Supplementary of "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

June 3, 2015

Abstract

Thanks for the excellent questions from the anonymous reviewers of our TPAMI submission. In answering their questions, we found some details and insights of our framework which have been overlooked before. Due to the page limits of our journal version, we use this document to further explain the details and insights and help our readers better understand our work.

1. Further, the proposed approach doesn't seem to truly get to the bottom of why subjective properties are tricky namely that two people might actually have a different understanding of the property. While the authors do refer to such possible disagreements in the introduction, the proposed method doesn't seem to consider this possibility. In other words, how does it make sense to consider a single global order when such an order might be unattainable since person A's "interestingness" will differ from person B's?

   This is a very good question. Indeed, since the properties are subjective, they are by definition person-dependent. However, in most applications when we learn a SVP prediction model using pairwise labels collected from many different annotators, we are modeling consensus. In other words, the model essentially aggregates the understandings of different people regarding a certain SVP so that the predicted SVP for an unseen data point can be agreed upon by most people. For example, in the case of video interestingness, YouTube may want to predict the interestingness of a newly uploaded video so as to decide whether or not to promote it. Such a prediction obviously needs to be based on consensus from the majority of the YouTube viewers regarding what defines interestingness. However, collecting consensus can be expensive; the proposed model in this paper thus aims to infer the consensus from as few labels as possible.

   It is also true that for a specific person, he/she would prefer a SVP prediction model that is tailor-made for his/her own understanding of the SVP, i.e. a person-specific prediction model. Such a model needs to be learned using his/her pairwise labels only.

For example, YouTube could recommend different videos for different registered users when they log in, if they provide some pairwise video interestingness labels for learning such a model (at present, this is done based on some simple rules from the viewing history of the user). This also has its own problem - it is much harder to collect enough labels from a single person only to learn the prediction model. There are solutions, e.g. categorising the users into different groups so that the labels from people of the same group can be shared. However this is beyond the scope of this paper and is being considered as part of ongoing work.

We have provide a discussion on this problem in Section 5 in the revised manuscript (Page 14).

2. It feels a little bit unsatisfying that the method requires we pick a fixed ratio of outliers. This would be more ok if the ratio can be automatically computed from the data somehow.

   Indeed, the pruning rate is a free parameter of the proposed model (in fact, the only free parameter) that has to set manually. As discussed in the beginning of Section 3.3, most existing outlier detection algorithms have a similar free parameter determining how aggressive the algorithm needs to be for pruning outliers. Automated model selection critieria such as BIC and AIC could be considered. However, as pointed out by [49], they are often unstable for the outlier detection problem with pairwise labels. We have carried out experiments to show that when BIC or AIC is employed, the selected model failed to detect meaningful outliers. Since a related comment is given by Reviewer 3, please refer to the Response Point 2 to Reviewer 3 for detailed experiment results and analysis on the alternative outlier detection methods including BIC. It is also worth pointing out that our results on the effect of the pruning rate show that the proposed model remains effective given a wide range of pruning rate values (see Fig. 3, 5, 9 and 10).

   We have now added a footnote in Section 3.3 to discuss why an automated model selection criterion such as BIC is not adopted.

3. I think cases of Raw performing similarly or better than MajVot1/ 2 should be explained in a little more detail, i.e. an intuition for such outcomes should be given.

   Thanks for the suggestion. Indeed, our results on both image and video interestingness experiments show that Raw performs similarly to majority voting. There is an intuitive explanation for that. When a pair of data points A and B receive multiple votes/labels of different pairwise orders/ranks, these multiple labels are converted into a single label corresponding to the order that receives the most votes. Since only one of the two orders is correct (either A>B or B>A), there are two possibilities: the majority voted label is correct, or incorrect, i.e. an outlier. In comparison, using Raw, all votes count, so the outlying votes would certainly having a negative effect on the learned prediction model, so would the correct votes/labels. Now let us consider which method is better. The answer is it depends on the outlier/error ratio of the labels. If the ratio is very

low, majority voting will get rid of almost all the outlying votes; MajVot would thus be advantageous over Raw which can still feel the negative effects of the outliers. However, when the ratio gets bigger, it becomes possible that the outlying label becomes the winning vote. For example, if A>B is correct, and received 2 votes and A<B is incorrect and received 3 outlying votes. Using Raw, those 2 correct votes still contribute positively to the model, whilst using MajVot, their contribution disappears and the negative impact of the outlying votes is amplified. Therefore, one expects that when MajVot makes more and more mistakes, its performance will get closer to that of Raw, until it reaches a tipping point where Raw starts to get ahead.

