Confidence and Diversity for Active Selection of Feedback in Image Retrieval

Bhavin Modi
bhavin_modi@hotmail.com
Adriana Kovashka
kovashka@cs.pitt.edu

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, USA

Abstract

Image search is a challenging problem because of the need to model any concept the user might want to retrieve. One recent solution to the problem allows the user to give feedback on the current set of results, by answering questions about how the relative attributes of individual returned images relate to his/her target image. We show how to ask more informative questions. In our active selection formulation that determines about which attribute the system should next ask a question, we account for the confidence of relative attribute models. In addition to asking about reliably modeled attributes, the system is also encouraged to ask diverse questions, by computing question diversity on both the attribute and image levels. We show that both of our novel active selection criteria, confidence and diversity, help improve search results on three datasets. Further, when used in combination, they boost performance more than either cue alone.

1 Introduction

Image search is one of the computer vision tasks that ordinary people often perform in their everyday lives. One approach to image retrieval allows the user to submit a text-based keyword query, followed by the system retrieving images that have been tagged with these keywords. Unfortunately, many relevant images will not be retrieved simply because the person who uploaded them did not provide the exact keywords that the search user named. An alternative to this approach is content-based image retrieval, which relies on image features. However, as any image understanding algorithm, content-based retrieval suffers from the frequent misalignment between low-level features and the user's concept. To fix the system's mistakes, the user can perform relevance feedback on the current set of results, i.e. mark which content is relevant and which content is not [47]. The user can also more specifically describe how the results should change to better satisfy his/her target (the desired image) [28], using statements that compare the relative attributes [39] of the target and results.

However, the user may not know what to give feedback on. [25, 28] show that search can be sped up through active selection of attribute-based questions. For example, their system might ask the user "Is the person you are looking for younger or older than the person in this image?" [28]'s active selection formulation is based on estimates of entropy, but they might be inaccurate, hence may lead to flawed selection of the next question to ask. We propose two ways to ameliorate the effect of unreliable entropy estimates. First, we

© 2017. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

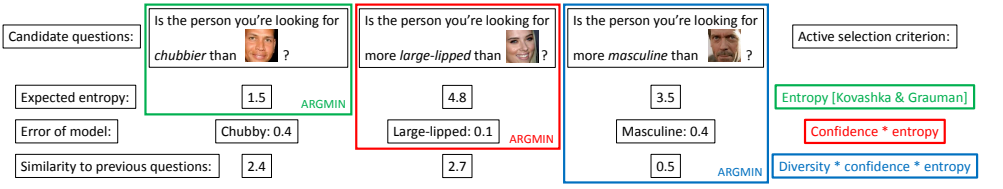

Figure 1: Illustration of our method. In addition to measuring the entropy over the relevance of each database image, we also measure the confidence of the model for the attribute that each question involves. We also discourage the system from asking questions too similar to previously asked questions. The smaller the entropy, error (the opposite of confidence), and similarity to previous questions, the more likely this question will be selected. Incorporating additional cues causes different questions to be selected (using argmin).

propose to account for the confidence of the offline-computed attribute prediction models. If an attribute model is inaccurate, then even if the system estimates that asking a question about this attribute would reduce entropy, the question might be uninformative. Second, unlike [28], we propose to explicitly account for potential redundancy in the questions. Even if information gain indicates we should keep asking questions about some attribute, once several have been asked, the actual information gain from continuing to ask about the same attribute might be low. Thus, in our active selection formulation, we weigh the expected entropy that might result from asking a question about some attribute by: (1) the confidence of the detection model for that attribute, and (2) a measure of how distinct this candidate question is from those the user has already answered. Our method is illustrated in Fig. 1.

We find that incorporating model confidence in the active selection formulation significantly improves the accuracy of search results. In conjunction with confidence, discouraging similar questions from being asked improves results further.

