Image retrieval with mixed initiative and multimodal feedback

Nils Murrugarra-Llerena and Adriana Kovashka
Department of Computer Science, University of Pittsburgh

BMVC 2018
Newcastle upon Tyne

Introduction

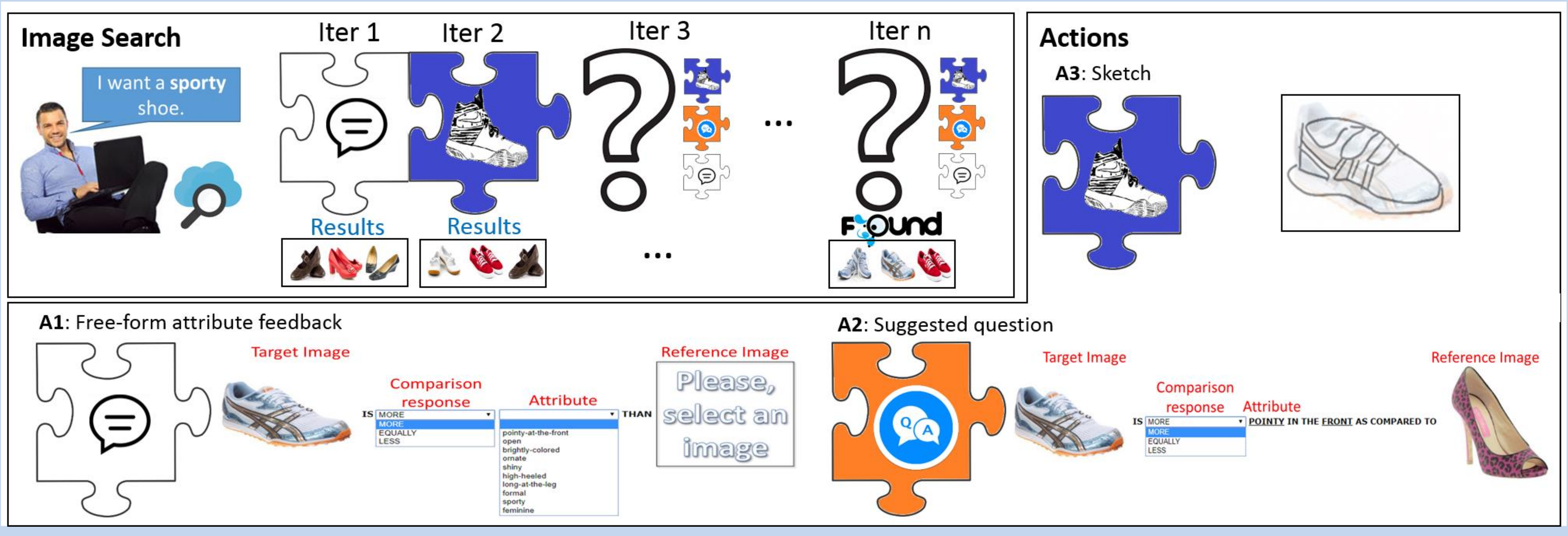
- We **combine** different image retrieval **interactions** to allow faster image search.
- We use a reinforcement learning approach which dynamically decides how to combine: drawing a **sketch**, providing **free-form attribute feedback,** or answering **attribute-based questions**. Our system optimizes informativeness and exploration capabilities for fast and accurate image retrieval.
- Our approach outperforms three different baselines on three datasets.
- We discover the RL agent prioritizes human-initiated feedback and complements it with machine-initiated feedback later in the search cycle.

Motivation

- In prior work, the way to guide the image retrieval is with feedback initiated by either the user (Kovashka et al, CVPR 2012) or system (Ferecatu and Geman, TPAMI 2009; Kovashka and Grauman, ICCV 2013), but **not both**.
- Lack of prior work to jointly explore textual and visual feedback.
- Is there an intelligent way to **combine** user and system **interactions** with **multimodal feedback**?