We have added a brief discussion on this in Section 4.1. on Page 9.

4. In Figure 3, why does the Kendall tau distance start to increase as the pruning rate increases, after 55% for URLR?

   Higher Kendall tau distance means worse prediction. Figure 3 (right) thus shows that our URLR's performance is improved when more and more outliers are pruned in the beginning; then after more than 55% of pairs are pruned, its performance starts to decrease. This result is expected: at low pruning rates, most of the pruned pairs are outliers; the model therefore benefits. Since the percentage of outliers would almost certainly be lower than 50%, when the pruning rate reaches 55%, most of the outliers have been removed, and the algorithm start to remove the correctly labelled pairs. With less and less correct labels available to learn the model, the performance naturally would decrease - when pruning rate gets close to 100%, it would not be possible to learn a meaningful model; the Kendall tau distance would thus shoot up.

   We have now added a sentence on Page 10 to give an explanation to this phenomenon.

5. Page 2 line 44 "For example, Figure 1 . . . " the authors try to argue that examples shown in Figure 1 are outliers. I don't quite agree. Authors are trying to study subjective attributes. These are good examples of subjective versus objective attributes. This doesn't seem to be about outliers vs. not. In fact, one source of outliers other than malicious workers is global (in)consistency, which is not mentioned here. The authors could draw from the concrete example of Figure 2.

   This is a very good point. It is certainly worthwhile to clarify the definition of outlier in the context of subjective visual property (SVP). In particular, since by definition a SVP is subjective, defining outlier, even making the attempt to predict SVP is self-contradictory - one man's meat is another man's poison. However, there is certainly a need for learning a SVP prediction model, hence this paper. This is because 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 (please see our Response Points 1 for more discussion on this). Therefore, Figure 1(a) can still be used to illustrate this outlier issue in SVP annotation, that is, you may have 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 thus 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. We have reworded the relevant text on Page 2 to avoid confusion.

6. A baseline to compare to might be to feed all individual constraints (without majority vote) to a rankSVM. SVMs already allow for some slack. So I would be curious to know if that takes care of some of the outliers already.

   Thanks. In fact, we do have one set of results on this. Specifically, in Sec 4.2 "Video interestingness prediction", as explained under "Experimental settings", we employed rankSVM model to replace Eq (9). Therefore, the model denoted as 'Raw' in this experiment is exactly the suggested baseline of feeding all constraints to a rankSVM. As shown in Fig. 5(a), the model is at par with Maj-Vot-1 but worse then the two global outlier detection methods Huber-LASSO-FL and our URLR. This result suggests that rankSVM does have some ability to cope with outliers. However, we are not sure this is due to the slack variables of rankSVM. This is because the slack variables are introduced to account for data noise [1] which is different from the outliers in the pairwise data.

7. For the scene and pubfig image dataset, the relative attribute prediction performance can only be evaluated indirectly by image classification accuracy with the predicted relative attributes as image representation." > Why is that? Can't you compute attribute prediction performance on a held out set of annotated pairs? Or is the concern that since the pairs may be noisily annotated, one can not think of them as GT? But is that not an issue with interestingness then? Please clarify in rebuttal.

   Thanks for this question. We stated in footnote 9 that "Collecting ground truth for subjective visual properties is always problematic. Recent statistical theories [61], [19] suggest that the dense human annotations can give a reasonable approximation of ground truth for pairwise ranking. This is how the ground truth pairwise rankings provided in [4] and [5] were collected." So for image and video interestingness as well as the age dataset, (dense) enough pairwise comparisons are available to give a reasonable approximation of the groundtruth. However, this is not the case for scene and pubfig image dataset: the collected pairs are much more sparse and cannot be used as an approximation to the groundtruth. In short, it is because they are too sparse rather than too noisy.
   In contrast, the indirect evaluation metric of downstream classification accuracy has clear unambiguous groundtruth, and directly depends on relative attribute prediction accuracy. So this evaluation is preferred.

