Photo-based Question Answering

Tom Yeh
MIT EECS, CSAIL
Cambridge, M.A., USA
tomyeh@mit.edu

John J. Lee
MIT EECS, CSAIL
Cambridge, M.A., USA
jjl@mit.edu

Trevor Darrell
UC Berkeley EECS, ICSI
Berkeley, C.A., USA
trevor@berkeley.edu

ABSTRACT

Photo-based question answering is a useful way of finding information about physical objects. Current question answering (QA) systems are text-based and can be difficult to use when a question involves an object with distinct visual features. A photo-based QA system allows direct use of a photo to refer to the object. We develop a three-layer system architecture for photo-based QA that brings together recent technical achievements in question answering and image matching. The first, template-based QA layer matches a query photo to online images and extracts structured data from multimedia databases to answer questions about the photo. To simplify image matching, it exploits the question text to filter images based on categories and keywords. The second, information retrieval QA layer searches an internal repository of resolved photo-based questions to retrieve relevant answers. The third, human-computation QA layer leverages community experts to handle the most difficult cases. A series of experiments performed on a pilot dataset of 30,000 images of books, movie DVD covers, grocery items, and landmarks demonstrate the technical feasibility of this architecture. We present three prototypes to show how photo-based QA can be built into an online album, a text-based QA, and a mobile application.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Systems and Software—*Question-answering (fact retrieval) systems***

General Terms

Algorithms, Design, Human Factors

1. INTRODUCTION

Many research systems [4, 32, 17, 16] and commercial systems [34, 22, 5, 6, 8, 1] have been developed to provide useful question answering (QA) services (Section 2.1). Compared to conventional keyword-based search engines, a QA service offers two unique advantages. First, it supports an interaction style more similar to how we naturally interact with other humans—stating our questions in plain sentences rather than with carefully chosen, sometimes cryptic, sets of keywords. Second, unlike keyword-based search engines that often return lengthy webpages through which users need to browse in order to locate the desired information, a QA service allows users to be specific about what they need to know so that they receive brief and concise answers directed to their particular needs.

However, current QA systems are based on text alone and can be difficult to use when questions are centered on physical objects with distinctive visual attributes. For example, a person who has just seen an interesting poster may want to ask the question *"where can I buy this poster?"* A text-based QA system would require the person to meticulously describe the visual details of the poster in order to identify it. The difficulty of this task stems from the fact that such questions are inherently dual-modal: it involves a verbal component that states its intent (where to buy) and a visual component that identifies the object (a specific poster). Unfortunately, with text as the only available input modality, users of current QA systems are often forced to express in words what would be best expressed visually.

We propose photo-based QA as a solution to the limitation of current text-based QA systems. By taking advantage of recent advances in QA (Section 2.1) and image matching technologies (Section 2.2), photo-based QA supports direct use of photos in phrasing questions and finding answers. In contrast, current text-based QA system are hard to use when visual objects are involved (Section 3.1). These problems highlight the unique usability benefits of photo-based QA (Section 3.2). Two factors play in our favor in developing useful photo-based QA systems. First, many online multimedia (i.e., image and text) data sources can supply a photo-based QA system with structured information to handle a variety of common photo-based questions automatically (Section 3.3). Second, many community human users are willing to look at photos and answer questions that the automatic process fails to find answers for (Section 3.4).

We describe a three-layer system architecture for photo-based QA. It draws inspiration from three popular QA approaches: template-matching, information retrieval, and human computation (Figure 1, Section 4). To evaluate the potential of the proposed architecture, we constructed a pilot dataset with more than 30,000 multimedia records of books, movies, grocery items, and buildings extracted from various online sources (Section 5.1) and measure the performance

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

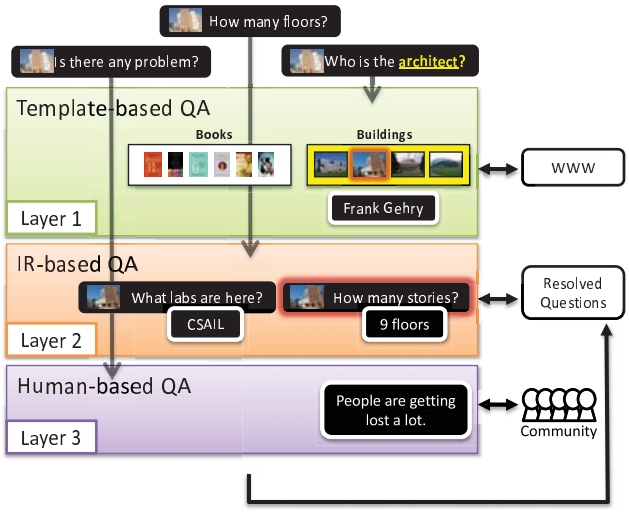

Figure 1: Three-layer system architecture for photo-based QA (Section 4). For a simple question, the template-based layer identifies its category (e.g., building), finds a matched image within the category, and forms a template to extract an answer from the Web. For a harder question, the IR-based layer searches for relevant photo-based questions already resolved. The human-based layer handles the rest of the questions too difficult for the first two layers.

of matching 600 camera phone images to these database images (Section 5.2, 5.3). We describe three prototype systems designed to demonstrate how photo-based QA can be integrated into an online photo album by introducing an extra question modality (Section 6.1), into a text-based QA service by embracing an additional photo modality (Section 6.2), and into a mobile application for ubiquitous access (Section 6.3).