Key idea

- We learn to **adaptively combine** different forms of **user feedback** (textual or visual) for interactive image retrieval.
- To select the best feedback, we train a RL agent which finds the target image fast.

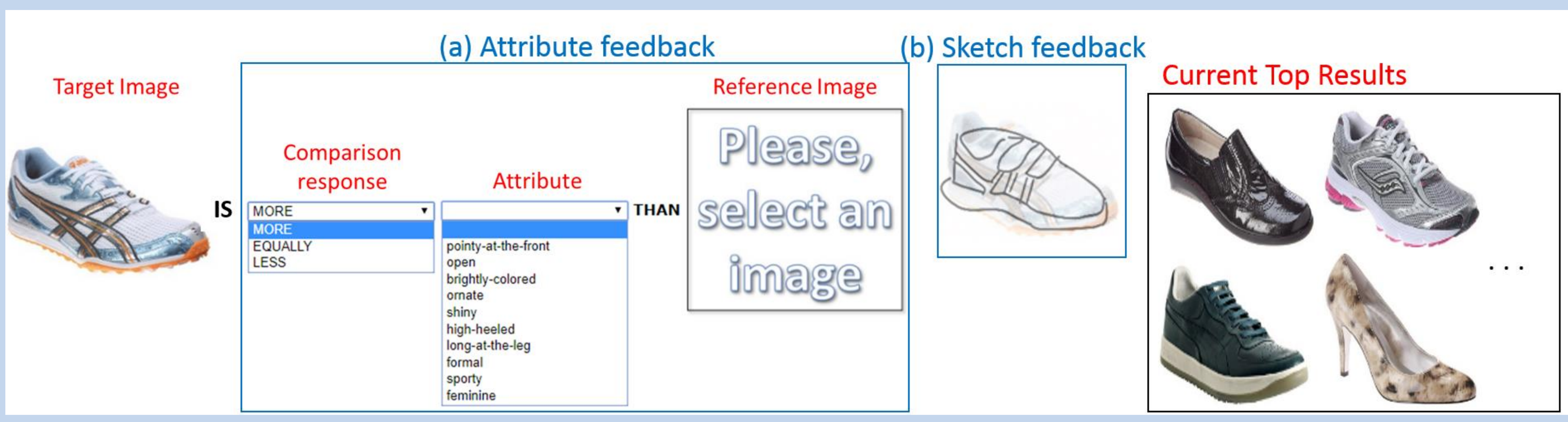


Related work

- Our focus is not on improving sketch-based retrieval, instead we focus on how to decide **when to request** a sketch.
- Rather than retrieve a query and return a single set of results, we engage users in **iterative image retrieval** and show results after each round.
- The most similar work to us is Yin et al (TPAMI 2005). However, it does not allow users to **describe comparatively** how the results should change (i.e. attributes). Instead, each image property is defined as desirable or not.
- Unlike Yin et al (TPAMI 2005), we consider both **textual and visual feedback**.

