# Implied Feedback: Learning Nuances of User Behavior in Image Search (Supplementary Material)

Devi Parikh
Virginia Tech
parikh@vt.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

This document contains additional implementation details of the image features, the list of attributes, the instructions provided to subjects in our user studies, the procedure for training the relative attributes and our proposed features that capture implicit tendencies of users. Note that the main paper is comprehensive and self contained. The following are supplemental details for interested readers.

## 1. Details of Image Features

We used the following low-level features to describe images.

- Scenes: Images are described by a 512 dimensional gist descriptor.
- Faces: Images are described by a 512 dimensional gist descriptor concatenated with a 30 dimensional color histogram. The color histogram is formed by concatenating 10 dimensional histograms from each of the 3 color channels in the Lab color space.
- Shoes: Images are described by a 960 dimensional gist descriptor concatenated with a 30 dimensional color histogram as described above.

## 2. List of Attributes Per Dataset

For relative-attributes based feedback, we used the following attributes.

- Scenes: natural, open and expanding space. These are the only three attributes in [5] that we find to be reliably understood by layman users (e.g. Mechanical Turk workers).
- Faces: white, dark hair, young, chubby, masculine looking (male), pointy nose, strong nose-to-mouth-lines, round jaw, visible forehead, big lips. These are a subset of the 29 attributes with relative attribute annotations provided by [1].
- Shoes: pointy-at-the-front, open, bright-in-color, covered-with-ornaments, shiny, high-at-the-heel, long-on-the-leg, formal, sporty, feminine. We used the annotations provided in [4].

## 3. Details of Procedure for Training Relative Attribute Predictors

The relative attributes were trained using the approach in [6]. We are given a set of training images $I = \{i\}$. Each image is represented by a feature-vector $\boldsymbol{x_i} \in \mathbb{R}^n$. We are also given a vocabulary of $M$ attributes $A = \{a_m\}$. For each attribute $a_m$, we are given two sets of pairs of images. The first is a set of ordered pairs of images $O_m = \{(i, j)\}$. $(i, j) \in O_m \implies i \succ j$, i.e. image $i$ has a stronger presence of attribute $a_m$ than $j$. The second is a set of of un-ordered pairs of images $S_m = \{(i, j)\}$. $(i, j) \in S_m \implies i \sim j$, i.e. $i$ and $j$ have similar relative strengths of $a_m$. Our goal is to learn $M$ ranking functions

$$r_m(\boldsymbol{x_i}) = \boldsymbol{w_m}^T \boldsymbol{x_i}, \tag{1}$$

for $m = 1, \dots, M$, such that the maximum number of the following constraints is satisfied:

$$\forall (i, j) \in O_m : \boldsymbol{w_m}^T \boldsymbol{x_i} > \boldsymbol{w_m}^T \boldsymbol{x_j} \tag{2}$$

$$\forall (i, j) \in S_m : \boldsymbol{w_m}^T \boldsymbol{x_i} = \boldsymbol{w_m}^T \boldsymbol{x_j}. \tag{3}$$

By introducing non-negative slack variables, a "learning to rank" objective similar to SVM classification is given in [3]. Following [6], we use a quadratic loss function and incorporate the similarity constraints in the formulation of [3], leading to the following optimization problem:

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right) \tag{4}$$

$$s.t. \quad \boldsymbol{w_m^T x_i} \geq \boldsymbol{w_m^T x_j} + 1 - \xi_{ij}; \forall (i,j) \in O_m \tag{5}$$

