Image retrieval with mixed initiative and multimodal feedback

Computer vision apps serve a variety of user needs: for example, they can automatically count calories [12], summarize vacation footage [20], "paint" [4], or help users find shoes they want to buy [9] via image search. While for calorie-counting or machine-painting the interaction between the user and the machine is limited to submitting a photograph, for image search the user needs to communicate with the system in a more fine-grained and unrestricted fashion, since success is defined by whether the system successfully "guessed" what the user wanted to find. A person can look for online shopping options on products they saw in a store, or even try to find a criminal they saw in an online database. The user's mental concept of what they wish to retrieve can be arbitrarily subtle hence difficult to capture, and in order to ensure that the system's model of the user's search concept is accurate, the user needs to be able to "explain" to the system how it should adjust its predictions.

Prior work has tackled this challenge in a number of ways. Some work has used semantic visual attributes (like "shiny" or "chubby") [2, 9, 11, 13] to allow the user to give precise language-based guidance to the system. Attributes provide an excellent channel for communication because humans naturally explain the world to each other with adjective-driven descriptions. Attributes have been shown promising as a tool for image search [5, 9, 10, 14, 17, 21]. For example, [9] show how a user can perform rich relevance feedback by specifying how the attributes of a results image should change to better match the user's target image. For example, the user might say "Show me people with longer hair than this one." Another approach has been to engage the user in question-answering with questions that the system estimated are most useful [3, 7]. Thus, in prior work, the initiative for what guidance to give to the system has been taken by either the user [8, 9, 10, 17, 21] or system [3, 7, 18] *but not both.* Another approach has been to allow the user to provide visual cues for what they are looking for, e.g. by drawing a sketch [1, 15, 22, 23]. The system can then retrieve visually similar results. Thus, the user can use either language or visuals to search, but it is not clear which modality is more informative.

In our work, we propose a framework where **either** the user **or** system can drive the interaction, and the input modality can be **either** textual **or** visual, depending on what seems most beneficial at any point in time. For example, the user can kick off the search using a sketch, then refine the results by explaining how the top retrieved images at a certain iteration differ from her mental model. Then the system might ask attribute-based questions, and give control back to the user when it runs out of informative questions to ask, so the user can provide some more free-form attribute feedback of her choosing. Since it is the system that must rank the results, we propose to leave the choice of what is most informative to the system. In other words, the system can decide to let the user lead and *explore*, if it cannot *exploit* any relevant information in a certain iteration. The system can request that the user provides multimodal feedback, i.e. textual **or** visual feedback. To make all these decisions, we train a reinforcement learning (RL) agent.

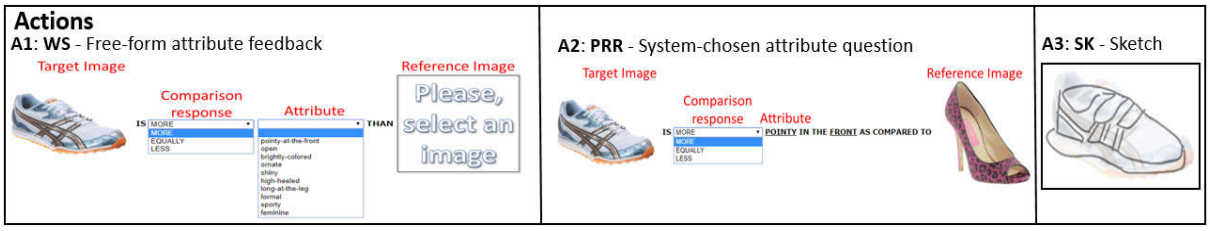

Figure 1. Available actions for our reinforcement learning agent.

In particular, the options that the reinforcement learning chooses between are: (1) sketch feedback, (2) free-form attribute feedback, or (3) system-chosen attribute questions, as shown in Figure 1. At each iteration, the system adaptively chooses one of these interactions and asks the user to provide

the corresponding type of feedback (e.g. it asks the user to choose an image and attribute to comment on). Our agent receives a state composed of top result images, proxies for the target image, and history of taken actions, when available. It interacts with the environment trying different actions. Over time, it learns to pick the most meaningful action, given a certain state. We guide our agent with information about whether the target image is among our top results.

For our experimental evaluation, we compare our RL agent on three datasets: Pubfig, Scenes and Shoes with three different baselines on ten iterations of feedback in Figure 2. We report percentile rank of the target image, defined as the fraction of database images ranked lower than the target (in the range [0, 1], higher is better). The three employed baselines are:

- *Whittle Search [9] (**WS**):* In each iteration, users select a (reference image, attribute) and compare target and reference images for the chosen attribute ("more / less / equally"). The relevance of database images which satisfy this feedback increases.
- *Pivot round-robin [7] (**PRR**):* In each iteration, PRR provides a (reference image, attribute) pair and users select a more / less / equally response.
- *Sketch retrieval [22] + Pivot round robin [7] (**SK_PRR**):* In the first iteration, we ask for a sketch. In later iterations, the system follows the pivot round-robin strategy. Sketches are simulated with edge maps [19]. They are converted to photographs using a GAN [6], and rank database images by their similarity to the target image, using the probabilities from a one-class SVM [16].