2 Related Work

Image retrieval. Web search is a task many people perform daily. *Image* search in particular has many uses: one might want to see what a certain actor looks like, explore a vacation destination, learn about plants and animals, buy clothing, etc. Using keywords to perform search is limited as relevant characteristics of the image will often not be described by the available keywords. Users can also query with images [6, 16] or sketches [57], but if the concept the user wishes to retrieve is not captured in any images that the user has at hand, or the user cannot sketch well, results will be imprecise.

Interactive image retrieval. To cope with the limitations of keyword-based and image-based queries, researchers proposed interactive image search [5, 7, 11, 12, 44, 47, 58] where users iteratively provide feedback to the system. For example, images in the dataset are first ranked by how well they match a set of keywords, the user is presented with the top-ranked results, and then he/she gives feedback on these results in order to refine them. Traditional relevance feedback methods allow the user to only provide coarse feedback, e.g. in the form of binary statements like "this result is relevant and that result is irrelevant."

The WhittleSearch system [28] proposes a more descriptive approach for interactive image search. The search scenario is as follows: A user is trying to find an image that represents a particular concept, but does not have the actual image at hand, so he/she describes the concepts via keywords. The system allows the user to make statements like "The shoes I am

looking for are *more formal* and *less bright in color* than these shoes," and point to one of the currently top-ranked images. The italicized terms are visual attributes for which the system has learned models offline. [28] show this approach results in more accurate searches.

Attributes. Attributes (e.g. "smiling" person, "serene" scene, "shiny" dress) [10, 31, 41] have been used to recognize unseen objects [19, 31, 39, 54], compactly represent object classes [45, 56], perform fine-grained recognition [1, 3], describe 3D structures [13], etc. Importantly, they have been found useful for image retrieval [17, 26, 27, 28, 30, 43, 46, 49].

Active learning for image retrieval. [28] incorporate active learning with attribute-based search. The user's feedback now takes the form of a 20-questions game. In particular, both the image and attribute for the relative comparison are chosen by the system. For example, the system can ask "Are the shoes you are looking for more or less formal than these shoes?" where the underlined terms are selected by the system. Previously, active learning for retrieval was explored in [7, 11, 50] for binary relevance feedback.

Confidence of models. In Sec. 3.2.1, we propose to incorporate model confidence in our active selection formulation. [37, 48, 53] measure how confident classifiers are in their decisions, and [23, 59] model classifier errrors. [32] use confidence as an uncertainty measure during active learning. [8, 9, 34] measure confidence over all model parameters and prioritize updates on the less confident ones. In contrast, our method uses confidence to *weigh* our measure of uncertainty. We believe we are the first to use confidence as a way to safeguard against unreliable uncertainty estimates.

Diversity. In Sec. 3.2.2, we encourage our method to select feedback on diverse content. [35, 36] use diverse ensembles to select items to label. [4] use a diversity criterion to select batches of samples for labeling, but the approach is computationally expensive. [55] use diversity for relevance feedback, but limit their selection only among the top-ranked documents. Our method is the first to use diversity in active learning for image retrieval. Rather than select within a homogeneous pool (e.g. documents), our method makes a two-level selection from a pool of heterogeneous questions (composed of both images and attributes). Further, our method makes its selection in real time.

3 Approach

We first describe [28]'s active selection method that chooses the feedback questions to ask the user. We incorporate additional selection cues into this method, in three ways which we discuss in turn: by accounting for the confidence of attribute models, by encouraging diversity, and by incorporating both confidence and diversity into the selection criterion.

The pipeline for any of the active selection methods (each corresponding to a separate image retrieval system) that we propose is as follows. The system receives an initial set of results, retrieved in response to an initial user query, or simply retrieved as a random sample of the dataset. The search task is for the system to retrieve an image that the user envisions, but does not have in hand, hence does not feed to the system. The goal for the active selection methods is to ask a series of questions that would allow the system to most quickly retrieve that image. The output of the system after a certain number of questions have been asked, or after the user terminates the search, is a final ranking of all database images with respect to their relevance according to the user's interactions with the system.

3.1 Active selection for attribute-based retrieval

[28] propose a method for image retrieval that asks the user *actively selected* questions about what the target image (the one he/she wishes to retrieve) should look like, and ranks images

based on how well they align with the user's answers. We next describe this method as we will reuse some of its notation and computations, and discuss its limitations.