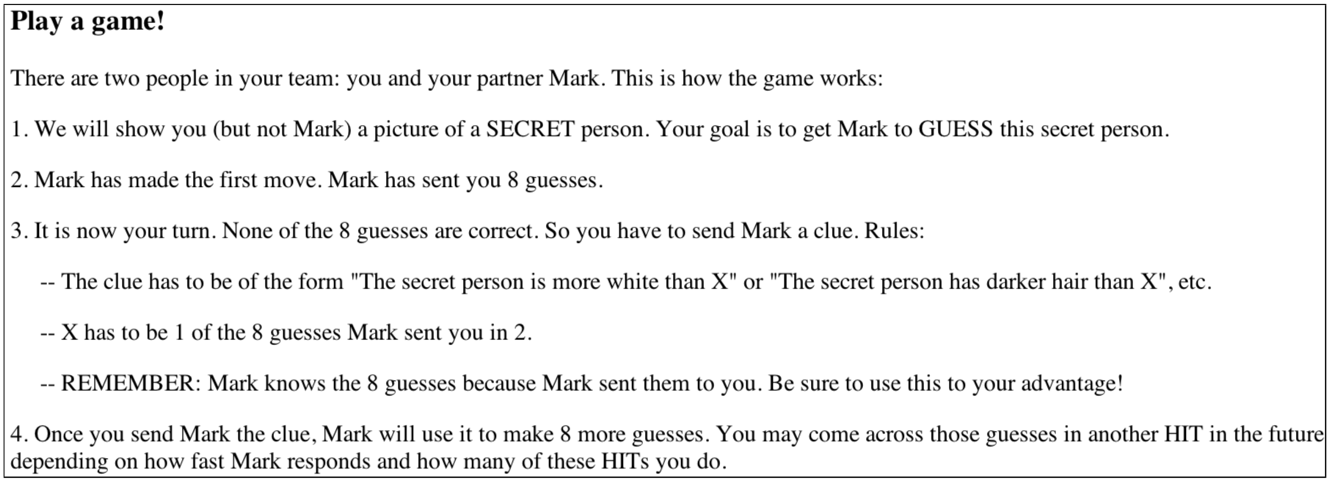
$$|\boldsymbol{w_m^T x_i} - \boldsymbol{w_m^T x_j}| \leq \gamma_{ij}; \forall (i,j) \in S_m \tag{6}$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0. \tag{7}$$

$C$ trades off maximizing the margin with satisfying the constraints. We solve the above primal problem using Newton's method [2].

## 4. Instructions Given to Subjects

As discussed in the main paper, our data collection task was disguised as a game between two players. Subjects were told that their goal is to help their partner guess what the "secret" (target) image is by giving him clues. See Figure 1. Our game-like interface is intended to bolster the realism of the data. The game aspect encourages the user to care about the quality of his response, much as he would if doing a search for his own purposes. In contrast, if he were to think he is simply participating in a data collection effort, it could dilute the very nuances in behavior that we are interested in modeling.



**Play a game!**

There are two people in your team: you and your partner Mark. This is how the game works:

1. We will show you (but not Mark) a picture of a SECRET person. Your goal is to get Mark to GUESS this secret person.

2. Mark has made the first move. Mark has sent you 8 guesses.

3. It is now your turn. None of the 8 guesses are correct. So you have to send Mark a clue. Rules:

-- The clue has to be of the form "The secret person is more white than X" or "The secret person has darker hair than X", etc.

-- X has to be 1 of the 8 guesses Mark sent you in 2.

-- REMEMBER: Mark knows the 8 guesses because Mark sent them to you. Be sure to use this to your advantage!

4. Once you send Mark the clue, Mark will use it to make 8 more guesses. You may come across those guesses in another HIT in the future depending on how fast Mark responds and how many of these HITs you do.

Figure 1: The instructions for our user studies simulate a game.

## 5. List of Features to Capture Implicit Relative Attributes-based Feedback (Section 3.3 in Paper)

The main paper presents the intuitions behind our proposed features, the mathematical form of the main features, as well as a description of their variations. Below we list out these variations explicitly.

In the following, $t_m$ is shorthand for the strength of the attribute $a_m$ in image $\boldsymbol{t}$ i.e. $r_m(\boldsymbol{t})$. Similarly, $p_m = r_m(\boldsymbol{p})$. The direction of feedback $q$ is $+1$ if the user said "more" and $-1$ if the user said "less". $\boldsymbol{p}^+$ denotes the reference image ranked consecutively to the chosen reference image $\boldsymbol{p}^*$ in the direction $q$ of the feedback. For any candidate reference image $\boldsymbol{p}$ and attribute $m$, $p_m^+$ is the value of the $m^{th}$ attribute in a candidate reference image closest to $\boldsymbol{p}$ along $m$, while ensuring that the attribute strength in the target image $t_m$ falls between the strength of the attribute in $\boldsymbol{p}$ and this reference image, i.e., $t_m \in [p_m, p_m^+]$. Since the relative attribute predictors are not perfect, it is possible that even though the user says "more $a_{m^*}$ than $\boldsymbol{p}^*$, $t_{m^*}$ for the true target image is actually less than $p_{m^*}^*$. In this case, $p_{m^*}^+ \neq r_{m^*}(\boldsymbol{p}^+)$. See Figure 2. In the following, $\boldsymbol{t}$ is the true target image during training, or the candidate target image being considered at test time.

