FashionAsk: A Multimedia based Question-Answering System

Wei Zhang, Lei Pang, Chong-Wah Ngo
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
{wzhang34-c, leipang3-c}@my.cityu.edu.hk, cscwngo@cityu.edu.hk

ABSTRACT
We demonstrate a multimedia-based question-answering system, named FashionAsk, by allowing users to ask questions referring to pictures snapped by mobile devices. Instead of asking verbose questions to depict visual instances, direct pictures are provided as part of the question. The significance of our system is that (1) our system is fully automatic such that no human expert is involved; (2) and we bypass the requirement of the name for the object under query, which is mostly unknown to the asker. To answer these multi-modal questions, FashionAsk performs a large-scale instance search and metadata pooling to name the instance under query, and then matches with similar questions from community-contributed QA websites as answers. This demonstration is conducted on a million-scale dataset of Web images and QA pairs in the domain of fashion products. Asking a multimedia question through FashionAsk can take as short as five seconds to retrieve the candidate answer as well as suggested questions.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms
Algorithms, Performance, Experimentation

Keywords
Multimedia question answering, instance naming, question matching

1. INTRODUCTION
Community-contributed question-answering (cQA) websites, such as Yahoo! Answers, has gained popularity in the past few years, and accumulated hundred of millions question-answer (QA) pairs. Most of the QA pairs are text-based, where a user posts question by typing and expects text answers in return by search or by other users. Traditionally, a user can get answers from the community either by posting a question and waiting human answers or by searching past archives. The first way requires long waiting time, and the latter requires certain knowledge of the object under query (e.g., the name). However, with the popularity of mobile devices for snapping pictures, there is a shift recently on how community users post questions. For instance, instead of asking a question by providing a long text description for a visual instance, direct visual examples are included as part of the question [9, 3]. For example, asking the question “Who is the designer of this rock-and-roll t-shirt?” by providing a picture is far convenient than describing the T-shirt verbosely.

In this demo, a system, named FashionAsk, capable of answering multimedia-based questions in the domain of fashion is demonstrated. Specifically, a question is answered by first naming the visual instance in the given picture, e.g., “Twisted Sister Rock-and-Roll t-shirt”, through a large-scale search of Web images. The original text question is then augmented with the name entity, and searched against cQA archives. To this end, the system returns potential answers in text, as well as shortlists a few suggested QA pairs. Our current system operates on a dataset with 1.1 millions Web images and 1.5 millions QA pairs.

Compared to existing cQA websites, which rely on text for searching similar QA pairs, our system is novel by answering multimodal questions with visual content analysis. Compared to recent mobile search engines, such as Google Goggles and SnapTell, the developed system is capable of handling instances of non-rigid motions and non-planar surfaces, such as t-shirts, sneakers, and handbags. Further-

Figure 1: Framework of FashionAsk.

1http://answers.yahoo.com/

2http://www.google.com/mobile/goggles/#text

3http://www.snaptell.com/
more, instead of returning relevant webpages to users, our system is more specific by allowing users to expect answers and suggested questions that might directly address the information needed. Note our system is fully automatic, which does not involve human experts’ interaction. Also our system can handle unknown products by offering facilities inferring the name of the product under query, which is extremely useful in the scenario of question answering.

The rest of the paper is organized as follows. Section 2 details the method and implementation details for the system. Section 3 demonstrates the system and a brief evaluation, and finally Section 4 concludes this paper.

2. TECHNOLOGY

Figure 1 shows the framework of FashionAsk. The system is composed of two main components: Instance Naming and QA Matching. The former infers the name of a fashion instance, while the latter searches for similar QA pairs.

2.1 Instance Naming

The basic idea is to search similar images of the given instance, and then infer the name by analyzing the metadata. The main challenges are that fashion instances could appear in different forms and exist in different backgrounds, as shown in Figure 5(a). Thus, our framework considers a spatial verification of instance by triangulation matching to address the challenges, which will be briefly elaborated next. 

