effective as majority voting[10].

Figure 3 (right) evaluates how the performances of *URLR* and *Huber-LASSO-FL* 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-FL* gets bigger when more comparisons are pruned showing *Huber-LASSO-FL* stops detecting outliers much earlier on. However, when the pruning rate is over 55%, since most outliers have been removed, inliers start to be pruned, leading to poorer performance.

Qualitative Results Some examples of outlier detection using *URLR* are shown in Figure 4. It can be seen

10. One intuitive explanation for this is that given a pair of data with multiple contradictory votes, using Raw, both the correct and incorrect votes contribute to the learned model. In contrast, with Maj-Vot, one of them is eliminated, effectively amplifying the other's contribution in comparison to Raw. When the ratio of outliers gets higher, Maj-Vot will make more mistakes in eliminating the correct votes. As a result, its performance drops to that of Raw, and eventually falls below it.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

that those in the green boxes are clearly outliers and are detected correctly by our URLR. The failure cases are interesting. For example, in the bottom case, ground truth indicates that the woman sitting on a bench is more interesting than the nice beach image, whilst our URLR predicts otherwise. The odd facial appearance on that woman or the fact that she is holding a camera could be the reason why this image is considered to be more interesting than the otherwise more visually appealing beach image. However, it is unlikely that the features used by URLR are powerful enough to describe such fine appearance details.

4.2 Video interestingness prediction

Datasets The video interestingness dataset is the YouTube interestingness dataset introduced in [5]. It contains 14 categories of advertisement videos (e.g. 'food' and 'digital products'), each of which has 30 videos. $10 \sim 15$ annotators were asked to give complete interesting comparisons for all the videos in each category. So the original annotations are noisy but not sparse. We used bag-of-words of Scale Invariant Feature Transform (SIFT) and Mel-Frequency Cepstral Coefficient (MFCC) as the feature representation which were shown to be effective in [5] for predicting video interestingness.

Experimental settings Because comparing videos across different categories is not very meaningful, we followed the same settings as in [5] and only compared the interestingness of videos within the same category. Specifically, from each category we used 20 videos and their paired comparisons for training and the remaining 10 videos for testing. The experiments were repeated for 10 rounds and the averaged results are reported.

Since MFCC and SIFT are bag-of-words features, we employed χ^2 kernel to compute and combine the features. To facilitate the computation, the χ^2 kernel is approximated by additive kernel of explicit feature mapping [58]. To make the results of this dataset more comparable to those in [5], we used rankSVM model to replace Eq (9) as the ranking model. As in the image interestingness experiments, we used Kendal tau rank distance as the evaluation metric, while we find that the same results can be obtained if the prediction accuracy in [5] is used. The pruning rate was again set to 20%.

Comparative Results Figure 5(a) compares the interestingness prediction methods given varying amounts of annotation, and Figure 5(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-FL* is much closer to our *URLR* now. This is because the advantage of *URLR* over *Huber-LASSO-FL* is stronger when $|E|$ is close to $|V|$. In this experiment, $|E|$ (1000s) is much greater than $|V|$ (20) and the advantage of enlarging the projection space Γ for $\boldsymbol{\gamma}$ (see Section 3.4.2) diminishes.

Qualitative Results Some outlier detection examples are shown in Figure 6. In the two successful detection examples, the bottom videos are clearly more interesting than the top ones, because they (1) have a plot, sometimes with a twist, and (2) are accompanied by popular songs in the background and/or conversations. Note that in both cases, majority voting would consider them inliners. The failure case is a hard one: both videos have cartoon characters, some plot, some conversation, and similar music in the background. This thus corresponds to a truly ambiguous case which can go either way.

