Image ranking in video sequences using pairwise image comparisons and temporal smoothing

Michael Burke
Mobile Intelligent Autonomous Systems
Modelling and Digital Sciences
Council for Scientific and Industrial Research
South Africa
Email: michaelburke@ieee.org

***Abstract*—The ability to predict the importance of an image is highly desirable in computer vision. This work introduces an image ranking scheme suitable for use in video or image sequences. Pairwise image comparisons are used to determine image 'interest' values within a standard Bayesian ranking framework, and a Rauch-Tung-Striebel smoother is used to improve these interest scores. Results show that the training data requirements typically associated with pairwise ranking systems are dramatically reduced by incorporating temporal smoothness constraints. Experiments on a coastal image dataset show that smoothed pairwise ranking can provide ranking results equivalent to standard pairwise ranking with less than half the training data.**

I. Background and Related Work

Video cameras are increasingly deployed in exploration, monitoring and surveillance applications. These cameras produce vast amounts of information, which needs to be condensed into manageable quantities for both storage and human-operator evaluation. While data compression can address the former, this does not aid operators, who are often faced with the daunting task of analysing lengthy video sequences. As a result, a system that automatically flags interesting images or information and presents this to an operator in a concise manner is highly desirable.

In the case of video or image sequences, a mechanism by which only interesting information is stored would not only help to remedy data storage challenges, but be particularly useful in reducing the workload of data end-users, if useful summaries or storyboards of the information obtained could be provided. This is particularly challenging though, in part due to the subjective nature of the term 'interesting'.

It can be hard to define 'interesting' images, as this is typically context dependent. A study investigating the feasibility of classifying images by scientific value to address bandwidth constraints on a Mars rover [1] has shown that domain experts from different fields value and rank images differently. Information theoretic approaches to novelty detection have been proposed previously [2], but these are typically measure and data dependent. For example, ranking images using entropy is unlikely to flag images of interest to humans, as images with high texture content will always have larger information content than images with only a single centred object, yet it is highly likely that the latter is more useful to an operator.

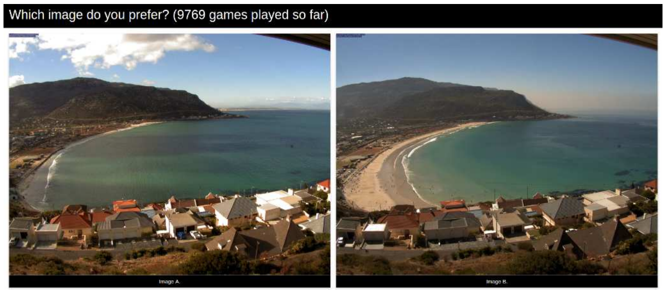

Which image do you prefer? (9769 games played so far)

Fig. 1. A pairwise comparison website is used to source image comparisons suitable for use in a Bayesian ranking system. For the sediment transfer studies example used in this paper, the right image is preferable, because regions of wet and dry sand are more easily distinguishable than those in the left image.

Despite this difficulty and the potentially subjective definition of interest, a great deal of work has been conducted in an attempt to rank the value of information. A common definition of interest relates to novelty, with the frequency of occurrence of an event or observation determining its interest value. Novelty detection is relatively well studied and detailed survey papers can be found on the topic [2].

However, information of interest to an end-user not only includes unique observations (novelty), but also observations that are most representative of the process or environment observed. For example, film makers would probably prefer a storyboard summary of a film to a collection of unique frames sampled from it. In the video domain case, end-users may value images of objects, people or animals more than an image of an unusually shaped cloud (unless of course, they were meteorologists). In addition, users may prefer good quality images over a blurry or overexposed image, but the detection of suspicious activity in a surveillance camera probably outweighs a user's desire for image quality.

The subjective nature of image ranking means that it is unlikely that an image interest detection algorithm can be designed from the bottom up, potentially incorporating image metrics for novelty, representativeness, aesthetic aspects and so on. Instead, a far more sensible approach would be to let operators in a specific domain label a selection of images that they find interesting and have a system that uses this to learn about the task at hand.