8. Related Work: The Bradley-Terry-Luce (BTL) model is the standard model for computing a global ranking from pairwise labels. It should be mentioned in the related work. See [52] or Hunter, D. R. (2004). MM algorithms for generalized BradleyTerry models. Annals of Statistics. Experiments: I would expect additional comparisons to

state-of-the-art (BTL or SVM-rank aggregation [52]). In particular the Bradley-Terry-Luce (BTL) model is extremely widely used and more robust to noise than LASSO based approaches [52]. E.g. "Generalized Method-of-Moments for Rank Aggregation" or "Efficient Bayesian Inference for Generalized Bradley-Terry Models" provide code for inference in BTL models. Such a method leads to a global ranking, which could be used to train an SVM. Alternatively, it can be used to find pairwise rankings that disagree with the obtained global ranking. These could be removed as outliers and a rank-SVM trained from the remaining pairwise labels. Such an experiment should be included as an additional state-of-the-art comparison in the updated version of the manuscript.

Thanks for the suggestion. Indeed, the Bradley-Terry-Luce (BTL) model is a very relevant global ranking model. We have now studied it carefully and made connections to the proposal URLR model. We also carried out new experiments to evaluate the BLT model for our Subjective Visual Property (SVP) prediction task.

More specifically, the BTL model is a probabilistic model that aggregates the ranking scores of pairwise comparisons to infer a global ranking by maximum likelihood estimation. It is closely related to the proposed global ranking model; yet it also has some vital differences. Let's first look at the connection. The main pairwise ranking model of Huber-LASSO used in this paper is a linear model (see Eq (10) and Eq (12)), which is

$$y_{ij} = \theta_i - \theta_j + \gamma_{ij} + \varepsilon_{ij} \tag{1}$$

In statistics and psychology [2, 3, 4, 5], such a linear model can be extended to a family of generalised linear models when only binary comparisons are available for each pair (i, j), i.e. either i is preferred to j or vice versa. In these generalised linear models, one assumes that the probability of pairwise preference is fully determined by a linear ranking/rating function in the following,

$$\pi_{ij} = \text{Prob}\left\{i \, is \, preferred \, over \, j\right\} = \Phi\left(\theta_i - \theta_j\right)$$

where $\Phi : \mathbb{R} \to [0, 1]$ can be chosen as any symmetric cumulated distributed function.

Different choices of Φ lead to different generalised linear models. In particular, two choices are worth mentioning here:

- *Uniform model,*

$$y_{ij} = 2\pi_{ij} - 1 \tag{2}$$

  This model is equivalent to use $y_{ij} = 1$ if i is preferred to j and $y_{ij} = -1$ otherwise in linear model. This model is used in this work to derive our URLR model.

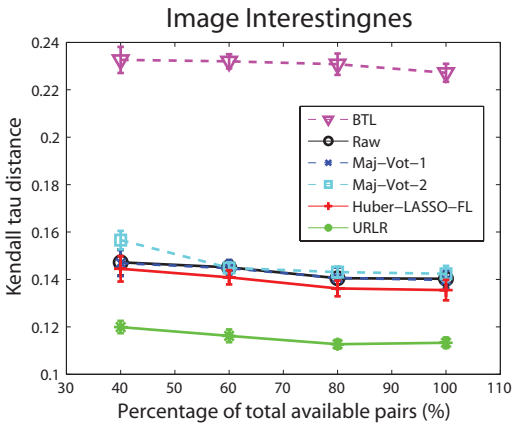

Figure 1: Comparing the BTL model with our model on image interestingness prediction

- *Bradley-Terry-Luce (BTL) model,*

$$y_{ij} = \log \frac{\pi_{ij}}{1 - \pi_{ij}} \tag{3}$$

So by now, it is clear that both our URLR and BTL generalise the linear model in Huber-LASSO. They differ in the choice of the symmetric cumulated distributed function Φ*.*

Although both of them are generalised from the same linear model, they are developed for very different purposes. The BTL model is introduced to describe the probabilities of the possible outcomes when individuals are judged against one another in pairs [6, 5]. It is primarily designed to incorporate contextual information in the global ranking model. For instance, in sports applications, it can be used to account for the home-field advantage and ties situations [3, 7]. In contrast, our framework tries to detected outliers in the pairwise comparisons and cope with the sparse labels. Consequently, from Eq (1) onwards, we introduce the outlier variable to model the outiers explicitly and introduce low-level feature variable to enhance our model's ability to detect outliers given sparse labels. None of these is in the BLT model, which means that it may not be suitable given sparse pairwise comparisons with outliers.