4.3 Relative attributes prediction

Datasets The *PubFig* [54] and Scene [55] datasets are two relative attribute datasets. PubFig contains 772 images from 8 people and 11 attributes ('smiling', 'round face', etc.). Scene [55] consists of 2688 images from 8 categories and 6 attributes ('openness', 'natrual' etc.). Pairwise attribute annotation was collected by Amazon Mechanical Turk [2]. Each pair was labelled by 5 workers and majority vote was used in [2] to average the comparisons for each pair[11]. A total of 241 and 240 training images for PubFig and Scene respectively were labelled (i.e. compared with at least another image). The average number of compared pairs per attribute were 418 and 426 respectively, meaning most images were only compared with one or two other images. The annotations for both datasets were thus extremely sparse. GIST and colour histogram features were used for PubFig, and GIST alone for Scene. Each image also belongs to a class (different celebrities or scene types). These datasets were designed for classification, with the predicted relative attribute scores used as image representation.

Experimental Settings We evaluated two different image classification tasks: multi-class classification where samples from all classes were available for training and zero-shot transfer learning where one class was held out during training (a different class was used in each trial with the result averaged). Our experiment setting was similar to that in [6], except that image-level, rather than class-level pairwise comparisons were used. Two settings were used with different amounts of annotation noise:

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

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

11. Thanks to the authors of [2] we have all the the raw pairs data before majority voting.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

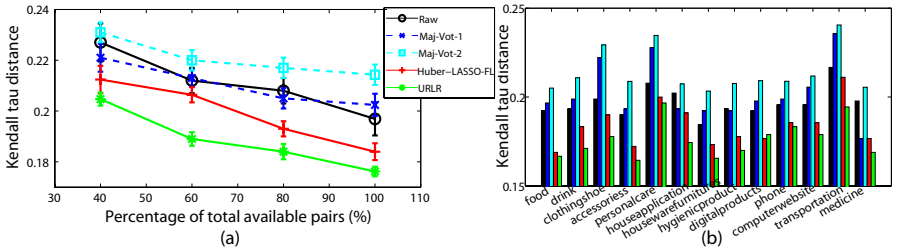


Figure 5. Video interestingness prediction comparative evaluation.

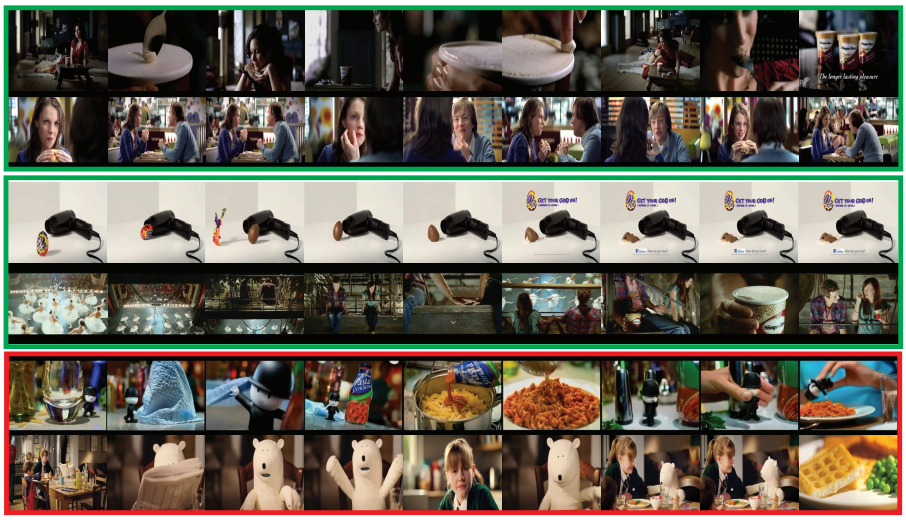


Figure 6. Qualitative examples of video interestingness outlier detection. For each pair, the top video was annotated as more interesting than the bottom. Green boxes indicate the annotations are correctly detected as outliers by our URLR and red box indicates a failure case (false positive). All 6 videos are from the 'food' category.

Evaluation metrics For Scene and Pubfig datasets, relative attributes were very sparsely collected and their prediction performance is thus evaluated indirectly by image classification accuracy with the predicted relative attributes as image representation. Note that for image classification there is ground truth and its accuracy is clearly dependent on the relative attribute prediction accuracy. For both datasets, we employed the method in [6] to compute the image classification accuracy.