This is equally challenging though, as the subjective nature of interest makes it hard to design a labelling mechanism suitable for capturing the intricacies of interest, short of arranging focus and discussion groups, which are unlikely to produce data in the volumes required for machine learning.

In an attempt to remedy this, crowd-sourcing systems that use relative image comparisons to infer user preference have been developed [3]. Here, pairwise image comparisons are used to rank images according to user preference. Pairwise ranking systems are often used for image ranking tasks because they can provide more stable and useful rankings than individual image-based scoring systems [4].

Pairwise ranking systems use binary comparison test results to infer an underlying rank and have been applied to a wide range of applications, including recommender systems [5], software simulation component evaluation [6], in sport [7], online gaming [8] and advertising [9].

An obvious approach to ranking using pairwise comparisons is to simply count the number of victories obtained by each compared item. Unfortunately, this ignores information about which items were compared with one another (a number of wins against an exceptionally poor opponent does not necessarily mean a player is skilled) and may fail to account for performance variability. Ranking systems that account for these factors include the Elo chess rating system [10] and TrueSkill$^{\text{TM}}$ [8], a Bayesian ranking scheme extension to Elo.

Pairwise comparisons are frequently used for image ranking tasks. For example, CollaboRank [11] uses pairwise comparisons to rank images according to a number of case-based queries (positiveness, perceived threat level, celebrity or film popularity), the Matchin approach [3] uses a two player pairwise comparison game to extract a global image 'beauty' rank and Streetscore [12] predicts the perceived safety of street scenes using binary answers to the question "Which place looks safer?"

Pairwise comparisons have also been used to rank abstract paintings according to the emotional responses they elicit [13], to evaluate the representativeness of images extracted from twitter timelines [14], and to determine appropriate facial expressions for portraits using images extracted from short video sequences [15]. Hipster wars [4] uses a pairwise comparison game to source style judgements to train an image-based style classifier in a fashion application. Unfortunately, the crowd-sourcing process used to obtain pairwise comparison results can be time consuming and expensive [16].

This work shows how the training data requirements for ranking can be reduced in ranking tasks where images to be compared are sampled from video or image sequences. This is often the case in exploration, monitoring and surveillance applications. This paper shows how the sequential nature of video sequences allows the addition of a temporal smoothing step to the traditional ranking process. Once ranked, additional rank prediction algorithms can be developed to identify task specific image features that are of importance to end users, and predict the interest value of previously unseen images.

The proposed approach is illustrated using a coastal monitoring application, where pairwise image comparisons (Figure 1) are used within a Bayesian ranking scheme to infer interest values for images in the corpus, and these interest values smoothed temporally using a Kalman smoother. Results show that this dramatically reduces the training data requirements to predict image rank.

The remainder of this paper is organised as follows. Section II describes TrueSkill$^{\text{TM}}$ and Kalman smoothing, showing how these techniques can be applied to image ranking in video sequences. Section III introduces an image ranking task in the coastal science domain, and presents experimental results obtained when applying the proposed ranking system to this dataset. Finally, conclusions and recommendations for future work are provided in Section IV.

II. Method

The proposed approach to image ranking combines a standard pairwise ranking scheme and Kalman smoothing. These subcomponents are briefly described below.

A. Pairwise ranking using TrueSkill$^{\text{TM}}$

As a baseline, this work uses the TrueSkill$^{\text{TM}}$ Bayesian ranking scheme [8] to compute image interest scores.

TrueSkill$^{\text{TM}}$ is a probabilistic skill rating system developed for online gaming that assumes players in a game have respective skills, w_1 and w_2, and that game outcomes can be predicted by the performance difference between skills, subject to Gaussian noise effects.

Let

$$t \sim \mathcal{N}(s, 1) \tag{1}$$

denote the performance difference between two players, with $s = w_1 - w_2$ the skill difference and the standard normal distribution accounting for potential player inconsistency. Using this model, game outcomes are given by $y = \text{sign}(t)$, with a positive y indicating a win for player 1, and a negative y indicating a loss.