In [28], images are ranked based on a probabilistic estimate of how well they satisfy all feedback constraints given by the user. A constraint is of the form $\{I_{p_m}, m, r\}$, where $m \in M$ is an attribute in the attribute vocabulary of size $|M|$, I_{p_m} is the "pivot" image for that attribute (defined below), and r is the user's response ("more," "less," or "equally"). For example, a question can be "Is the target image you are looking for more/less/equally m than image I_{p_m}?" and an answer can be "My target image is r m than/as image I_{p_m}."

A pivot image is that image which currently indicates the system's best guess about the strength of attribute m in the target image. [28] keep a list of one pivot per attribute, denoted by $\mathcal{P}$. The pivot is a node in a binary search tree for that attribute, and is initialized as the root of the tree, then updated in accordance with the user's answers on this attribute. The system's goal is to select the attribute and pivot about which it should ask a question next.

To select the pivot I_{p_m}, [28] propose the following active selection formulation:

$$I_p^* = \arg\min_{I_{p_m} \in \mathcal{P}} \sum_r P(r|I_{p_m}, \mathcal{F}) \;\; H(\mathcal{F} \cup (I_{p_m}, r)) \tag{1}$$

The last term above captures the uncertainty that the system has, about whether any database image is relevant to the user's query, based on all feedback statements $\mathcal{F}$ and the candidate statement. The latter includes the candidate question on pivot I_{p_m} (for attribute m) and the (yet unknown) user-given response r. The entropy is computed as follows in [28], where N is the number of images in the dataset, I_i is a database image and y_i denotes whether I_i is relevant ($\ell = 1$) or not ($\ell = 0$) to the target.

$$H(\mathcal{F}) = -\sum_{i=1}^{N} \sum_{\ell \in \{0,1\}} P(y_i = \ell | I_i, \mathcal{F}) \log P(y_i = \ell | I_i, \mathcal{F}), \tag{2}$$

The probability that an image is relevant to the user's query, $P(y_i = \ell | I_i, \mathcal{F})$, is computed as the product of probabilities measuring to what degree an image satisfies all user-given feedback statements on attribute strengths. Let $a_m(I_i)$ denote the predicted attribute strength for attribute m in image I_i. To compute the probability that I_i satisfies a given constraint involving a pivot I_{p_m}, [28] compute the difference $a_m(I_i) - a_m(I_{p_m})$, and transform it into a probability [42]. The predicted attribute strengths are obtained from learned ranking functions a_m trained with about 200 pairs of images O_m per attribute.

The first term in Eq. 1 is the probability that a user provides a particular response $r \in$ {"more," "less," "equally"} to the question about I_{p_m}. To compute this probability, [28] "borrow" the probability of being relevant for the currently highest-ranked image, i.e. the system's current best guess about what the target image looks like.

To summarize, [28] compute the information gain of asking a question about pivot I_{p_m} (which uniquely identifies the corresponding attribute m) as the expected entropy over the relevance of all database images. This entropy depends on the the probability of satisfying all feedback constraints, and on the predicted attribute strengths a_m which may differ from the true (perceived by annotators and users) strengths A_m.

3.2 Choosing better questions with confidence and diversity

The attribute models a_m underlie all aspects of the attribute selection function of [28], including relevance probabilities and probabilities of particular user responses. Thus, entropy estimates may be flawed due to the misalignment between the attribute prediction functions

a_m and the true A_m. These flaws could cause entropy for some pivots to be misleadingly low (i.e. the pivots would be erroneously very likely to be chosen).

One way to deal with flaws in entropy is to explicitly account for how confident we are in a given attribute ranker's predictions. We propose how to do so in Sec. 3.2.1. Another way is to introduce a prior, e.g. one that discourages very similar questions from being asked in sequence, even if their entropy is very low. We describe how to encourage diversity in the questions being asked in Sec. 3.2.2. In Sec. 3.2.3, we combine modeling confidence and encouraging diversity into one active selection formulation.