Comparative Results Without the ground truth of relative attribute values, different models were evaluated indirectly via image classification accuracy in Figure 7. The following observations can be made: (1) Our URLR always outperforms *Huber-LASSO-FL*, *Maj-Vot-1* 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-FL* is better than *Maj-Vot-1* and Raw often gives better results than majority voting. (3) For PubFig, *Maj-Vot-1* [5] is better than *Raw* given more outliers, but it is not the case for Scene. This is probably because the annotators were more familiar with the celebrity faces in PubFig and 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. some people are unsure 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*, and *Huber-LASSO-FL* to a lesser extent.

Qualitative Results Figure 8 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.

Running Cost Our algorithm is very efficient with a unified framework where all outliers are pruned simultaneously and the ranking function estimation has a closed form solution. Using URLR on PubFig, it took only 1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

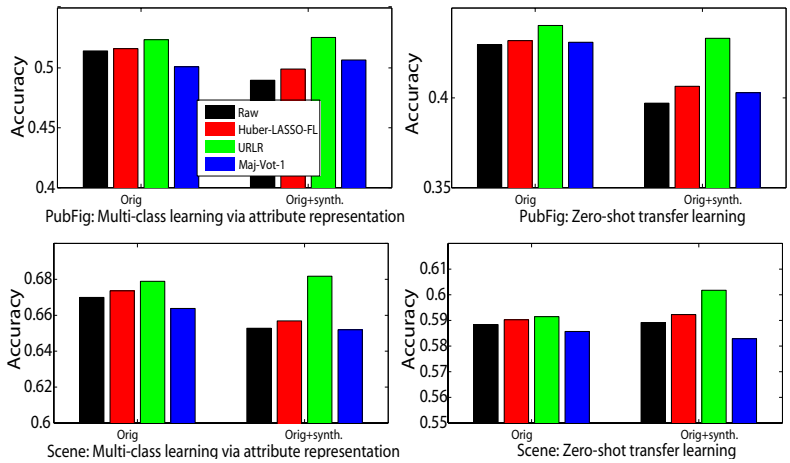


Figure 7. Relative attribute performance evaluated indirectly as image classification rate (chance = 0.125).

Figure 8. Qualitative results on image relative attribute prediction.

minutes to prune 240 images with 10722 comparisons and learn the ranking function for attribute prediction on a PC with four 3.3GHz CPU cores and 8GB memory.

4.4 Human age prediction from face images

In this experiment, we consider age as a subjective visual property of a face. This is partially true – for many people, given a face image predicting the person's age can be subjective. The key difference between this and the other SVPs evaluated so far is that we do have the ground truth, i.e. the person's age when the picture was taken. This enables us to perform in-depth evaluation of the significance of our URLR framework over the alternatives on various factors such as annotation sparsity, and outlier ratio (we now know the exact ratio). Outlier detection accuracy can also now be measured directly.

Dataset The FG-NET image age dataset[12] was employed which contains 1002 images of 82 individuals labelled with ground truth ages ranging from 0 to 69. The training set is composed of the images of 41 randomly selected people and the rest used as the test set. All experiments were repeated 10 times with different training/testing splits to reduce variability. Each image was represented by a 55 dimension vector extracted by active appearance models (AAM) [56].

Crowdsourcing errors We used the ground truth age to generate the pairwise comparisons without any error. Errors were then synthesised according to human error patterns estimated by data collected by an online pilot study[13]: 4000 pairwise image comparisons from 20 willingly participating "good" workers were collected as *unintentional errors*. So we assume they are not contributing random or malicious annotations. Thus the errors of these pairwise comparisons come from the natural data ambiguity. The human unintentional age error pattern was built by fitting the error rate against true age difference between collected pairs. As expected, humans are more error-prone for smaller age difference. Specifically, we fit quadratic polynomial function to model relation of age difference of two samples towards the chance of making an unintentional error. We then used this error pattern to generate unintentional errors. *Intentional errors* were introduced by 'bad' workers who provided random pairwise labels. This was easily simulated by adding random comparisons. In practice, human errors in crowdsourcing experiments can be a mixture of both types. Thus two settings were considered: *Unint.*: errors were generated following the estimated human unintentional error model resulting in around 10% errors. *Unint.+Int.*: random comparisons were added on top of *Unint.*, giving an error ratio of around 25%, unless otherwise stated. Since the ground-truth age of each face image is known to us, we can give an upper bound for all the compared methods by using ground-truth age of training data to generate a set of pairwise comparisons.