Treating skill estimation under this model as a Bayesian inference problem provides a posterior over skills,

$$p(w_1, w_2|y) = \frac{p(w_1)p(w_2)p(y|w_1, w_2)}{\int\int p(w_1)p(w_2)p(y|w_1, w_2)\mathrm{d}w_1\mathrm{d}w_2}, \tag{2}$$

where $p(w_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ is a Gaussian prior over player skills and

$$p(y|w_1, w_2) = \int\int p(y|t)p(t|s)p(s|w_1, w_2)\mathrm{d}s\mathrm{d}t \tag{3}$$

the likelihood of a game outcome given skills. The model above is easily extended to multiple players by chaining games together in a large graph. This is illustrated graphically in the factor graph of Figure 2.

Equation (2) is an intractable posterior, but can be estimated numerically. Expectation propagation [17] is used for inference in the original formulation.

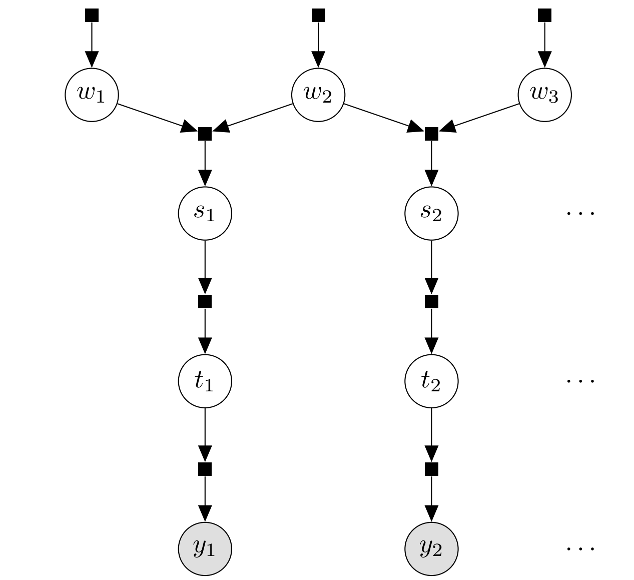


Fig. 2. The TrueSkill$^{\text{TM}}$ factor graph is extended each time a game is played, adding a connection between players.

B. Temporal Smoothing

Temporal smoothing refers to the process whereby knowledge of the temporal behaviour of a state is used in conjunction with a set of noisy measurements to produce an improved estimate of the underlying state.

Let $z_{1:K}$ denote a set of noisy measurements obtained at time steps $1 \ldots K$, and assume that our goal is to estimate an underlying state, x_k. If the state change is Markovian, we can use knowledge of the transition density, $p(x_k|x_{k-1})$, together with an observation model, $p(z_k|x_k)$, within a sequential Bayesian smoothing framework to provide a posterior distribution over the state, $p(x_k|z_{1:K})$, conditioned on the sequence of measurements.

Smoothing problems of this form can be solved by combining a sequential filtering operation with a backward smoothing stage [18]. Filtering estimates the density over the state, x_k, conditioned on measurements, $z_{1:k}$, by combining a prediction step,

$$p(x_k|z_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1})\mathrm{d}x_{k-1}, \quad (4)$$

with an update stage,

$$p(x_k|z_{1:k}) = \frac{p(z_k|x_k)p(x_k|z_{1:k-1})}{\int p(z_k|x_k)p(x_k|z_{1:k-1})\mathrm{d}x_k}. \quad (5)$$

The Kalman filter [19] provides an analytical solution to problems of this form when prior, transition and observation densities are Gaussian. Kalman filtering only considers historical measurements (up to time step k). Information about future observations is incorporated using backward pass to provide a density over the smoothed state,

$$p(x_k|z_{1:K}) = \int \frac{p(x_{k+1}|x_k)p(x_k|z_{1:k})}{p(x_{k+1}|z_{1:k})} p(x_{k+1}|z_{1:K})\mathrm{d}x_{k+1}. \quad (6)$$

The Rauch-Tung-Striebel (RTS) smoothing recursions [20] provide an analytical solution to smoothing problems when prior, transition and observation densities are Gaussian.