**Instance Search** Figure 2 depicts the retrieval model used for large scale instance search. Our model is based on Bag-of-visual-Words model [7]. First, a hierarchical vocabulary tree [4] with one million leaf nodes is trained on MIR-Flickr dataset [1]. To enable robust search, each word is indexed with its spatial location and a 32-bit Hamming signature [2]. To support efficient large-scale search, the index files are distributed to several machines. During online search, a query is matched against images in the dataset by traversing the inverted file with Hamming signature verification. For each matched image, triangulation meshes are further constructed (as Figure 3) to characterize the spatial consistency using DT [10]. This technique is different from traditional techniques such as WGC [2] and RANSAC [6], which assume a strict linear transformation model. In contrast, the triangulation-based matching does not assume such models, but locally measures the relative positioning of visual words to derive coherency scores for image ranking. The matching is found to be more effective than other methods in handling objects with non-rigid shapes and non-planar surfaces. For example, despite the flapping wings (Figure 3 (left)) violates any linear transformation, corresponding features on each local wing are still consistently depicted by the graph. Figure 3 middle and right present two additional examples for both relevant and irrelevant images. We can see the graph is insensitive to scale/rotation and small variations that remains stable topology, while being sensitive to irrelevant images with random matching locations.

**Name Entity Extraction** With the list of retrieved images, name entity extraction (NE) is processed by parsing the metadata (title and description) of each image with the Berkeley Parser [5]. The noun phrases are then extracted directly as candidate names. The likelihood of a phrase being the name of the instance is measured by its phrase frequency weighted by the coherency scores of instance search.

2.2 Questions Matching

As questions in cQA websites are asked in natural language, similar questions are varied in terms of lexical, syntactic, and semantic features. Hence, retrieving similar questions is not trivial, and requires natural language processing. For the consideration of speed, a two-stage question matching method is developed. In the first stage, the keywords extracted from user’s question together with top name candidates are used to retrieve a small set of similar questions from a QA dataset with the BM25 retrieval model, which ranks documents based on a probabilistic model. In the second stage, the original question is augmented by replacing the pronoun, such as “this”, with candidate names, and
The system can be accessed through a Web Interface:

3. SYSTEM AND USER INTERFACE

Interface: The system can be accessed through a Web browser operated either on a PC client or a mobile phone. As the interface shown in Figure 5(a), a user can either upload a picture, or pick an image randomly listed by our system, and then ask a fashion related question. Users can also crop the region-of-interest from an image when issuing questions. The returned page will show the name of the visual instance, the most likely answer, and a few suggested questions retrieved from our database, as illustrated in Figure 5(b).

System Setup: The system currently consists of 1.1 millions Web images, crawled from Flickr by querying several fashion related keywords (e.g., t-shirt, sneaker, and handbags) and 142 popular tags. Our dataset also has 1.5 millions QA pairs crawled from Yahoo! Answers, one quarter of which are related to fashion, and the others are randomly crawled as distracters. To enable fast response to user’s query, we use several machines serving the million scale instance search subsystem. For QA matching part, we implement the text processing with the help of Lucene.

Performance: Based on current setup, the performance of our system is as follows when testing on 180 multimedia questions involving 50 fashion instances: the mean Average Precision of instance search is 18.8%; the accuracy of finding the right instance name as the top-1 candidate is 21.7%; and the accuracy of returning a right answer or a related question in top-10 list is 19.8%. The system is currently run on two machines with 2.67GHz CPU and 16GB main memory.

Speed Efficiency: By current setup, instance naming consumes 1.7 seconds (including feature extraction, quantization, and search), and question matching takes about 2.9 seconds to rank similar questions. In total, the user needs to wait approximate five seconds to get the result. This time could be further reduced to sub-second, if more than ten machines are employed to parallelize instances searching and questions matching.

Interested readers can access our online video demonstration\footnote{http://www.youtube.com/watch?v=6SF2LPU4Cds} for more details.

4. CONCLUSIONS

This demo presents a first step towards large scale multimodal question answering system, which enables question asking on instances even the names are unknown. We show that with the help of instance search, questions can be asked in an much easier way such that hands can be freed from heavy typing. We also show that if only a narrow domain (such as fashion) is targeted, the performance of multimodal question answering can be quite reasonable.

5. REFERENCES