In Sec. 4, we experimentally verify the contribution to search accuracy of all three active selection cues. In order to demonstrate that it is namely the confidence and diversity rather than some other aspect of the method that increases search accuracy, we follow [28]'s approach, apart from the active selection formulation (Eq. 1). We show that while our modifications are simple, they greatly boost the accuracy of search.

3.2.1 Incorporating attribute models' confidence

We want to pick such pivots that will reduce the entropy (uncertainty) of the system, but also weigh the expected entropy estimates by our confidence in a given attribute ranker's predictions. For example, if we are confident a ranker a_m is well-aligned to perceived attribute strengths A_m, then we will multiply the entropy resulting from a question on pivot I_{p_m} by a low score (since we are minimizing Eq. 1). Otherwise we will multiply the entropy by a high value, hence the corresponding pivot will have less chance to be picked as the best pivot.

We want to know how reliable our attribute model for some pivot is. We calculate the error rates for each attribute model on the ground truth annotations O_m provided by [28], using 5-fold cross validation. We denote the error rate per attribute as ε_m. This error rate is always in the range [0, 1]. Low error corresponds to high confidence.

The expected entropy computed in Eqn. 2 however need not lie in the [0, 1] range, so we compute normalized expected entropy:

$$U(\mathcal{F} \cup (I_{p_m}, r)) = \frac{\sum_r P(r|I_{p_m}, \mathcal{F}) \;\; H(\mathcal{F} \cup (I_{p_m}, r))}{\sum_m \sum_r P(r|I_{p_m}, \mathcal{F}) \;\; H(\mathcal{F} \cup (I_{p_m}, r))}. \tag{3}$$

We then select the pivot about which the system should ask the next question as that pivot which corresponds to a confident attribute ranker and low expected entropy:

$$I_p^* = \arg\min_{I_{p_m} \in \mathcal{P}} (\varepsilon_m \cdot U(\mathcal{F} \cup (I_{p_m}, r))). \tag{4}$$

In this way, our method will not be misled by attribute pivots with erroneously low expected entropy due to unreliable attribute rankers.

3.2.2 Encouraging question diversity

The ultimate goal of active selection questions is to obtain high search accuracy. We do not know what the user's target is and cannot measure accuracy during selection, so we use entropy as a proxy. However, entropy is only an approximate cue and might be flawed, so we build safeguards into the selection function.

One prior for what defines a useful question is one that we have not asked before. Imagine that a method keeps asking questions about the same attribute but using different pivots (which become progressively finer approximations to the degree of attribute strength for m that the user desires, and progressively lower nodes in the search tree). On one hand, this

exploration of the same attribute could indicate that if we could only get to the exact level of strength of this particular attribute, we would find the exact image that the user is looking for. On the other hand, several attributes are likely to be important in the user's search query, so getting "stuck" on the same attribute may be undesirable.

To prevent excessive exploration of the same attribute (or of different but related attributes), we encourage the questions that the system asks to be diverse. Thus, for each candidate question involving attribute m and pivot I_{p_m}, we compute how different it is from the previously asked questions. We define the difference between questions as the product of the difference of the attributes, times the difference of the images involved in the questions.

Let $Q_{p_m} = \{I_{p_m}, m\}$ be the question we are evaluating as a potential question for feedback, and $Q_k = \{I_{p_z}, z\} \in \mathcal{Q}$ (where $z \in M$ is an attribute index and I_{p_z} is the corresponding pivot) be some question we asked previously. The set $\mathcal{Q}$ is almost the same as the feedback statement set $\mathcal{F}$, but ignores the user response. We define the difference between Q_{p_m} and Q_k as:

$$\text{Diff}(Q_{p_m}, Q_k) = \text{Dist}(I_{p_m}, I_{p_z})(1 - \text{NormCorr}(m, z)), \tag{5}$$

where

$$\text{Dist}(I_{p_m}, I_{p_z}) = \sqrt{\sum_n (a_n(I_{p_m}) - a_n(I_{p_z}))^2}, \tag{6}$$

$$\text{NormCorr}(m, z) = \frac{\text{Corr}(m, z) - \min_{z \in M} \text{Corr}(m, z)}{\max_{z \in M} \text{Corr}(m, z) - \min_{z \in M} \text{Corr}(m, z)}, \tag{7}$$

and $\text{Corr}(i, j)$ is Kendall's coefficient to measure correlation between the predictions of the attribute rankers a_i and a_j on the full set of database images.