To verify this, we took the suggestion by Reviewer 3 and employed the matlab codes from the website of [7] "Efficient Bayesian Inference for Generalized BradleyTerry Models" to carry out experiments. The results on image interestingness prediction are compared in Fig 1. It shows that the performance of BTL is much worse than the other alternatives. Similar results were obtained on video interestingness prediction and age estimation.

As explained above, it is actually not fair to compare the BTL model to the other models because BTL was not designed for outlier detection and could not cope with the amount

of outliers and the level of spareness in our SVP data. We therefore decide not to include the new results in the revised manuscript. However, from our analysis above, it is also clear that we could use the BTL model (Eq (3)) to generalise the linear model in place of the uniform model, and use it in our outlier detection framework. In this way, we can have the better of both worlds: the ability of BTL to incorporate contextual information such as the home-filed advantage in sports can also be taken advantage of in our framework whilst preserving our model's strength on robustness against outliers and sparse labels. However, this is probably beyond the scope of this paper and is better left to the future work. In the revised manuscript, we have now added the following paragraph in Section 5, where we discuss that BTL is an alternative model that can be integrated into our framework as part of the future work.

> *"Note that our model is only one of the possible solutions to inferring global ranking from pairwise comparisons. In particular, one widely studied alternative is the (Bradley-Terry-Luce (BTL) model [61,62,63], 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 [61]. 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."*

9. 3.3 Regularization path. On the one hand the authors say that "Setting a constant λ value independent of dataset is far from optimal because the ratio of outliers may vary for different crowdsourced datasets", but using the regularization path this is exactly what is done in the end. It is true that the experiments show that the proposed method is fairly robust w.r.t. the outlier ratio. Nonetheless, I would like to see an experiment using a (modified) BIC for selecting the outlier ratio. This would be a valuable extension over the ECCV work.

   Thanks. As discussed in the beginning of Section 3.3, most existing outlier detection algorithms have a similar free parameter as λ to determine how aggressive the algorithm needs to be for pruning outliers. Automated model selection critieria such as BIC and AIC could be considered. However, as pointed out by [49], they are often unstable for the outlier detection problem with pairwise labels.

   We have evaluated alternative methods including the modified BIC and AIC for image and video interestingness prediction. The results suggest those automated models such as AIC and BIC failed to identify any outliers - they prefer the model that include all input pairwise comparisons. To find out why it is the case, we carried out a controlled

experiment using synthetic data to investigate how different factors affect the performance of different methods for determining the outlier ratio. Specifically, we compare , BIC and with our Regularization Path model.

Experiments design. *we use a complete graph G with 30 nodes. Our framework is simplified into the following ranking model,*

$$Y_{ij} = \theta_i - \theta_j + \gamma_{ij} + \varepsilon_{ij}$$

Let $\theta \sim U(-1, 1)$, $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $\gamma_{ij} = \pm L$. We simulate the outlier pairs by randomly sampling, that is, each pair's true ranking is reversed (i.e. becoming an outlier/error) with a probability p which will determine the outlier ratio. The magnitude of outliers in relation to that of the noise is another factor which could potentially affect the performance of different methods on outlier detection. So we define the outlier-noise-ratio $ONR := L/\sigma$, where $\sigma = 0.1$ in our experiment and L is varied in our experiment to give different ONR values.

Evaluation protocols and results. *We first compare three methods that require the manual setting of a free parameter corresponding to the outlier ratio. These include our formulation (Eq (8)) with Regularization Path (i.e. the proposed model), IPOD hard-threshold [8][1] with Regularization Path, and our formulation with orthogonal matching pursuit [9]. Using our model with Regularization Path, λ is decreased from ∞ to 0 and the graph edges are order according to how likely it corresponds to an outlier. The top $p\%$ edge set Λ_p are detected as outliers. By varying p, ROC (receiver-operating-characteristic) curve can be plotted and AUC (area under the curve) is computed. Similarly, IPOD hard-threshold can also be solved using the same Regularization Path strategy. And orthogonal matching pursuit can be used to solve our formulation for outlier detection in place of Regularization Path. As shown in Figure 2, the results of our formulation with Regularization Path are consistently better than those of IPOD hard-threshold + Regularization Path and our formulation + orthogonal matching pursuit. Specifically, it shows that (1) when there are small portions of outliers, all the methods can reliably prune most of outliers; (2) in all experiments, IPOD-hard threshold and orthogonal matching pursuit have similar performance, whilst our formulation + Regularization Path is consistently better than the other alternatives, especially when there are large portions of outliers (high values of p); (3) the higher the ONR, the better performance of outlier detection for all three methods.*