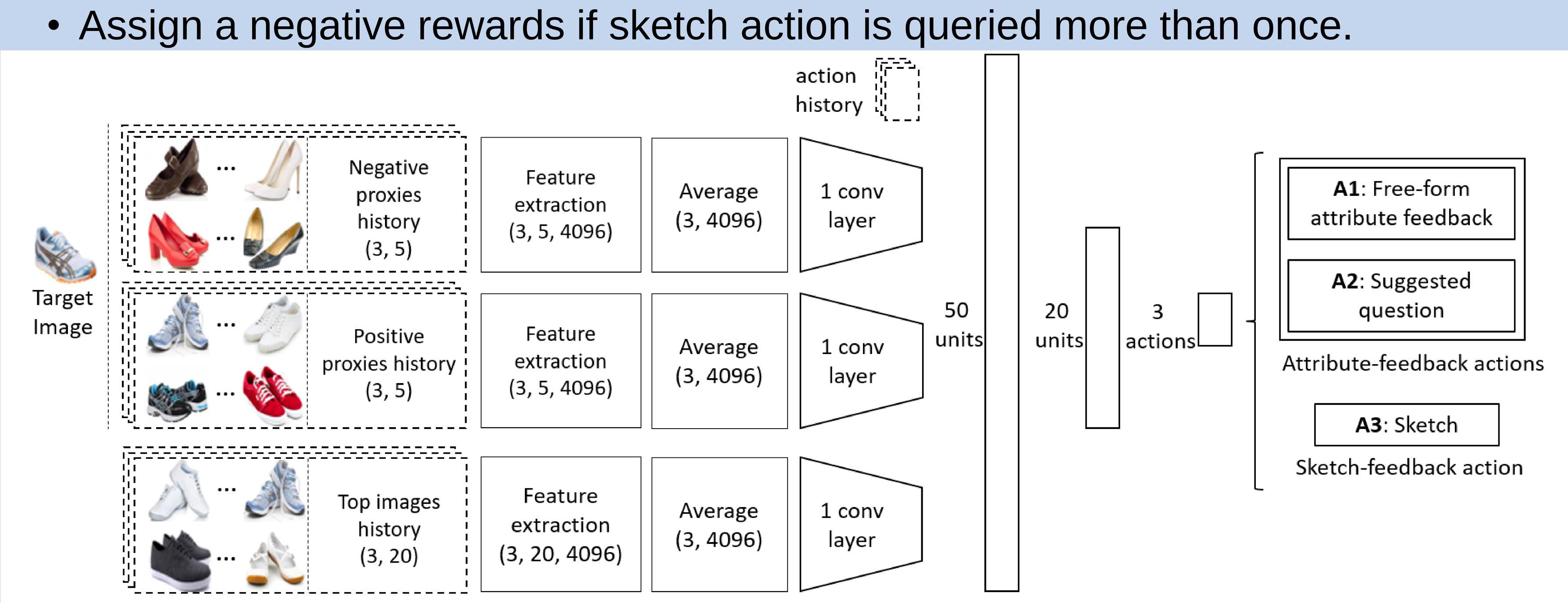
Approach

Interactions



Reinforcement learning agent

- We formulate reinforcement learning as a Markov decision process (MDP).
- **Actions:** attribute-based and sketch-based feedback
- **State:** history of positive and negative proxies of the target image, current top images and actions. Image is represented using features from **AlexNet**.
- **Rewards:**
  - Distance to positive proxies should decrease.
  - Distance to negative proxies should increase.
  - Assign a negative rewards if sketch action is queried more than once.