We score candidate question Q_{p_m} with the smallest difference between it and any previous question, as we want to ensure that the question we are asking is not too similar to *any* previously asked question, not just from the most different question:

$$\text{Diff}(Q_{p_m}, \mathcal{Q}) = \min_{Q_k \in \mathcal{Q}} \text{Diff}(Q_{p_m}, Q_k). \tag{8}$$

We normalize this quantity so it is in the range [0, 1] and use this as our final measure of the diversity that the candidate question $Q_{p_m} = \{I_{p_m}, m\}$ can introduce:

$$D_m(Q_{p_m}) = \frac{\text{Diff}(Q_{p_m}, \mathcal{Q})}{\sum_m \text{Diff}(Q_{p_m}, \mathcal{Q})}. \tag{9}$$

We then select the best pivot as that pivot which is both *maximally different/diverse* from (i.e. *minimally similar* to) previously asked questions, and minimizes entropy:

$$I_p^* = \arg\min_{I_{p_m} \in \mathcal{P}} ((1 - D_m(Q_{p_m})) \cdot U(\mathcal{F} \cup (I_{p_m}, r))), \tag{10}$$

where we convert difference to similarity.

3.2.3 Incorporating confidence and diversity

Since we believe that both the confidence and diversity cues are useful for selecting informative questions, we also combine them as follows:

$$I_p^* = \arg\min_{I_{p_m} \in \mathcal{P}} (\varepsilon_m \cdot (1 - D_m(Q_{p_m})) \cdot U(\mathcal{F} \cup (I_{p_m}, r))). \tag{11}$$

The intuition behind this approach is that asking diverse questions may not always be best. If the attribute for which we are asking seemingly redundant questions has the attribute prediction model with the lowest error rate (highest confidence), then continuing to ask questions about it may still be informative.

4 Experimental Setup and Results

Our paper's contributions lie in proposing an attribute-based active selection formulation for visual search. Note that attributes have previously been shown useful for search [17, 28, 30, 43, 46, 49], and active learning for visual recognition has received consistent attention [14, 15, 18, 22, 24, 33, 40, 51, 52]. However, [25, 28] is the only prior work which proposes an attribute-based active selection method for search, hence it is the only relevant method against which we can compare. We present an evaluation of how much each of the pivot selection methods proposed in Sec. 3 contributes to the accuracy of the user's search results.

We use three datasets: **PubFig**, which contains 769 images (after de-duplication) from [29] and 11 attributes (e.g. "chubby," "narrow-eyed," "young") from [39]; **Scenes** with 2,688 images and 6 attributes ("natural," "open") from [38, 39]; and **Shoes** with 12,807 images (after de-dup.) from [2] and 10 attributes ("brightly-colored," "feminine," "sporty") [28].

We extract features from the *fc6*, *fc7* and *fc8* layers of CaffeNet [20]. We compare the attribute prediction error rates of all three features, and select (on a validation set) the best one to use (*fc6* for **Shoes** and **Pubfig** and *fc7* for **Scenes**). We train attribute ranking models for all attributes, using the data provided by [28] and the SVM Rank code of [21]. The data of [28] consists of about 200 image pairs per attribute, where in each pair, one image has the attribute more than the other. We use the same features for all methods.

We conduct our search experiments on Amazon Mechanical Turk (MTurk). In order to reliably measure the accuracy of search results, we tell the MTurk workers what target image to look for. A total of 50 images from each of the three datasets are selected as target images. Three distinct users do the same search, i.e. look for the same target image. This results in a total of 50x3x3=450 search tasks.[1] We compare how well four methods rank the target image. We create MTurk tasks that each contain a single target image and links to four separate search interfaces, one for each of the methods. The links are presented in random order to ensure there is no bias towards/against the methods that the user employed first. Each task involves the user answering questions that one of the four methods selects.