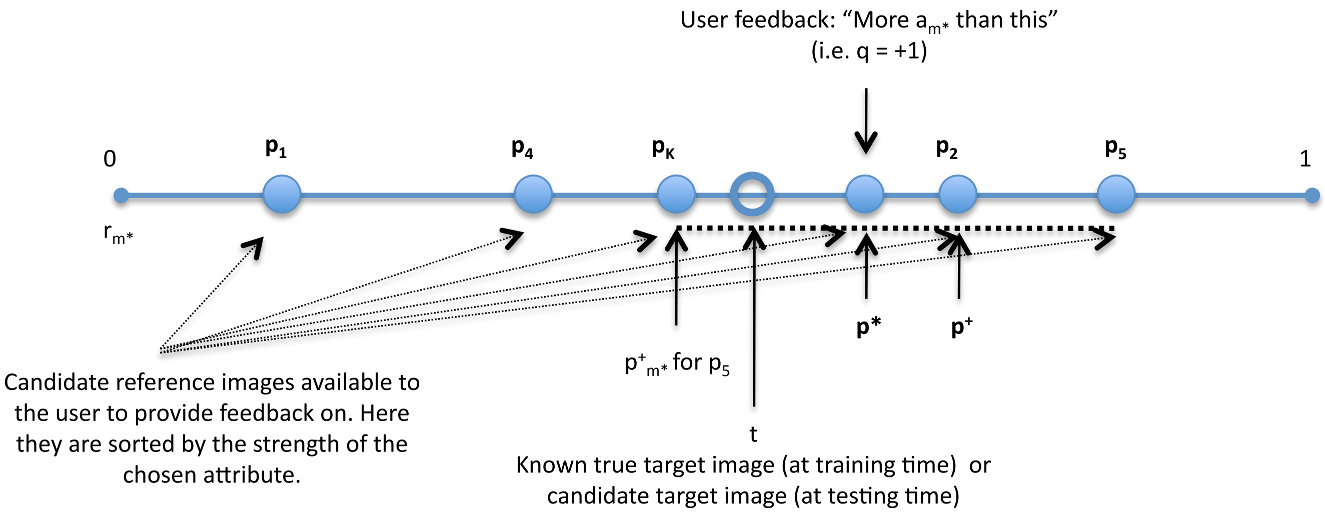

User feedback: “More a$_{m*}$ than this”
(i.e. q = +1)
0
p$_1$
p$_4$
p$_K$
p$_2$
p$_5$
1
r$_{m*}$
p*
p$^+$
p$^+{}_{m*}$ for p$_5$
t
Known true target image (at training time) or candidate target image (at testing time)
Candidate reference images available to the user to provide feedback on. Here they are sorted by the strength of the chosen attribute.

Figure 2: Illustration to visualize our notation.

1. **Satisfies:** $q.(\text{sign}(t_{m^*} - p^*_{m^*}))$
   This feature is 1 if the target image satisfies the user provided feedback statement, and -1 otherwise. If the user says “more $a_{m^*}$ than $\boldsymbol{p}^*$” i.e. $q = 1$ and the target image does have a stronger presence of attribute $a_{m^*}$ i.e. $t_{m^*} > p^*_{m^*}$, this feature is 1. If the the user says “less $a_{m^*}$ than $\boldsymbol{p}^*$” i.e. $q = -1$ and the target image does have a weaker presence of attribute $a_{m^*}$ i.e. $t_{m^*} < p^*_{m^*}$, this feature is 1. Otherwise, this feature is -1.

2. **Soft-satisfies and closeness:** $q.(t_{m^*} - p^*_{m^*})$
   This is a softer version of the feature described above. It is positive if the target image satisfies the user's feedback, and negative otherwise. But it also captures by what amount the target image satisfies the user's feedback. If the user says “I want an image that is furrier than this image”, this feature captures whether the target image is significantly more furry than the reference image, or just a tad bit more furry, or sometimes a little less furry, etc.