C. Image ranking in video sequences

This section shows how TrueSkill$^{\text{TM}}$ and RTS smoothing can be used for image ranking. Although developed for online gaming, TrueSkill$^{\text{TM}}$ is directly applicable to image ranking using pairwise comparisons.

Here, games are image comparisons presented to a human labeller, and game winners are the images selected as preferable in each comparison. The inferred skills can be considered to be image 'interest' scores, with images of greater interest to a user more likely to be preferred in pairwise comparisons.

TrueSkill$^{\text{TM}}$ makes no assumptions about the underlying process producing images, treating images independently for the purposes of scoring. However, in many applications, the images to be ranked are captured in sequences or video.

As a result, temporal interest consistency is to be expected in a sequence, as interest scores for subsequent images are unlikely to change significantly. The posterior interest scores inferred using the TrueSkill$^{\text{TM}}$ algorithm are normally distributed, parametrised by a mean interest, z_k and corresponding uncertainty, R_k, and consequently perfectly suited for fixed interval smoothing.

Modelling the change in image interest over subsequent images in a sequence as a Gaussian random walk,

$$p(w_k|w_{k-1}) = \mathcal{N}(w_k|w_{k-1}, Q), \quad (7)$$

where Q is a tunable transition uncertainty ($Q = 5\mathrm{e}{-5}$ gave good results in experiments), allows for fixed interval smoothing using a Rauch-Tung-Striebel (RTS) smoother [20].

Here, the goal is to find the posterior density over image interest, conditioned on all image interest measurements in a sequence, $p(w_k|z_{1:K})$. Initially, RTS smoothing uses a Kalman filter [19] forward pass step to calculate

$$p(w_k|z_{1:k}) = \mathcal{N}(m_k, P_k). \quad (8)$$

For the Gaussian random walk used here, the simplified recursive Kalman filter update equations

$$\begin{aligned} \hat{m}_k &= m_{k-1}, & (9)\\ \hat{P}_k &= P_{k-1} + Q, & (10)\\ m_k &= \hat{m}_k + \hat{P}_k(\hat{P}_k + R_k)^{-1}(y_k - \hat{m}_k), & (11)\\ P_k &= \hat{P}_k - \hat{P}_k(\hat{P}_k + R_k)^{-1}\hat{P}_k, & (12) \end{aligned}$$

are used to find filtered means and variances, while the RTS backward pass recursions are

$$\begin{aligned} \tilde{m}_k &= m_k + P_k\hat{P}_k^{-1}(\tilde{m}_{k+1} - \hat{m}_{k+1}), & (13)\\ \tilde{P}_k &= P_k + P_k\hat{P}_k^{-1}(\tilde{P}_{k+1} - \hat{P}_{k+1})P_k\hat{P}_k^{-1}, & (14) \end{aligned}$$

resulting in the posterior over image interest values in the sequence,

$$p(w_k|z_{1:K}) = \mathcal{N}(\tilde{m}_k, \tilde{P}_k), \quad (15)$$

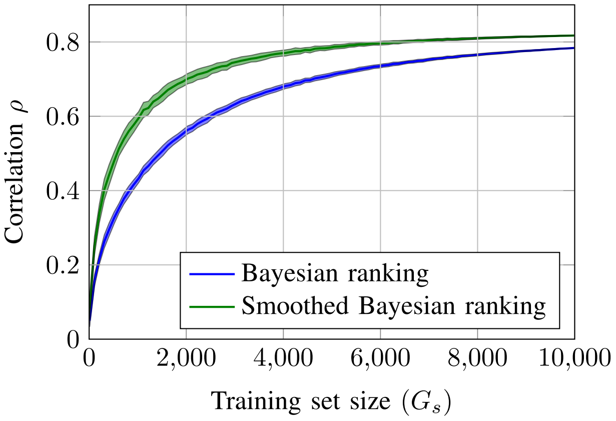


Fig. 3. Interest score correlation with the baseline image interest predictions grows as the training set increases in size. Shaded traces indicate 1-sigma curves.