12. http://www.fgnet.rsunit.com/

13. http://www.eecs.qmul.ac.uk/~yf300/survey4/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

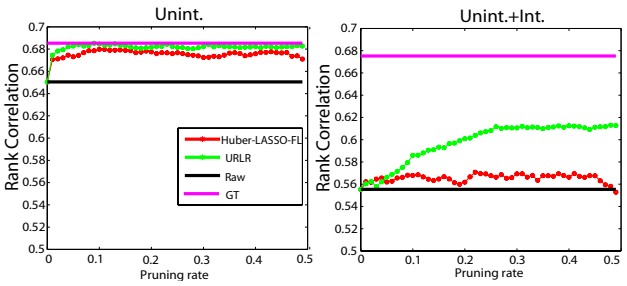


Figure 9. Comparing *URLR* and *Huber-LASSO-FL* on ranking prediction under two error settings. Note that the ranking prediction accuracy is measured using Kendall tau rank correlation which is very similar to Kendall tau distance (see [59]). With rank correlation, the higher the value the better the performance.

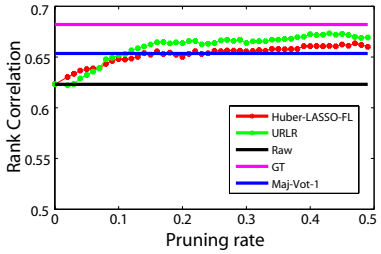


Figure 10. Comparing *URLR* and *Huber-LASSO-FL* against majority voting (5 comparisons per pair).

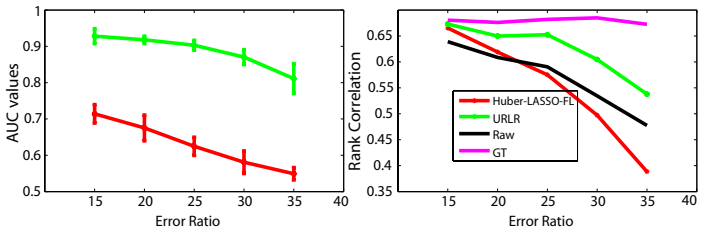


Figure 11. Effect of error ratio. Left: outlier detection performance measured by area under ROC curve (AUC). Right: rank prediction performance measured by rank correlation.

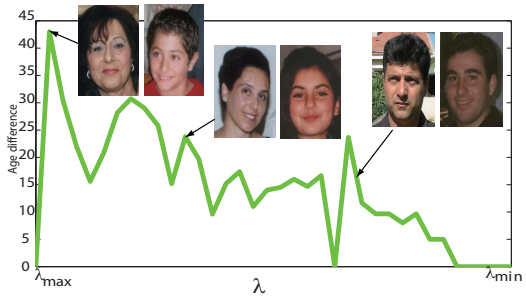


Figure 12. Relationship between the pruning order and actual age difference for URLR.

This outlier-free dataset is then used to learn a kernel ridge regression with Gaussian kernel. This ground-truth data trained model is denoted as *GT*.

Quantitative results Four experiments were conducted using different settings to show the effectiveness of our URLR method quantitatively.

(1) URLR vs. Huber-LASSO-FL. In this experiment, 300 training images and 600 unique comparisons were randomly sampled from the training set. Figure 9 shows that *URLR* and *Huber-LASSO-FL* improve over *Raw* indicating that outliers are effectively pruned using both global outlier detection methods. Both methods are robust to low error rate (Figure 9 Left: 10% in Unint.) and are fairly close to GT, whilst the performance of *URLR* is significantly better than that of *Huber-LASSO-FL* given high error ratio (Figure 9 Right: 25% in Unint.+Int.) because of the using low-level feature representation to increase the dimension of projection space dimension for γ from 301 for *Huber-LASSO-FL* to 546 for *URLR* (see Section 3.4.2). This result again validates our analysis that higher $\dim(\Gamma)$ leads to better chance of identifying outliers correctly. It is noted that in Figure 9(Right), given 25% outliers, the result indeed peaks when p is around 25; importantly, it stays flat when up to 50% of the annotations are pruned.