3. **Tightness:** $\log\left(\frac{|t_{m^*} - p^*_{m^*}|}{p^+_{m^*} - p^*_{m^*}}\right)$
   Let's say the user has 8 reference images to choose from to provide feedback. Let's say the target image is more shiny than 4 of the 8 images. Perhaps the user would comment on the shiniest of these 4 images to say “What I want is more shiny than this image” – since that is the tightest and most informative constraint he can provide. This means that the target image is closer to the chosen reference image than the reference image shinier than the chosen reference image. The numerator in the above feature is the distance between the target image and the chosen reference image along the chosen attribute $a_{m^*}$, while the denominator is the distance between the chosen reference image and the reference image consecutive to the chosen image when all images are sorted by the strength of the chosen attribute present in them. If our hypothesis is true, the ratio above would like between 0 and 1. We show the numerator and denominators of the above feature in the illustration in Figure 3. We use log to control the scale of the feature.

4. **Relative closeness:** $\frac{|p^*_{m^*} - t_{m^*}|}{\max_p p_{m^*} - \min_p p_{m^*}}$
   Perhaps the target image usually lies at distance from the reference image much smaller than the entire range of attribute values covered by the candidate reference images. If this is true, the above ratio (between 0 and 1) is usually close to 0. We show the numerator and denominators of the above feature in the illustration in Figure 4.

5. **Optimize attribute and reference image selection for closeness:** $\frac{\min_{p,m} |p_m - t_m|}{|p^*_{m^*} - t_{m^*}|}$
   We now consider a hypothesis that expands on the one in Feature 2. Among all the choices of attributes and reference images, perhaps a users picks the reference image and attribute so that the target image can be as close to the reference image as possible along the chosen attribute. If this hypothesis is true, the above ratio (between 0 and 1) would usually be close to 1, because the distance between the chosen reference image and the target along the chosen attribute (denominator) would be close to the minimum possible given all the choices of reference images and attributes

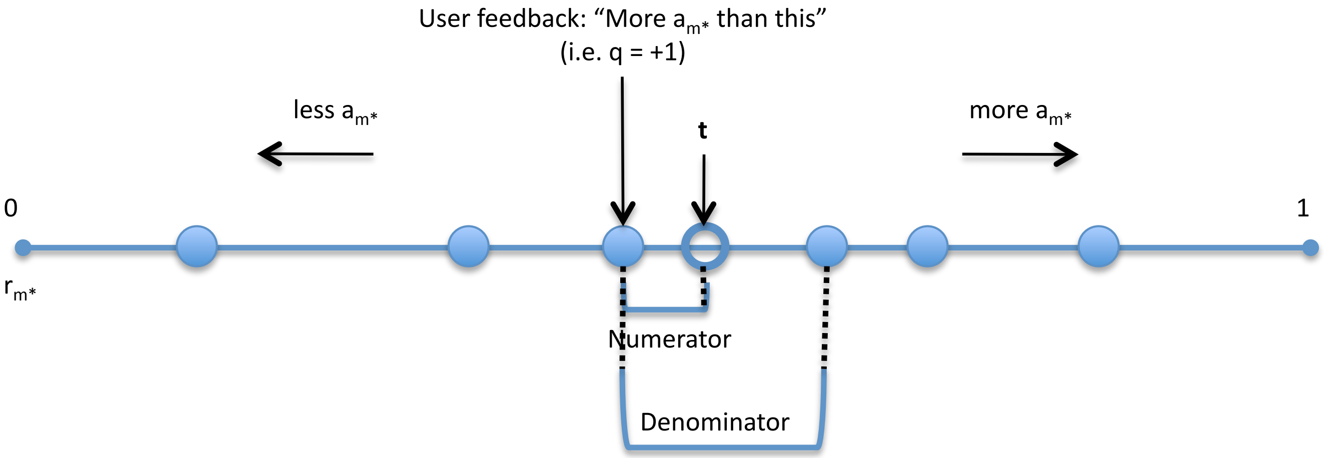


Figure 3: Illustration to visualize feature 3.

