Ranking Images on Semantic Attributes using Human Computation

Jeroen H.M. Janssens
Tilburg center for Cognition and Communication, Tilburg University
PO Box 90153, 5000 LE Tilburg, The Netherlands
jeroen@jeroenjanssens.com

Abstract

We investigate to what extent a large group of human workers is able to produce collaboratively a global ranking of images, based on a single semantic attribute. To this end, we have developed CollaboRank, which is a method that formulates and distributes tasks to human workers, and aggregates their personal rankings into a global ranking. Our results show that human workers can achieve a relatively high consensus, depending on the type of the semantic attribute.

1 Semantic Ranking Problem

With the Internet being the largest, and fastest growing image database available, responding adequately to a user query remains a constant challenge. Although the precision of the responses has already improved substantially over the past few years, image search may be further improved by ranking the images in a sensible and understandable way. An attractive solution is to rank the images according to their contents, i.e., semantics, on a single attribute. For instance, when searching for an image of an expensive car, ranking images on the semantic attribute "price" would facilitate the search considerably.

We define a semantic ranking problem (SRP) as the problem of obtaining a ranking of images, belonging to the same class, based on a single semantic attribute. Jörgensen identified three types of semantic attributes: perceptual, interpretive, and reactive [5]. Perceptual attributes are directly related to a visual stimulus (e.g., color, shape). Interpretive attributes require both interpretation of perceptual cues and a general level of knowledge or inference from that knowledge (e.g., the artist of a painting). Reactive attributes describe a personal response or emotion (e.g., the attractiveness of a face). While ranking images on perceptual attributes is usually trivial for a computer, the latter two types are more challenging, or even impossible. In the domain of Content Based Image Retrieval this is called the semantic gap [7]. We propose that the semantic gap may be bridged by employing a large group of human workers, i.e., human computation.

To investigate to what extent human computation can help solving an SRP, we have developed a method called CollaboRank. CollaboRank may be compared to Matchin [3], in the sense that both methods obtain a ranking of images with the help of human workers, or players. Matchin is an online game that shows two images to two players. Both players should click on the image that they think the other player prefers. When both players choose the same image, it is concluded that that image is more beautiful than the other. A single global ranking of "beautifulness" is computed using TrueSkill [4]. CollaboRank differs from Matchin because each task concerns a well-defined attribute that relates to the semantics of the image, and not the image itself. This allows us to assume a higher level of transitivity in the global ranking, and therefore allows us to distribute tasks containing more than two images. This also influences how CollaboRank computes a global ranking.

The remainder of the paper is as follows. Section 2 describes CollaboRank. Experiments and results are presented in Section 3. Section 4 presents our conclusion and gives directions for future research.

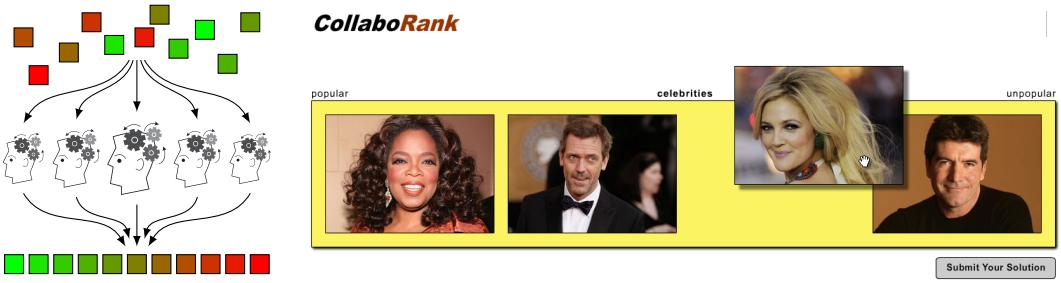

Figure 1: Left: A global ranking of images may be obtained through human computation by distributing small tasks to human workers and subsequently aggregating their personal rankings. Right: A screenshot of a CollaboRank task containing images of the class "celebrities", that should, in this case, be ranked on the semantic attribute "popularity". Human workers perform the task by re-ordering the images, and pressing the submit button to return their personal ranking.

2 CollaboRank

The CollaboRank method enables human workers to rank collaboratively a large set of images I, that share a common class c, on a semantic attribute α and to produce a global ranking $\mathcal{R}^g$. A central element is the global preference matrix $\mathbf{P}^g$, which is used for the three steps of CollaboRank: (1) formulating tasks for human workers, (2) aggregating their personal rankings, and (3) computing a global ranking $\mathcal{R}^g$. There exists one $\mathbf{P}^g$ per class-attribute combination. It represents the image preference relation, i.e., the relative order of each image pair $\langle i, j \rangle$. The preference of image i_i over image i_j is denoted by p^g_{ij}, where $p^g_{ij} = \frac{1}{2}$ indicates equal preference for i_i and i_j ($i_i \sim i_j$); $p^g_{ij} > \frac{1}{2}$ indicates that i_i is preferred to i_j ($i_i \succ i_j$); and $p^g_{ij} < \frac{1}{2}$ indicates the reverse. A preference relation is a useful concept for modeling decision processes, and therefore also for aggregating personal rankings into a global ranking $\mathcal{R}^g$ [1].