(2) Comparison with Maj-Vot-1. Given the same data but each pair compared by 5 workers (instead of 1) under the Unint.+Int. error condition, Figure 10 shows that *Maj-Vot-1* beats *Raw*. This shows that for relative dense graph, majority voting is still a good strategy of removing some outliers and improves the prediction accuracy. However, *URLR* outperforms *Maj-Vot-1* after the pruning rate passes 10%. This demonstrates that aggregating all paired comparisons globally for outlier pruning is more effective than aggregating them locally for each edge as done by majority voting.

(3) Effects of error ratio. We used the Unint.+Int. error model to vary the amount of random comparisons and simulate different amounts of errors in 10 sampled graphs from 300 training images and 2000 unique sampled pairs from the training images. The pruning rate was fixed at 25%. Figure 11 shows that *URLR* remains effective even when the true error ratio reaches as high as 35%. This demonstrates that although a sparse outlier model is assumed, our model can deal with non-sparse outliers. It also shows that URLR consistently outperforms the alternative models especially when the error/outlier ratio is high.

What are pruned and in what order? The effectiveness of the employed regularisation path method for outlier detection can be examined as λ decreases to produce a ranked list for all pairwise comparisons according to the outlier probability. Figure 12 shows the relationship between the pruning order (i.e. which pair is pruned first) and ground truth age difference and illustrated by examples. It can be seen that overall outliers with larger age difference tend to be pruned first. This means that even with a conservative pruning rate, obvious outliers (potentially causing more performance degradation in learning) can be reliably pruned by our model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

5 Conclusions and Future Work

We have proposed a novel unified robust learning to rank (URLR) framework for predicting subjective visual properties from images and videos. 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 feature based rank prediction formulation also provides our model with an advantage over the conventional robust ranking methods without features for outlier detection: it can be applied with a large number of candidates in comparison but a sparse sampling in crowdsourcing. The effectiveness of our model in comparison with state-of-the-art alternatives has been validated on the tasks of image and video interestingness prediction and predicting relative attributes for visual recognition. Its effectiveness for outlier detection has also been evaluated in depth in the human age estimation experiments.

By definition subjective visual properties (SVPs) are person-dependent. When our model is learned using pairwise labels collected from many people, we are essentially learning consensus – given a new data point the model aims to predict its SVP value that can be agreed upon by most people. However, the predicted consensual SVP value could be meaningless for a specific person when his/her taste/understanding of the SVP is completely different to that of most others. How to learn a person-specific SVP prediction model is thus part of the on-going work. Note that our model is only one of the possible solutions to inferring global ranking from pairwise comparisons. Other models exist. In particular, one widely studied alternative is the (Bradley-Terry-Luce (BTL) model [60], [61], [62]), which aggregates the ranking scores of pairwise comparisons to infer a global ranking by maximum likelihood estimation. The BTL model is introduced to describe the probabilities of the possible outcomes when individuals are judged against one another in pairs [60]. It is primarily designed to incorporate contextual information in the global ranking model. We found that directly applying the BTL model to our SVP prediction task leads to much inferior performance because it does not explicitly detect and remove outliers. However, it is possible to integrate it into our framework to make it more robust against outliers and sparse labels whilst preserving its ability to take advantage of contextual information. Other new directions include extending the presented work to other applications where noisy pairwise labels exist, both in vision such as image denoising [63], iterative search and active learning of visual categories [30], and in other fields such as statistics and economics [19].

References

[1] J. Donahue and K. Grauman, "Annotator rationales for visual recognition," in *ICCV*, 2011.

[2] A. Kovashka, D. Parikh, and K. Grauman, "Whittlesearch: Image search with relative attribute feedback," in *CVPR*, 2012.