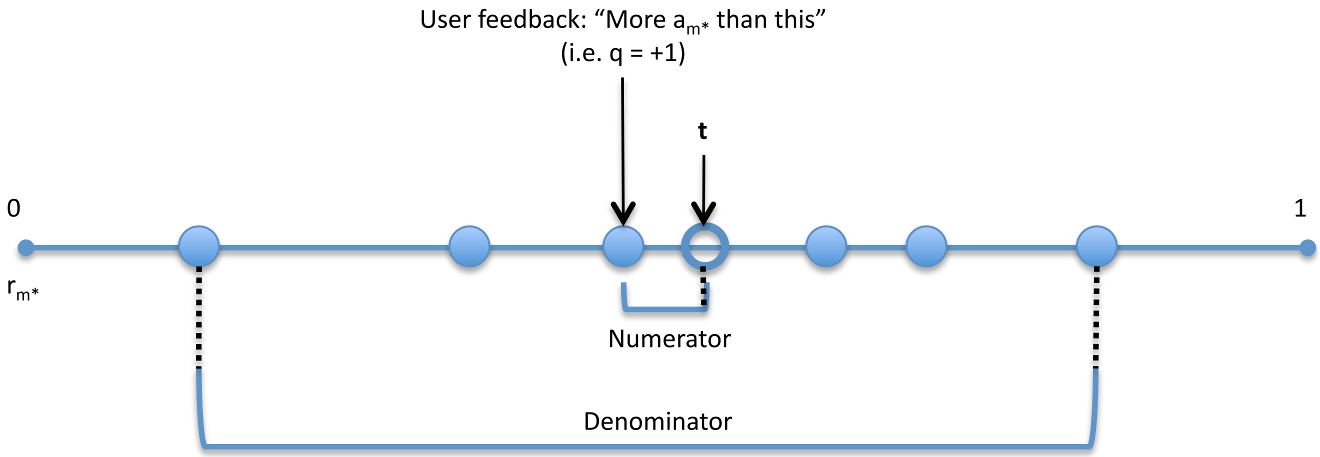


Figure 4: Illustration to visualize feature 4.

(numerator). The next two features (6 and 7) are similar in spirit, but replace the $\min$ operator with average and $\max$ operators.

6. $\frac{|p^*_{m^*} - t_{m^*}|}{\text{avg}_{p,m}|p_m - t_m|}$

7. $\frac{|p^*_{m^*} - t_{m^*}|}{\max_{p,m}|p_m - t_m|}$

8. **Optimize reference image selection for closeness:** $\frac{\min_p |p_{m^*} - t_{m^*}|}{|p^*_{m^*} - t_{m^*}|}$
   Another hypothesis, one that is a bit looser than the one presented for Feature 5, is that whatever the reason to pick the attribute, once it is picked, users pick the reference image that allows the target image to be as close to the reference image as possible. So the above ratio (between 0 and 1) is usually close to 1. Note that the $\min$ operator is only over the choice of reference images; the attribute is fixed. The next two features (9 and 10) are similar but replace the $\min$ operator with average and $\max$.

9. $\frac{|p^*_{m^*} - t_{m^*}|}{\text{avg}_p|p_{m^*} - t_{m^*}|}$

10. $\frac{|p^*_{m*} - t_{m*}|}{\max_p |p_{m*} - t_{m*}|}$

11. **Optimize attribute selection for closeness:** $\frac{\min_m |p^*_m - t_m|}{|p^*_{m*} - t_{m*}|}$
The flip hypothesis is that whatever the reason to pick the reference image, once it is picked, users pick the attribute that allows the target image to be as close to the reference image as possible. This time the $\min$ operator is over the choice of attributes for a fixed chosen reference image. Again, if this hypothesis is true, the above ratio (between 0 and 1) would usually be close to 1. The next two features (12 and 13) are similar but with the $\min$ operator replaced by average and $\max$.

12. $\frac{|p^*_{m*} - t_{m*}|}{\text{avg}_m |p^*_m - t_m|}$

13. $\frac{|p^*_{m*} - t_{m*}|}{\max_m |p^*_m - t_m|}$

14. **Optimize attribute and reference image selection for tightness:** $\frac{\min_{p,m} |p_m - p^+_m|}{|p^*_{m*} - p^+_{m*}|}$
We now consider an extension of the hypothesis in Feature 3. Maybe users pick the attribute and reference image such that the target image falls in the smallest interval formed by any two consecutive candidate reference images (sorted by their relative attribute values). In this case, the above ratio (between 0 and 1) would usually be close to 1. The next feature (15) is similar in spirit, but replaces the $\min$ operator with the average.