A task is formulated by selecting θ images whose position in the ranking is least certain. We first select an image m for which the summed entropy of all the image pairs $\langle m, \cdot \rangle$ is highest. Subsequently we select $\theta - 1$ times an image k that has the highest entropy of p^g_{mk}, i.e., image pair $\langle m, k \rangle$. Figure 1 shows a screenshot of CollaboRank illustrating how a task may be performed by drag-and-dropping the images in the desired order.

Aggregation of a personal ranking submitted by a human worker is done as follows. First, the personal ranking $\mathcal{R}^p$ is transformed into a personal preference matrix $\mathbf{P}^p$, which contains the (binary) pairwise preferences of the images contained in the task. Second, assuming that each human worker is weighted the same, we aggregate the personal ranking matrix into the global preference matrix by assigning the average preference of images i_i and i_j, with regard to all the submitted personal rankings, to p^g_{ij}.

When sufficient tasks have been performed, a global ranking of images can be computed. With the Matchin game [3], it is assumed that computing a global ranking of images is the same as computing the skills of chess players, which is why TrueSkill[4] is used. However, as skill implies practice, images should compete continuously to maintain their relative rank (i.e., once $(i_a \succ i_b)$, and later on $(i_b \succ i_c)$, $(i_a \succ i_b)$ may not hold anymore). Since we can assume a higher level of transitivity, due to more specific tasks, we instead adopt the Greedy-Order algorithm for the computation of a global ranking [2]. The algorithm can be best described by interpreting the global preference matrix as a directed weighted graph, where initially, the set of vertices V is equal to the set of images I, and each edge $u \to v$ has weight p^g_{uv}. Each vertex $v \in V$ is assigned a potential value $\pi(v)$, which is the weighted sum of the outgoing edges minus the weighted sum of the ingoing edges. That is, $\pi(v) = \sum_{u \in V} p^g_{vu} - \sum_{u \in V} p^g_{uv}$. The algorithm then chooses the vertex t with the maximum potential. The vertex, and thus the corresponding image, is assigned the rank $\mathcal{R}^g(t) = |V|$, making it appear first in the ranking. The vertex and its edges are deleted from the graph, and the potential value π of the remaining vertices are updated. This process is repeated until the graph is empty.

Table 1: Results of a CollaboRank experiment with 21 human workers for 20 minutes. Number of personal rankings is denoted by $\#\mathcal{R}^p$. Baseline consensus is 0.212.

Class	Attribute	Type	$\#\mathcal{R}^p$	Time (sec.)	Consensus
Emotions	positiveness	reactive	295	11.3 ± 7.9	0.745
Surveillance video's	threat level	inter. / react.	284	17.0 ± 9.9	0.628
Textures	smoothness	perceptive	280	10.9 ± 8.2	0.591
Movies	popularity	interpretive	289	11.6 ± 8.4	0.556
Celebrities	popularity	reactive	300	12.9 ± 9.6	0.514

3 Experiments and Results

To evaluate CollaboRank, and thus to answer our question whether human computation aids in solving an SRP, we conducted an experiment with 21 male and female human workers aged 20 – 25. In order to have better control over the environment, and to be able to receive feedback from the human workers, we decided to conduct the experiment in a class room instead of using the Internet (e.g., Amazon's Mechanical Turk or as a game). In total we used five different image classes, 50 images each, that had to be ranked on different semantic attributes: (1) faces from Japanese women[1] on positiveness of emotion, (2) stills from surveillance camera's[2,3,4] ranked on threat level, (3) textures from different materials[5] on smoothness of the surface, (4) movies[6] on popularity, and (5) celebrities[7] on popularity. These semantic attributes vary in difficulty as they belong to the perceptual, interpretive, or reactive type. The human workers were instructed to rank the images given a specific class and attribute, where each class was addressed in turn.

We define consensus among human workers as the average Kendall's tau rank correlation coefficient [6] between all submitted personal ranking and their corresponding intersections of the global ranking. It should be noted that although originally this coefficient may range from -1 (complete disagreement) to $+1$ (complete agreement), consensus ranges from 0 to $+1$. This is due to the fact that the computation of the global ranking is based on the personal rankings. The baseline consensus is what a group of random rankers on average achieves. The baseline increases (i.e., obtaining a higher consensus becomes easier) with either a larger image set size, a smaller task size, or a smaller number of submitted personal rankings. In our case, i.e., 50 images, 4 images per task, and 300 submitted personal rankings, the baseline consensus is 0.212.