In contrast, BIC utilises the relative quality and likelihood functions of statistical models themselves to determine a fixed λ. Therefore, the true positive rate (TPR) and false

[1] *Strictly speaking, IPOD hard-threshold is not a Lasso solver, since it replaced the soft-thresholding with hard-thresholding. However, for comparison convenience, we still compare it with our RP.*

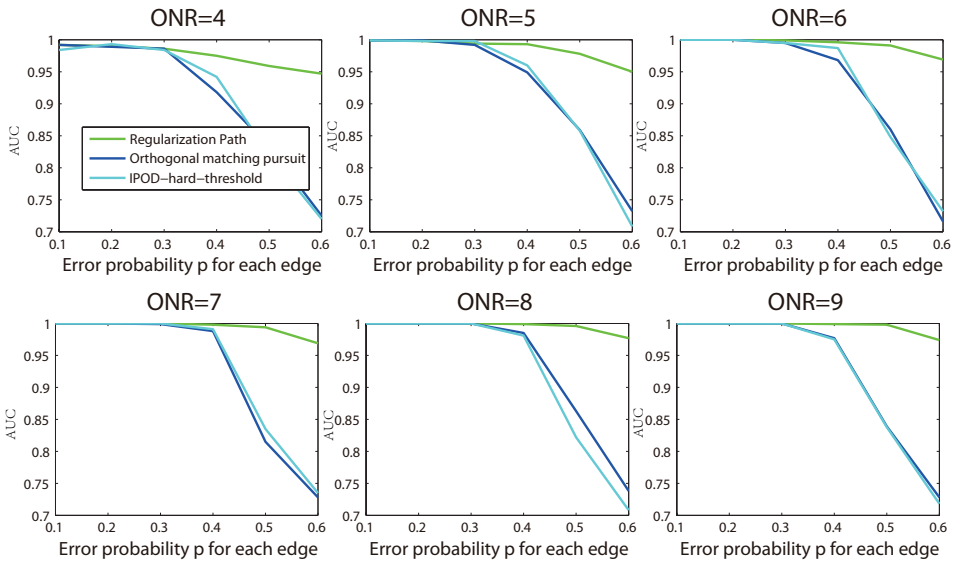

Figure 2: Effects of outlier/error probability (p) and outlier-noise ratio (ONR) on our formulation + Regulaization Parth (denoted as Regulaization Parth), IPOD-hard threshold + Regulaization Parth and our formulation + Orthogonal matching pursuit.

	ONR=4	ONR=5	ONR=6	ONR=7	ONR=8	ONR=9
p=0.1	0.002/0	0.494/0.012	1/0.003	1/0.026	1/0.025	1/0.031
p=0.2	0/0	0/0	0.3/0.016	0.9/0.05	1/0.064	1/0.037
p=0.3	0/0	0/0	0/0	0/0	0/0	0.5/0.06
p=0.4	0/0	0/0	0/0	0/0	0/0	0/0
p=0.5	0/0	0/0	0/0	0/0	0/0	0/0
p=0.6	0/0	0/0	0/0	0/0	0/0	0/0

Table 1: The outlier detection results of our formulation + BIC. The results are presented as TPR/FPR. The error probability and ONR are: $p \in [0.1, 0.6]$ and $ONR \in [4, 9]$ respectively.

positive rate (FPR) for BIC are reported. The results are listed in Table 1. It shows that when using our formulation with BIC, only when there are very small portions of outliers and the outlier-noise-ratio is extremely high, BIC can reliably prune most of outliers. Otherwise, it tends to consider all pairs inliers. As mentioned above, using BIC in place of Regularization Path also leads to no outliers being pruned in our SVP prediction experiments. This thus suggests that the real outlier ratio (roughly corresponds to p=0.2, see Response Point 10 to Reviewer 2) and/or outlier-noise-ratio (ONR) are too high for BIC to work.