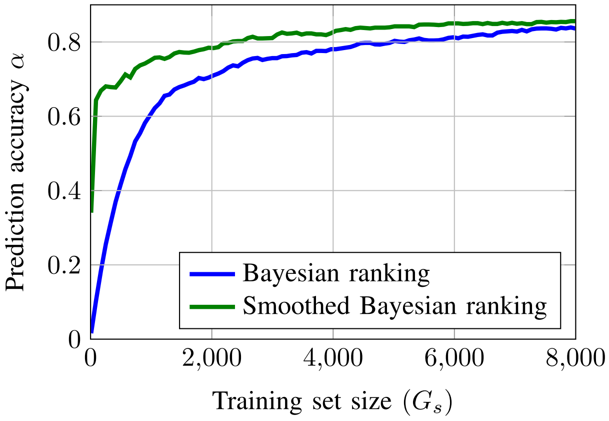


Fig. 4. The prediction accuracy obtained for each method increases as the training set size increases. Accuracy was tested on 2000 randomly selected comparisons held out from the labelled corpus (G).

conditioned on the TrueSkill$^{\text{TM}}$ estimates. This smoothed density incorporates the temporal interest consistency likely to be present in video, thereby providing an improved interest estimate. The filter is initialised with large uncertainty, $P_0 = 10^9$ and $m_0 = z_1$.

III. An experimental case study

The proposed approach to image interest ranking was tested on a sequence of 1900 images captured in Fish Hoek, Cape Town. These images were captured for a sediment transfer study, where the goal was to segment and label image areas as either wet or dry sand, so as to study the motion of sand over time.

Unfortunately, this segmentation process is challenging, as it can only occur when certain conditions are met: the image is captured in daylight; the tide must low enough for sand regions to be visible; wave swash needs to be in the backwash stage and glare should not interfere with the regions to be labelled. As a result, an automated input stage that ranks images according to their suitability for sand segmentation is desired.

As a baseline, 10000 image pairwise comparisons, G, were performed by a domain expert[1] using the web interface shown in Figure 1. These comparisons were augmented by adding all possible (774336) day/night image combinations, G_{dn}, in the dataset to produce a baseline comparison set, G_a. Images captured at night are easily detected, and guaranteed to be less important than daylight images, so provide a useful mechanism to introduce connections into the TrueSkill$^{\text{TM}}$ factor graph, thereby reducing the uncertainty in image interest estimates.

Figure 3 shows the Pearson correlation coefficient, ρ, calculated between baseline image interest means estimated using G_a and those estimated with an increasing training set size. Image interest scores were calculated by selecting a subset of comparisons, G_s, at random without replacement from the full set of 10000 pairwise comparisons (G) and applying TrueSkill$^{\text{TM}}$ Bayesian ranking with and without smoothing. Experiments were repeated 100 times for 100 increasing subset sizes to produce the shaded error trace in Figure 3. It is important to note that the correlation after 10000 training comparisons is not unity, because the baseline skills were calculated using the augmented comparison set (G_a), which includes day/night comparisons.

It is clear that smoothing dramatically reduces the training set size required, with smoothed ranking only requiring 3743 comparisons to achieve results equivalent to those obtained using TrueSkill$^{\text{TM}}$ and 10000 comparisons. This can also be observed in Figure 4, which shows the prediction accuracy,

$$\alpha = \frac{\text{Number correct predictions}}{\text{Total number comparisons}}, \tag{16}$$