The methods we compare are: (1) ORIGINAL [28]; (2) CONFIDENCE (Sec. 3.2.1); (3) DIVERSITY (Sec. 3.2.2); and (4) CONFIDENCE+DIVERSITY (Sec. 3.2.3). For each method, the search stops if ten questions have been asked or the user finds his/her assigned target image (the target shows up on the first page of search results containing 4 rows of 9 images).

We store and use as our measure of accuracy the *final rank of the target image* after the ten questions asked, where lower is better (ideally the target image would be ranked at position 1). In Table 1, we show the average and median of the target image ranks for each method. In Fig. 2, we show boxplots over the final ranks, for **PubFig**, **Scenes**, and **Shoes**.

From the results in Table 1, we observe that both our confidence and diversity cues help improve the average rank of the target image, compared to the original method of [28]. CONFIDENCE achieves best or second-best results in 4 of the 6 rows. However, the diversity cue helps achieve additional gains in performance, via CONFIDENCE+DIVERSITY. In 5/6 cases, CONFIDENCE+DIVERSITY combined achieve better results than confidence alone.

[1] Due to human errors while using the search interface we ended up with 443 tasks.

Dataset	ORIGINAL [28]	CONFIDENCE	DIVERSITY	CONF+DIV
PubFig (mean)	158 (± 15)	*137* (± 11)	154 (± 14)	**103** (± 9)
PubFig (median)	64	93	**56**	*58*
Scenes (mean)	807 (± 53)	*653* (± 44)	819 (± 51)	**617** (± 43)
Scenes (median)	657	*535*	691	**488**
Shoes (mean)	*2674* (± 233)	**2658** (± 208)	2703 (± 244)	2845 (± 259)
Shoes (median)	1816	1950	*1717*	**1688**

Table 1: Mean and median target image rank on three datasets, with standard error in parentheses. Ranks are different across datasets due to the different dataset sizes. Lowest is best. In **bold** is the best performer per row, and in *italics* the second-best.

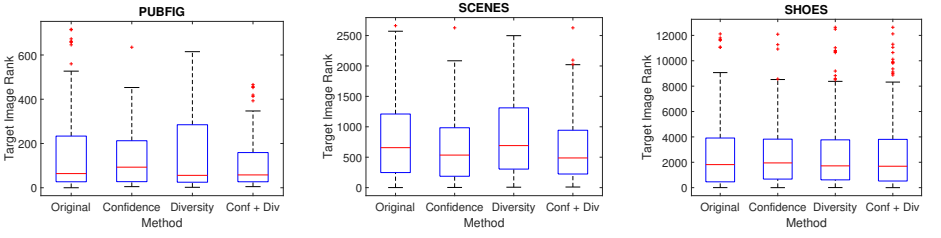

Figure 2: Target image rank for all 443 tasks. Lower is better. The red lines indicate the median target ranks, and the bottom and top edges show the 25th and 75th percentiles.

On **Shoes**, we observe that confidence alone is more useful than diversity alone (mean row), but both cues combined achieve the best performance (median row). On the **PubFig** dataset, DIVERSITY achieves best performance (median row).

Overall, in 4/6 cases, confidence improves performance with respect to [28], diversity in 3/6 of cases, and both in 5/6 of cases. Our method reduces the median target image rank obtained by [28] by 26% for **Scenes**, 9% for **PubFig**, and 7% for **Shoes**. If we average the six rows in Table 1, we obtain: ORIGINAL: 1029, CONFIDENCE: 1004, DIVERSITY: 1023, CONF+DIV: 967 (lowest is best). Thus CONFIDENCE+DIVERSITY improves upon either metric alone, and either metric alone improves over ORIGINAL.