[3] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *CVPR*, 2011.

[4] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool, "The interestingness of images," in *ICCV*, 2013.

[5] Y.-G. Jiang, YanranWang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," in *AAAI*, 2013.

[6] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011.

[7] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Robust relative attributes for human action recognition," *Pattern Analysis and Applications*, 2013.

[8] K. Chen, C. Wu, Y. Chang, and C. Lei, "Crowdsourceable QoE evalutation framework for multimedia content," in *ACM MM*, 2009.

[9] Y. Ma, T. Xiong, Y. Zou, and K. Wang, "Person-specific age estimation under ranking framework," in *ACM ICMR*, 2011.

[10] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Inf. Retr.*, 2010.

[11] X. Chen and P. N. Bennett, "Pairwise ranking aggregation in a crowdsourced setting," in *ACM International Conference on Web Search and Data Mining*, 2013.

[12] O. Wu, W. Hu, and J. Gao, "Learning to rank under multiple annotators," in *IJCAI*, 2011.

[13] C. Long, G. Hua, and A. Kapoor, "Active visual recognition with expertise estimation in crowdsourcing," in *ICCV*, 2013.

[14] A. Kittur, E. H. Chi, and B. Suh., "Crowdsourcing user studies with mechanical turk," in *ACM CHI*, 2008.

[15] A. Kovashka and K. Grauman, "Attribute adaptation for personalized image search," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[16] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *NIPS*, pp. 2424–2432, 2010.

[17] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *ICML*, pp. 889–896, 2009.

[18] J. Whitehill, T. fan Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *NIPS*, 2009.

[19] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial hodge theory," *Math. Program.*, 2011.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. of the Royal Statistical Society, Series B*, 1996.

[21] I. Gannaz, "Robust estimation and wavelet thresholding in partial linear models," *Stat. Comput.*, vol. 17, pp. 293–310, 2007.

[22] F. L. Wauthier, N. Jojic, and M. I. Jordan, "A comparative framework for preconditioned lasso algorithms," in *Neural Information Processing Systems*, 2013.

[23] Y. Ke, X. Tang, , and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, 2006.

[24] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?," in *CVPR*, 2011.

[25] F. Liu, Y. Niu, and M. Gleicher, "Using web photos for measuring video frame interestingness," in *IJCAI*, 2009.

[26] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009.

[27] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[28] A. Shrivastava, S. Singh, and A. Gupta, "Constrained semi-supervised learning via attributes and comparative attributes," in *ECCV*, 2012.

[29] A. Parkash and D. Parikh, "Attributes for classifier feedback," in *ECCV*, 2012.

[30] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers and attributes via relative feedback," in *CVPR*, 2013.

[31] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *CVPR Workshops*, 2008.

[32] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes.," in *CVPR*, 2012.

[33] W. V. Gehrlein, "Condorcet's paradox," *Theory and Decision*, 1983.

[34] L. Liang and K. Grauman, "Beyond comparing image pairs: Setwise active learning for relative attributes," in *CVPR*, 2014.

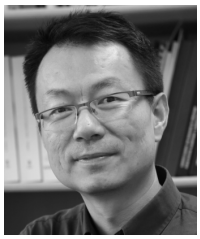
[35] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, "Hodgerank on random graphs for subjective video quality assessment," *IEEE TMM*, 2012.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2456887, IEEE Transactions on Pattern Analysis and Machine Intelligence