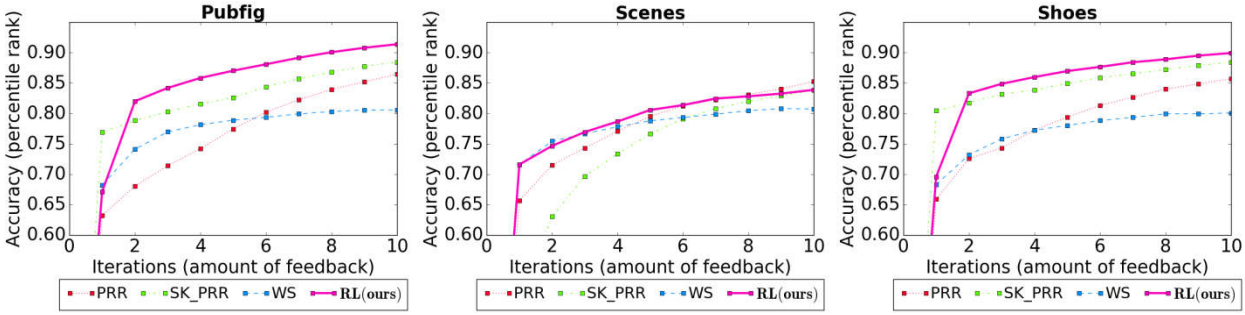

Figure 2. Percentile rank plots for Pubfig, Scenes, and Shoes. Our mixed-initiative RL agent outperforms the other baselines on Pubfig and Shoes, and performs competitively for Scenes.

In order to understand the success of our mixed-initiative RL agent, we count its predicted actions per iteration in Figure 3. We observe that SK (sketch) and WS actions are mainly performed in iterations 1 and 2, because these are the exploration-like actions. Then, after iteration 3, the PRR is the most common one. Once the most beneficial human knowledge is acquired, having a computer suggest feedback (in the form of questions) helps reduce the search space the fastest.

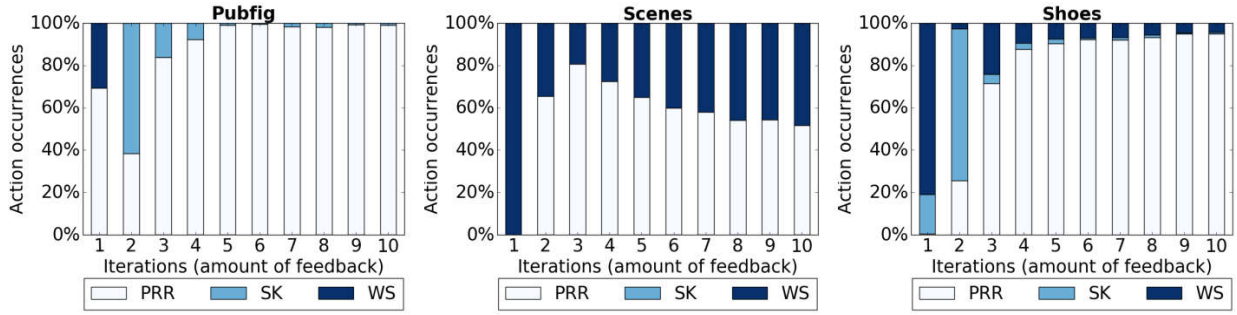

Figure 3. Percentage of actions predicted by our approach in the test set.

To conclude, from experiments in Figure 3, we find that our model prefers human-initiated feedback in former iterations, and complements it with machine-based feedback requests (e.g. questions) in later iterations. We outperform standard image retrieval approaches allowing faster image retrieval.

References

[1] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. Transactions on visualization and computer graphics (TVCG), 2011.

[2] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing Objects by Their Attributes. In Computer Vision and Pattern Recongition (CVPR). IEEE, 2009.

[3] Marin Ferecatu and Donald Geman. A statistical framework for image category search from a mental picture. Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2009.

[4] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[5] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

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

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

[9] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. International Journal of Computer Vision (IJCV), 2015.

[10] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Describable visual attributes for face verification and image search. Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2011.

[11] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In Computer Vision and Pattern Recognition (CVPR). IEEE, 2009.

[12] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P. Murphy. Im2calories: Towards an automated mobile vision food diary. In International Conference on Computer Vision (ICCV). IEEE, 2015.

[13] Devi Parikh and Kristen Grauman. Relative attributes. In International Conference of Computer Vision (ICCV). IEEE, 2011.

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

[15] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. Transactions on Graphics (TOG), 2016.

[16] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. Neural Computing (NC), 2001.

[17] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In Computer Vision and Pattern Recognition (CVPR). IEEE, 2011.

[18] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research (JMLR), 2001.

[19] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In International Conference on Computer Vision (ICCV). IEEE, 2015.

[20] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[21] Aron Yu and Kristen Grauman. Just noticeable differences in visual attributes. In International Conference on Computer Vision (ICCV). IEEE, 2015.

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

[23] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. International Journal of Computer Vision (IJCV), 2017.