Fig. 2 shows more in-depth results which indicate that for **PubFig** and **Scenes**, CONFIDENCE+DIVERSITY achieves consistently low ranks with small variance, while for **Shoes**, the difference in the performance of the methods is smaller.

To get a practical sense of the utility of our approach, our best method reduces the rank of the target image by 128 images for **Shoes** (using median scores), 169 images for **Scenes**, and 8 images for **PubFig**. Assume it takes a user 1 second to examine a single image result. Then *compared to the method of [28], our method saves the user up to 3 minutes* (169 seconds), in terms of time to browse the results and find one's target.

In Fig. 3, we show how the uncertainty over the relevance of each database image reduces, as the user answers more questions actively selected by the system. While entropy generally decreases for all methods, it *decreases faster* for several of the methods we propose, compared to the ORIGINAL method. We see that the CONFIDENCE and CONFIDENCE+DIVERSITY methods have entropies lower than ORIGINAL. On **Shoes**, DIVERSITY has the most useful behavior compared to other datasets, but in general, it and ORIGINAL have somewhat unstable behavior, which may explain their weaker performance compared to CONFIDENCE and CONFIDENCE+DIVERSITY. On **Pubfig**, combining diversity with confidence achieves the lowest (best) entropy at the end of the search.

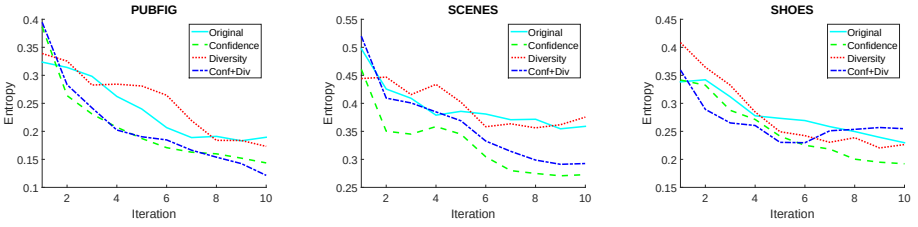

Figure 3: Entropy of being relevant per image, averaged across searches. Lower is better.

Figure 4: An example of how the search proceeded for one **PubFig** target image, for the ORIGINAL, CONFIDENCE, and DIVERSITY methods. In each column, the images on the left are pivots, and the images on the right are the target (not seen by the system except for evaluation). Lower rank is better.

In Fig. 4, we show a real search done on MTurk. We show the first five questions asked and the final target image rank for each method. Note that the CONFIDENCE method (middle) was able to find the target image with only 7 questions, compared to 10 for the other methods. This comes at the expense of asking more similar questions: ORIGINAL asked about 7 attributes total and DIVERSITY asked about 10, while CONFIDENCE asked about 3.

5 Conclusion

We showed that both incorporating the confidence of attribute rankers in the active selection, as well as discouraging newly asked questions from being too similar to previously asked questions, are useful active selection cues and improve the accuracy of searches. In the future, we will explore more priors for the active selection, as well as mixed-initiative collaboration of feedback by the user and questions by the system.

Acknowledgment. We are grateful to Ray Mooney for valuable ideas, and Nils Murrugarra Llerena for logistic support.

References

[1] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *European Conference on Computer Vision (ECCV)*, pages 401–416. Springer, 2014.

[2] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision (ECCV)*, pages 663–676. Springer, 2010.

[3] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision (ECCV)*, 2010.

[4] Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *International Conference on Machine Learning (ICML)*, volume 3, pages 59–66, 2003.

[5] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *International Conference on Multimedia*, pages 1605–1608. ACM, 2010.

[6] Ondřej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2007.

[7] Ingemar J Cox, Matt L Miller, Thomas P Minka, Thomas V Papathomas, and Peter N Yianilos. The bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, 2000.

[8] Mark Dredze and Koby Crammer. Active learning with confidence. In *46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 233–236. Association for Computational Linguistics, 2008.

[9] Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *International Conference on Machine Learning (ICML)*, pages 264–271. ACM, 2008.

[10] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] Marin Ferecatu and Donald Geman. Interactive search for image categories by mental matching. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2007.