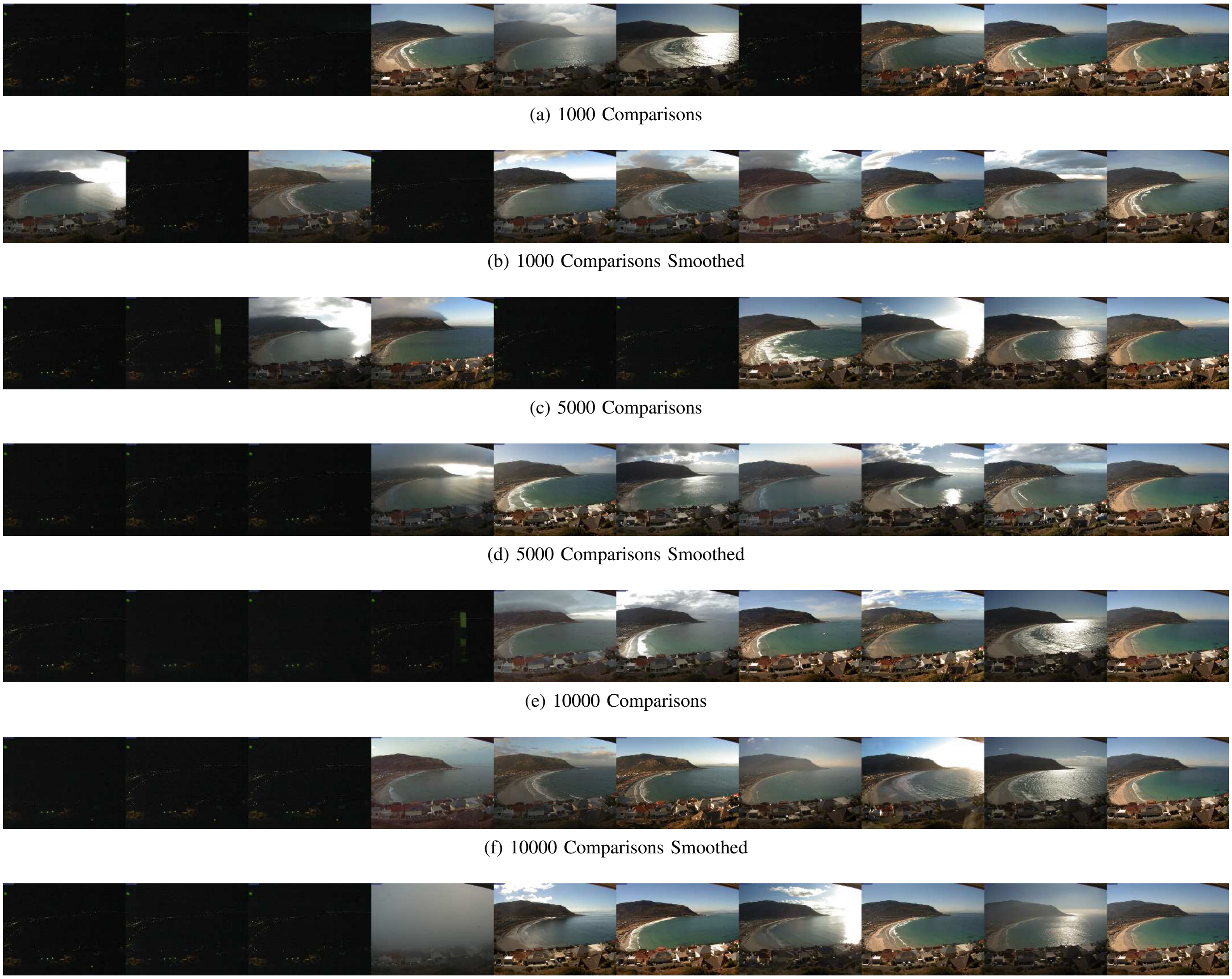
obtained when 2000 previously unseen comparison results (selected at random from G) are predicted using the posterior interest means obtained for TrueSkill$^{\text{TM}}$ and smoothed ranking. Here, only 4368 comparisons are required to obtain equivalent results to standard Bayesian ranking on the full corpus.