[36] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," in *ACM MM*, 2012.
[37] M. Maire, S. X. Yu, and P. Perona, "Object detection and segmentation from joint embedding of parts and pixels," in *ICCV*, 2011.
[38] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *ICML*, 2007.
[39] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li, "Browserank: letting web users vote for page importance," in *ACM SIGIR*, 2008.
[40] Z. Sun, T. Qin, Q. Tao, and J. Wang, "Robust sparse rank learning for non-smooth ranking measures," in *ACM SIGIR*, 2009.
[41] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao, "Interestingness prediction by robust learning to rank," in *ECCV*, 2014.
[42] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning:Data Mining, Inference, and Prediction (2nd)*. Springer, 2009.
[43] A. N. Hirani, K. Kalyanaraman, and S. Watts, "Least squares ranking on graphs," *arXiv:1011.1716*, 2010.
[44] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.
[45] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.
[46] A. Belloni, V. Chernozhukov, and L. Wang, "Pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, pp. 791–806, 2011.
[47] Y. She and A. B. Owen, "Outlier detection using nonconvex penalized regression," *Journal of American Statistical Association*, 2011.
[48] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
[49] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *JASA*, 2001.
[50] J. Fan, R. Tang, and X. Shi, "Partial consistency with sparse incidental parameters," *arXiv:1210.6950*, 2012.
[51] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, 2004.
[52] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, 2006.
[53] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing," in *ACM MM*, 2013.
[54] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.
[55] A. Oliva and A. Torralba., "Modeling the shape of the scene: Aholistic representation of the spatial envelope," *IJCV*, vol. 42, 2001.
[56] Y. Fu, G. Guo, and T. Huang, "Age synthesis and estimation via faces: A survey," *TPAMI*, 2010.
[57] S. Mika, B. Scholkopf, A. Smola, K.-R. Muller, M. Scholz, and G. Ratsch, "Kernel PCA and de-noising in feature spaces," in *NIPS*, pp. 536–542, 1999.
[58] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," in *IEEE TPAMI*, 2011.
[59] B. Carterette, "On rank correlation and the distance between rankings," in *ACM SIGIR*, 2009.
[60] R. Hunter, "Mm algorithms for generalized bradley-terry models," *The Annals of Statistics*, vol. 32, p. 2004, 2004.
[61] H. Azari Soufiani, W. Chen, D. C. Parkes, and L. Xia, "Generalized Method-of-Moments for Rank Aggregation," in *NIPS*, 2013.
[62] F. Caron and A. Doucet, "Efficient bayesian inference for generalized bradley-terry models," 2012.
[63] S. X. Yu, "Angular embedding: A robust quadratic criterion," *TPAMI*, 2012.

Yanwei Fu received the PhD degree from Queen Mary University of London in 2014, and the MEng degree in the Department of Computer Science & Technology at Nanjing University in 2011, China. He is currently a Post-doc of Disney Research, Pittsburgh. His research interest is image and video understanding.

Timothy M. Hospedales received the PhD degree in neuroinformatics from the University of Edinburgh in 2008. He is currently a lecturer (assistant professor) of computer science at Queen Mary University of London. His research interests include probabilistic modelling and machine learning applied variously to problems in computer vision, data mining, interactive learning, and neuroscience. He has published more than 30 papers in major international journals and conferences. He is a member of the IEEE.



Tao Xiang received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a reader (associate professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision and machine learning. He has published over 100 papers in international journals and conferences and co-authored a book, Visual Analysis of Behaviour: From Pixels to Semantics.

Jiechao Xiong is currently pursuing the Ph.D. degree in statistics in BICMR & School of Mathematical Science, Peking University, Beijing, China. His research interests include statistical learning, data science and topological and geometric methods for high-dimension data analysis.

Shaogang Gong is Professor of Visual Computation at Queen Mary University of London, a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil in computer vision from Keble College, Oxford University in 1989. His research interests include computer vision, machine learning and video analysis.

Yizhou Wang Dr. Yizhou Wang is a Professor of Computer Science Department at Peking University, Beijing, China. He is a vice director of Institute of Digital Media at Peking University, and the director of New Media Lab of National Engineering Lab of Video Technology. He received his Ph.D. in Computer Science from University of California at Los Angeles (UCLA) in 2005. Dr. Wang's research interests include computational vision, statistical modeling and learning, pattern analysis, and digital visual arts.

Yuan Yao received his Ph.D. in mathematics from the University of California, Berkeley, in 2006. Since then he has been with Stanford University and in 2009, he joined the School of Mathematical Sciences, Peking University, Beijing, China, as a professor of statistics. His current research interests include topological and geometric methods for high dimensional data analysis and statistical machine learning, with applications in computational biology, computer vision, and information retrieval.