[12] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: interactive concept learning in image search. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 29–38. ACM, 2008.

[13] David F. Fouhey, Abhinav Gupta, and Andrew Zisserman. 3D shape attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, 2014.

[15] Efstratios Gavves, Thomas Mensink, Tatiana Tommasi, Cees G. M. Snoek, and Tinne Tuytelaars. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[16] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 817–824. IEEE, 2011.

[17] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[18] Paril Jain and Ajay Kapoor. Active learning for large multi-class problems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[19] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1629–1636, 2014.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[21] Thorsten Joachims. Training linear svms in linear time. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226. ACM, 2006.

[22] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Breaking the interactive bottleneck in multi-class classification with active selection and binary feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[23] Mayank Kabra, Alice Robie, and Kristin Branson. Understanding classifier errors by examining influential neighbors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3917–3925, 2015.

[24] Christoph Kading, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[26] Adriana Kovashka and Kristen Grauman. Attribute adaptation for personalized image search. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[27] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[28] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision (IJCV)*, 115(2):185–210, 2015.

[29] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009.

[30] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Describable Visual Attributes for Face Verification and Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[31] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[32] Mingkun Li and Ishwar K Sethi. Confidence-based active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1251–1261, 2006.

[33] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[34] Varvara Logacheva and Lucia Specia. Confidence-based active learning methods for machine translation. *European Chapter of the Association for Computational Linguistics (EACL)*, page 78, 2014.

[35] Prem Melville and Raymond J Mooney. Constructing diverse classifier ensembles using artificial training examples. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 3, pages 505–510. Citeseer, 2003.

[36] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *International Conference on Machine Learning (ICML)*, page 74. ACM, 2004.

[37] Michael Muhlbaier, Apostolos Topalis, and Robi Polikar. Ensemble confidence estimates posterior probability. In *Multiple Classifier Systems*, pages 326–335. Springer, 2005.

[38] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42 (3):145–175, 2001.

[39] Devi Parikh and Kristen Grauman. Relative attributes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 503–510, 2011.

[40] Amar Parkash and Devi Parikh. Attributes for classifier feedback. In *European Conference on Computer Vision (ECCV)*, pages 354–368. Springer, 2012.

[41] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision (ECCV)*, pages 85–100. Springer International Publishing, 2016.

[42] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[43] Nikita Prabhu and R. Venkatesh Babu. Attribute-graph: A graph based approach to image ranking. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[44] Nikhil Rasiwasia, Pedro J Moreno, and Nuno Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia*, 9(5):923–938, 2007.

[45] Mohammad Rastegari, Ali Farhadi, and David Forsyth. Attribute discovery via predictable discriminative binary codes. In *European Conference on Computer Vision (ECCV)*, pages 876–889. Springer, 2012.

[46] Mohammad Rastegari, Ali Diba, Devi Parikh, and Ali Farhadi. Multi-attribute queries: To merge or not to merge? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[47] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[48] Walter J Scheirer, Anderson Rocha, Ross J Micheals, and Terrance E Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1689–1695, 2011.

[49] Behjat Siddiquie, Rogerio Feris, and Larry Davis. Image Ranking and Retrieval Based on Multi-Attribute Queries. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[50] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *ACM International Conference on Multimedia*, pages 107–118. ACM, 2001.

[51] Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[52] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision (IJCV)*, 108(1-2), 2014.

[53] Rong Wang and Bir Bhanu. Learning models for predicting recognition performance. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1613–1618. IEEE, 2005.

[54] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016.

[55] Zuobing Xu, Ram Akella, and Yi Zhang. *Incorporating diversity and density in active learning for relevance feedback*. Springer, 2007.

[56] Felix X Yu, Liangliang Cao, Rogerio Schmidt Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 771–778, 2013.

[57] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[58] Hanwang Zhang, Zheng-Jun Zha, Shuicheng Yan, Jingwen Bian, and Tat-Seng Chua. Attribute feedback. In *ACM International Conference on Multimedia*, pages 79–88. ACM, 2012.

[59] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3566–3573, 2014.