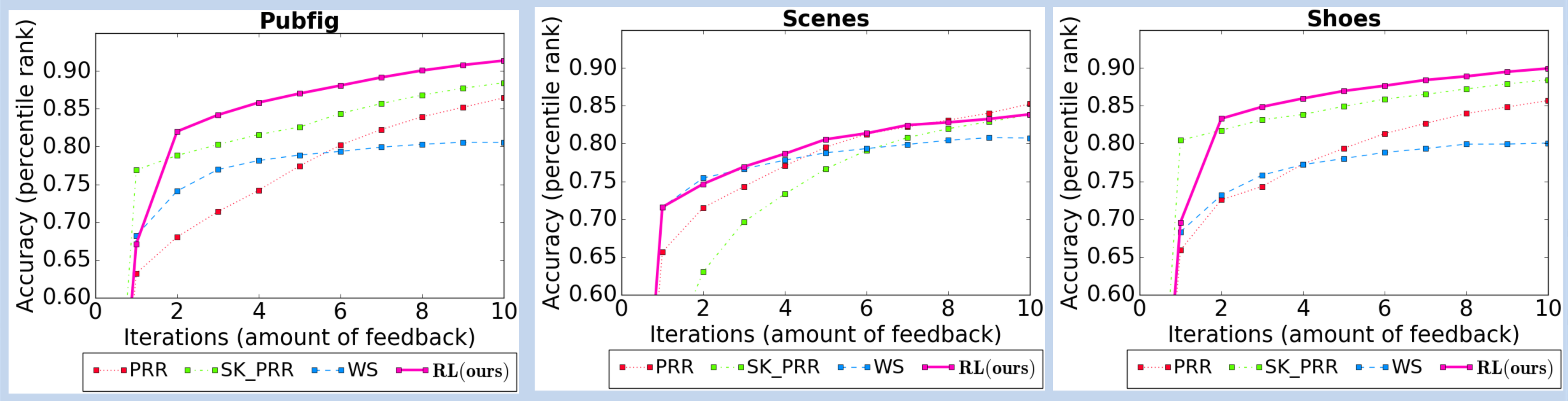


- **Learning:**
  - We use Q-learning, which receives a state and predicts the best action.
  - We employ a neural network with convolutions, because they capture information about image features and ordering.
  - We follow a **replay-memory** mechanism to collect many instances as the agent is running. It is also useful to remove short-term correlation and makes our algorithm more robust and stable.
  - We generate **random actions** with probability decreasing from 1 to 0.1 as training progresses. Random actions are useful at initial stages to *explore* the problem. Later on, this information is *exploited* by the agent.

Evaluation

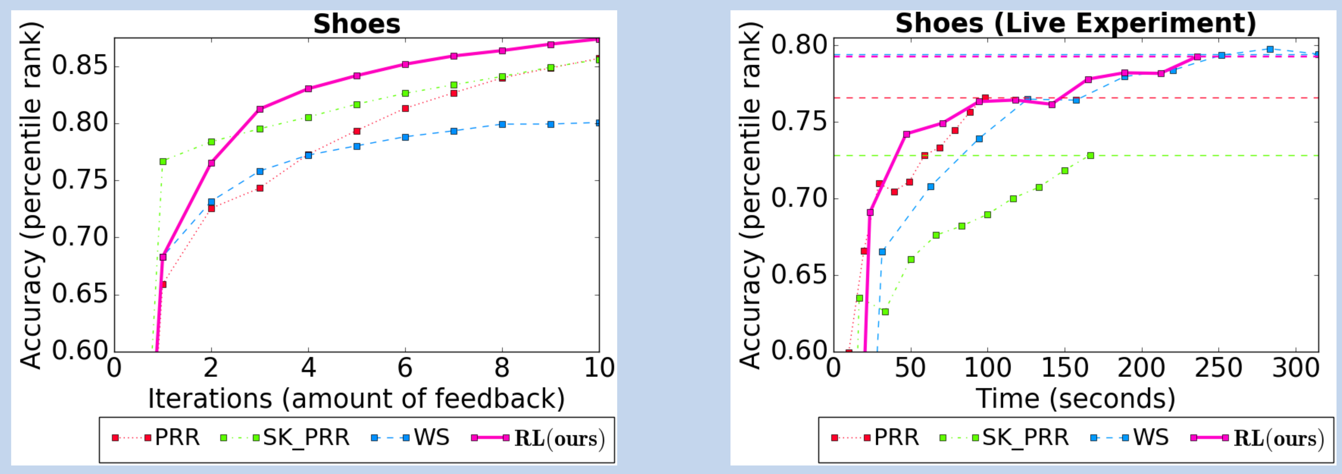
We first conduct **simulated experiments**, where we compare our reinforcement learning method with three different baselines using **percentile rank** (higher is better)**:**

- *Whittle Search (**WS**):* In each iteration, users select a (reference image, attribute) and compare target and reference for the chosen attribute ("more / less / equally").
- *Pivot round-robin (**PRR**):* In each iteration, PRR provides a (reference image, attribute) pair and users select a more / less / equally response.
- *Sketch retrieval + Pivot round robin (**SK_PRR**):* In the first iteration, we ask for a sketch. In later iterations, the system follows the pivot round-robin strategy.



We also conduct **live experiments**, where we recruit university and Amazon Mechanical Turk participants to conduct 100 searches total via a web interface.