After only 20 minutes, the 21 human workers had completed 1,448 tasks in total, which corresponds to 8,688 pairwise comparisons. Table 1 shows for each of the five classes and their attributes, the attribute type, the number of personal rankings, the average time in seconds it took a human worker to complete a task, and the consensus among human workers. Figure 2 shows from each class the top ranked and the bottom ranked image.

Human workers achieved the highest consensus for ranking emotions on positiveness, although it was not always clear to them whether a face expressing "surprise" should be evaluated positive or negative. Tasks of the surveillance video's class took on average longest to complete, which may be explained by the fact that normally: (1) these are not images but video's and (2) these are evaluated by trained security guards. Nevertheless, a relatively high consensus was achieved for this class. The relatively low consensus for the textures class was unexpected, because smoothness is an objective attribute. It is possible that, for some human workers, the perceived smoothness of a material did not coincide with its actual smoothness. The least consensus was achieved for movies and celebrities, which indicates that "popularity" is a subjective and ambiguous attribute. As a case in point, after the experiment, human workers had a discussion whether popularity should be interpreted as well-liked or as well-known.

[1] Japanese Female Facial Expression database: `http://www.kasrl.org/jaffe.html`
[2] i-LIDS bag detection: `http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html`
[3] PETS 2006 Benchmark Data: `http://www.cvg.rdg.ac.uk/PETS2006/data.html`
[4] BOSS : On Board Wireless Secured Video Surveillance: `http://www.celtic-boss.org`
[5] CG Textures: `http://www.cgtextures.com`
[6] Internet Movie Database: `http://www.imdb.com`
[7] Forbes Celebrity Top 100: `http://www.forbes.com/celebs`

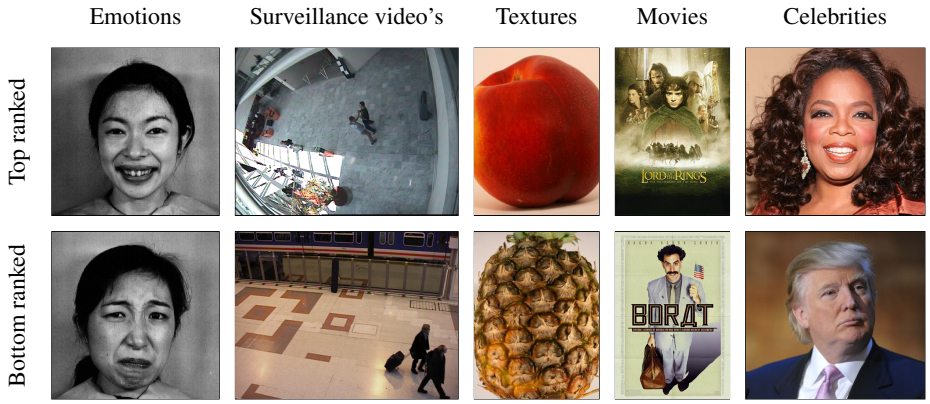

Figure 2: The top and bottom ranked images of the five classes. Please note that the textures class also contained images of other materials, such as wood, metal, and stone.

4 Conclusion and Future Research

Since human workers are able to extract the full semantic contents of an image, they are able to rank a small subset of images. However, distributed human computation alone cannot solve an SRP. A method is required that enables human workers to perform collaboratively global ranking tasks. We demonstrated that CollaboRank effectively fulfills this requirement, through formulating and distributing tasks to the human workers, and aggregating the personal rankings into a global ranking. From our obtained results, we may conclude that human computation can help solving an SRP with a relatively high consensus, depending on the type of the semantic attribute.

We suggest two directions in which future research can develop. The first direction could be taking a more Bayesian approach to the formulation of tasks, and the aggregation of personal rankings. Perhaps the TrueSkill algorithm could be extended such that a higher level of transitivity can be assumed, making it more suitable for CollaboRank. The second direction could be transforming CollaboRank into a game, in order to stimulate participation. As with Matchin, CollaboRank could be a two-player game, where both players are shown the same images, and where the reward could be based on the correlation of both personal rankings. Alternatively, one player could first rank the images, whereafter the second player should guess the appropriate semantic attribute.

References

[1] F. Chiclana, F. Herrera, and E. Herrera-Viedma. Preference relations as the information representation base in multi-person decision making. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 459–464, 1996.

[2] W.W. Cohen, R.E. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

[3] S. Hacker and L. von Ahn. Matchin: Eliciting user preferences with an online game. In *CHI '09: Proceedings of the 27th International Conference on Human Factors in Computing Systems*, pages 1207–1216, New York, NY, USA, 2009. ACM.

[4] R. Herbrich, T. Minka, and T. Graepel. Trueskill$^{\mathrm{TM}}$: A Bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, Cambridge, MA, 2007.

[5] C. Jörgensen. Attributes of images in describing tasks. *Information Processing & Management*, 34(2-3):161 – 174, 1998.

[6] M.G. Kendall. *Rank correlation methods*. London, Charles Griffin & Company, Limited, 1962.

[7] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.