Due to the space constraint, we could not include all these results and analysis in the revised manuscript. On Page 6, we have now added a footnote (Footnote 3) to refer the readers to find additional results and discussion on this outlier ratio problem in the

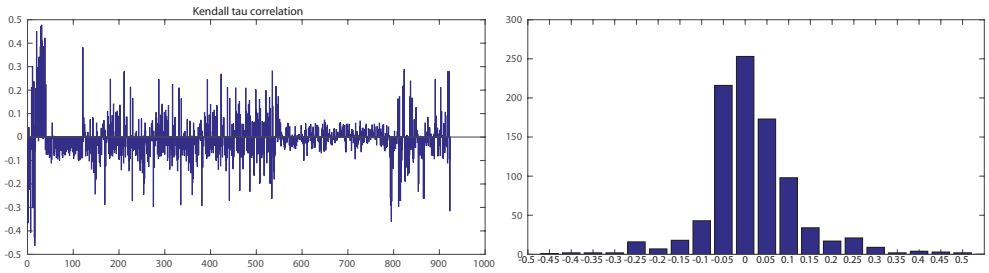

Figure 3: Kendall tau correlations of each feature dimension with the ground truth interestingness value. (left) X-axis: each dimension; Y-axis: Correlation values; (right): histogram of the correlation for different features.

attribute	pleasant_scene	attractive	memorable	is_aesthetic	is_interesting	on_post-card	buy_painting	hang_on_wall
corr	-0.4060	-0.4273	-0.4618	0.4487	0.4715	0.4767	0.4085	0.4209

Table 2: The pruned attribute features.

project webpage at http://www.eecs.qmul.ac.uk/~yf300/ranking/index.html.

10. Page 9, Col. 2, Line 52: The authors talk about global image features (GIST), but Page 8, Line 45 indicates that the ground truth annotations such as "central object", etc. were used. Using the complete ground truth annotation seems to be problematic, as it also contains an attribute "is interesting" and others such as "is aesthetic" and "is unusual". When using this ground truth, I believe such labels should be excluded and only content attributes used. (such as: indooroutdoor, contains a person, etc.).

Thanks for the suggestion. We have updated this experiment as suggested. Specifically, we first examined how each of the 932 attribute features are correlated to the groundtruth interestness value of each image. Figure 3. shows that (1) only small number of these attribute features have strong correlation with the interestingness value. (2) the histogram of kendall tau correlations[2] *of all features is roughly Gaussian as shown in Fig. 3(right).*

So as suggested, for more fair comparisons, we remove the attribute features [12] whose kendall tau correlations are higher than 0.4 or lower than -0.4. This will lead to deletion the features listed in Table 2. These pruned features include those suggested by Reviewer

[2] *Note that here, we employ kendall tau correlation rather than the Spearman correlation (Spearman correlation of "is interesting" vs. groundtruth is 0.63 as reported in [10]) since Spearman correlation is much more sensitive to error and discrepancies in data and Kendall tau correlation [11] generally have better statistical properties.*

3 ("is_interesting" and "is aesthetic"). but not the "unusual" attribute feature which has a low correlation value of -0.0226.

We repeat the image interestingness experiments with the updated features. It is noticed that this has little effect on the results (still within the variances).

References

[1] *Relevance Ranking for Vertical Search Engines*, 2014.

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

[3] T.-K. Huang, R. C. Weng, and C.-J. Lin, "Generalized bradley-terry models and multi-class probability estimates," *The Journal of Machine Learning Research*, vol. 7, pp. 85–115, 2006.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, 2004.

[5] R. Bradley and M. Terry, "Rank analysis of incomplete block designs, I. the method of paired comparisons," 1952.

[6] R. Hunter, "Mm algorithms for generalized bradley-terry models," *The Annals of Statistics*, vol. 32, p. 2004, 2004.

[7] F. Caron and A. Doucet, "Efficient bayesian inference for generalized bradley-terry models," 2012.

[8] Y. She and A. B. Owen, "Outlier detection using nonconvex penalized regression," *Journal of American Statistical Association*, 2011.

[9] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Information theory*, 2007.

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

[11] M.G.Kendall and J.D.Gibbons, *Rank correlation methods*. Ox, 1990.

[12] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," in *NIPS*, 2011.