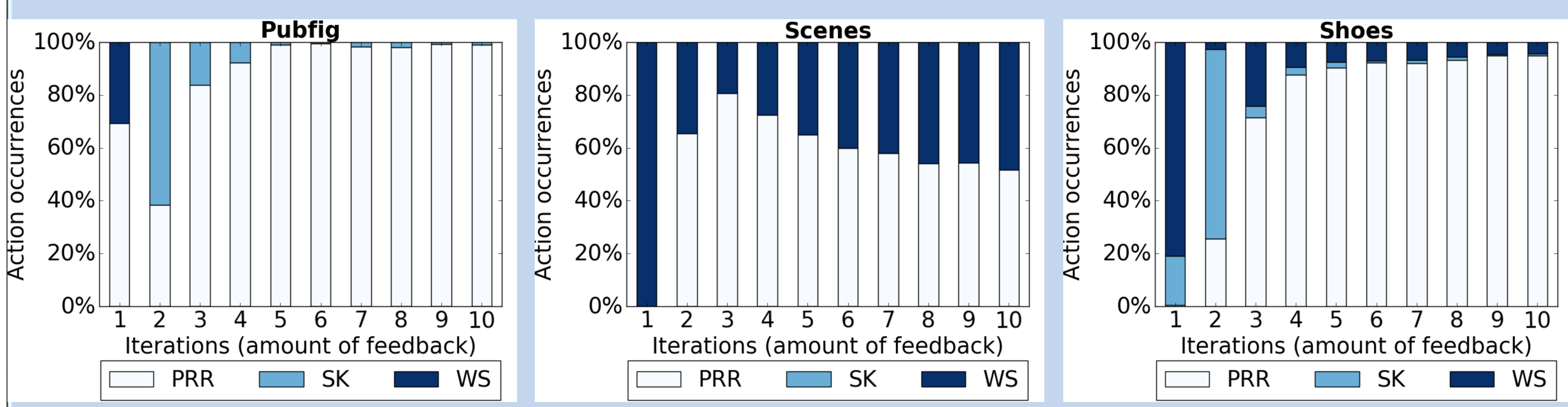
- Our RL agent queries the next action using a REST API.
- We only run experiments for Shoes because there were no appropriate sketch annotations for Pubfig and Scenes.
- We replace sketch-to-image generation with direct sketch-to-image retrieval in a joint space, to avoid memory problems due to multiple queries for GAN conversion.
- We verify the benefit of our adaptive feedback strategy in this realistic scenario.



Qualitative results

In order to understand the success of our approach, we visualize some generated sketch-to-image conversions, show the predicted actions on our test split, and finally we present some sketches provided by live users.

- From our **sketch-to-photo generated images**, we observe the most realistic ones correspond to Pubfig, then Shoes and finally Scenes. This order also corresponds to the relative performance of our method, best on Pubfig and Shoes.



- To understand our **mixed-initiative RL agent**, we count the predicted action per iteration. We observe that:
  - The SK and WS actions (sketch and free-form attribute feedback) are mainly performed in iterations 1 and 2, because they are *exploration-like* actions.
  - After iteration 3, PRR (system-chosen attribute questions) is most common.
- Our RL agent learned to prioritize **human-initiated** feedback **early on**, and complement it with **machine-initiated** feedback in **later** iterations.

User-drawn sketch:

Sketch to image:

- From our **live user sketches**, we observe that many users do a good job. For example, the sketch in column 2 has finer details such as the flower ornaments. Similarly, the boot in column 4 was drawn with laces on top and in the middle. Finally, the sneaker from column 5 contains shoelaces and details on the bottom.

Summary and contributions

We explored choosing interactions in a mixed-initiative image retrieval system. Our approach selects the **most appropriate interaction per iteration** using RL.

- Our model **prefers human-initiated** feedback in early iterations, and complements it with **machine-based feedback** requests (i.e. questions) in later iterations.
- We **outperform** standard image retrieval methods with **simulated** and **real** users.