15. $\frac{|p^*_{m*} - p^+_{m*}|}{\text{avg}_{p,m} |p_m - p^+_m|}$

16. **Optimize reference image selection for tightness:** $\frac{\min_p |p_{m*} - p^+_{m*}|}{|p^*_{m*} - p^+_{m*}|}$
Again, a relaxed hypothesis is that whatever the reason to pick the attribute, once it is picked, maybe users pick the reference image such that the target image falls in the smallest interval formed by any two consecutive candidate reference images (sorted by the picked relative attribute value). So the above ratio (between 0 and 1) is usually close to 1. The next feature (17) is similar, but with an average operator instead of $\min$.

17. $\frac{|p^*_{m*} - p^+_{m*}|}{\text{avg}_p |p_{m*} - p^+_{m*}|}$

18. **Optimize attribute selection for tightness:** $\frac{\min_m |p^*_m - p^{*+}_m|}{|p^*_{m*} - p^+_{m*}|}$
On the flip side, whatever the reason to pick the reference image, once it is picked, maybe users pick the attribute such that the target image falls in the smallest interval formed by any two consecutive candidate reference images (sorted by their relative attribute values). So the above ratio (between 0 and 1) is usually close to 1. The next feature (19) is similar, but with an average operator instead of $\min$.

19. $\frac{|p^*_{m*} - p^+_{m*}|}{\text{avg}_m |p^*_m - p^{*+}_m|}$

20. **"Like-this-but":** $\frac{\frac{|p^*_{m*} - t_{m*}|}{\max_{m \neq m*} |p^*_m - t_m|}}{\max_{p,m} \frac{|p_m - t_m|}{\max_{m' \neq m} |p_{m'} - t_{m'}|}}$
Our next hypothesis is that when a user says "I want something with more $a_{m*}$" maybe he really means "I want something *like this*, but more $a_{m*}$". So they pick the reference image and attribute such that it has a high difference from the target image along the picked attribute, but is similar to the target image along other attributes. The numerator in the above expression is small when the distance of the target image from the chosen reference image along the chosen attribute is much larger than the difference between the two along all other attributes. The denominator searches over all available reference images and attributes to make this quantity as large as possible. If the user picks the attribute and reference image that do maximize this quantity, the above ratio (between 0 and 1) would be close to 1. The remaining features (21-31) below capture a similar idea, but replace the $\max$ operators above with averages, and optimize over just reference images or just attributes or both.

21. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p^*_m-t_m|}}{\mathrm{avg}_{p,m}\frac{|p_m-t_m|}{\max_{m'\neq m}|p_{m'}-t_{m'}|}}$

22. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}{\max_{p,m}\frac{|p_m-t_m|}{\mathrm{avg}_{m'\neq m}|p_{m'}-t_{m'}|}}$

23. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}{\mathrm{avg}_{p,m}\frac{|p_m-t_m|}{\mathrm{avg}_{m'\neq m}|p_{m'}-t_{m'}|}}$

24. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p^*_m-t_m|}}{\max_{p}\frac{|p_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p_m-t_m|}}$

25. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p^*_m-t_m|}}{\mathrm{avg}_{p}\frac{|p_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p_m-t_m|}}$

26. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}{\max_{p}\frac{|p_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p_m-t_m|}}$

27. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}{\mathrm{avg}_{p}\frac{|p_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p_m-t_m|}}$

28. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p^*_m-t_m|}}{\max_{m}\frac{|p^*_m-t_m|}{\max_{m\neq m^*}|p^*_m-t_m|}}$

29. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\max_{m\neq m^*}|p^*_m-t_m|}}{\mathrm{avg}_{m}\frac{|p^*_m-t_m|}{\max_{m\neq m^*}|p^*_m-t_m|}}$

30. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}{\max_{m}\frac{|p^*_m-t_m|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}$

31. $\dfrac{\frac{|p^*_{m^*}-t_{m^*}|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}{\mathrm{avg}_{m}\frac{|p^*_m-t_m|}{\mathrm{avg}_{m\neq m^*}|p^*_m-t_m|}}$

We stress that all the hypotheses listed above are simply possible behaviors that we want our features to expose to the rank learning algorithm. Ultimately, their impact will be entirely learned, and is not hand-coded by us.

## References

[1] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013.

[2] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 2007.

[3] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[4] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image search with relative attribute feedback. In *CVPR*, 2012.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

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