As a sanity check on the proposed image ranking scheme, Figure 5 shows images (10 evenly spaced samples) in rank order as the training set increases in size. The rank becomes more reliable as the training set increases in size, with night images that were initially ranked above day images, sorted correctly. This occurs quite rapidly when smoothing is applied. After including a suitable amount of training data, higher ranked images clearly exhibit properties of interest for the sand segmentation task: images are captured in daylight; the tide is quite low; and the best images show clear discrepancies between wet and dry sand.

IV. Conclusions

This work has shown how the sequential nature of image sequences allows the incorporation of a temporal smoothing step into a standard Bayesian ranking framework, which reduces training data requirements significantly. As a result, the

[1] The unfortunate author of this paper, who had to label 10000 image pairs, likes to consider himself a domain expert.



(a) 1000 Comparisons

(b) 1000 Comparisons Smoothed

(c) 5000 Comparisons

(d) 5000 Comparisons Smoothed

(e) 10000 Comparisons

(f) 10000 Comparisons Smoothed

(g) Baseline

Fig. 5. The figure shows images drawn from the test corpus, and ranked using an increasing number of training samples. Images on the left are of low interest, while images on the right are considered more important. The baseline images were ranked by applying TrueSkill^{TM} using the augmented set of comparisons (G_a).

proposed interest scoring process has the potential to produce a substantial amount of training data for follow on interest detection algorithms, from only a limited amount of hand-labelled training data.

In future work, image interest scores will be used to train a predictive model of image interest, and to identify domain-specific image features that elicit human interest.

References

[1] R. Castano, K. Wagstaff, L. Song, and R. Anderson, "Validating rover image prioritizations," *The Interplanetary Network Progress Report*, vol. 42, p. 160, 2005.

[2] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[3] S. Hacker and L. von Ahn, "Matchin: Eliciting user preferences with an online game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 1207–1216.

[4] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *European conference on computer vision*. Springer, 2014, pp. 472–488.

[5] S. Balakrishnan and S. Chopra, "Two of a kind or the ratings game? adaptive pairwise preferences and latent factor models," in *2010 IEEE International Conference on Data Mining*, Dec 2010, pp. 725–730.

[6] J. Wienß, M. Stein, and R. Ewald, "Evaluating simulation software components with player rating systems," in *Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013, pp. 41–50.

[7] S. Motegi and N. Masuda, "A network-based dynamical ranking system for competitive sports," *Scientific Reports*, vol. 2, p. 904, 2012.

[8] R. Herbrich, T. Minka, and T. Graepel, "Trueskill^{TM}: A Bayesian skill rating system," in *Advances in neural information processing systems*, 2006, pp. 569–576.

[9] D. H. Stern, R. Herbrich, and T. Graepel, "Matchbox: large scale online Bayesian recommendations," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 111–120.

[10] A. E. Elo, *The rating of chessplayers, past and present*. Arco Pub., 1978.

[11] J. H. Janssens, “Ranking images on semantic attributes using human computation,” in *NIPS workshop on computational social science and the Wisdom of crowds*, 2010.

[12] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, “Streetscore – predicting the perceived safety of one million streetscapes,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2014, pp. 793–799.

[13] A. Sartori, “Affective analysis of abstract paintings using statistical analysis and art theory,” in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 384–388.

[14] C.-L. Wen *et al.*, “Event-centric twitter photo summarization,” Master’s thesis, Massachusetts Institute of Technology, 2014.

[15] J.-Y. Zhu, A. Agarwala, A. A. Efros, E. Shechtman, and J. Wang, “Mirror mirror: Crowdsourcing better portraits,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 234, 2014.

[16] (2016, August) Dynamo guidelines for academic requesters on Amazon Mechanical Turk. Online. [Online]. Available: http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters

[17] T. P. Minka, “A family of algorithms for approximate Bayesian inference,” Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[18] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, vol. 3.

[19] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[20] H. E. Rauch, C. Striebel, and F. Tung, “Maximum likelihood estimates of linear dynamic systems,” *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.