2. RELATED WORK

In this section, we review research and commercial systems for question answering (Section 2.1) and image matching (Section 2.2)—two core components of our proposed photo-based QA system—and previous research efforts on combining the two technologies (Section 2.3).

2.1 Question Answering

2.1.1 Research systems

Question answering has been an active research field in which many promising systems have been created. These research systems generally adopt one of these three approaches: natural language processing (NLP), information retrieval (IR), and template matching [4]. NLP-based QA systems aim to handle questions in a way that mimics human intelligence: they fully parse a question's sentence structure, convert it to an internal semantic representation, and apply formal logic based on this representation to derive an answer for the question [15]. IR-based QA systems treat question answering as an information retrieval problem: they search a large corpus of text for specific paragraphs, phrases, or

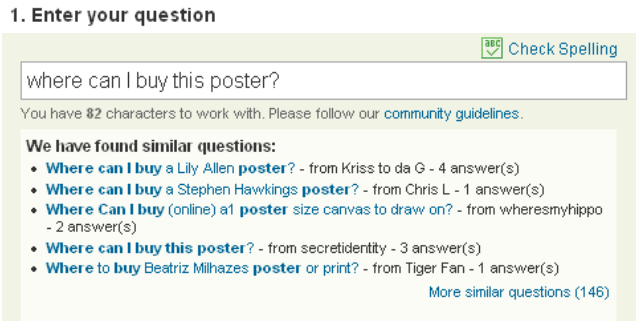

Figure 2: Asking a question on Yahoo! Answers (Section 2.1.2)

words that are relevant to a given question [32]. Template-based QA systems do not process text directly, but rather convert questions into templates and use these templates to extract knowledge from structured data stored in databases [29] or embedded in HTML tables [17]. In Section 4 we describe how we integrate IR-based and template-based QA into a coherent architecture for photo-based QA.

2.1.2 Commercial systems

Question answering has enjoyed commercial success in many real-world systems. Some systems provide answers based on private, special-purpose databases. For example, some customer service lines employ speech-based QA systems to handle common, easy questions while directing specific, more difficult requests to appropriate human agents.

Some systems offer answers based on public knowledge available on the Web. For example, Ask.com [5] is a QA-based search engine that accepts queries in natural-language sentences, instead of keywords, and retrieves a list of relevant webpages as answers. AskMeNow [6] specifically targets mobile users. It accepts SMS messages and handles questions about popular topics such as weather, sports, and stock quotes, giving answers based on information it extracts from third-party websites.

Most interesting are systems that leverage human computation to provide answers, such as the US-based Yahoo! Answers [34] and the Korea-based Naver Knowledge [22]. They create and maintain active knowledge-sharing communities where people are free to ask and answer questions about virtually any topic. What makes these systems valuable is the capability to look up relevant questions resolved in the past. Figure 2 shows a typical interface on Yahoo! Answers for entering questions. Based on the content of a question asked, it suggests several similar questions that have already been resolved by the community. These human-based QA systems are successful because they harness the community's knowledge and store a tremendous amount of useful information that can later be queried by automatic methods.

Moreover, human-based QA has been brought to mobile platform by a number of startups such as Chacha [8]. These companies handle questions sent by mobile users and deliver answers back to them in a timely manner. These answers come from qualified human researchers who have signed up with these companies to offer their services in exchange for monetary rewards. As demonstrated by these companies, the demand for mobile-based QA can not be overstated. In

Section 6.3 we present a prototype system as an example showing how mobile QA systems can also benefit from the inclusion of photos to provide photo-based QA.

2.2 Image Matching

2.2.1 Research systems

Steady progress has been made on image matching technology over the past few years. The objective of image matching is to match an input image to a large database of images and find a list of images visually similar to the input image. Several research groups have reported encouraging empirical results on matching various types of images, such as images of posters [9], magazine covers [25], video frames [27], CD covers [23], grocery items [41], and buildings [24].

For posters, [9] analyzed the color patterns and layouts of 192 synthesized poster images and matched them against 10,000 frames in a video sequence. For magazine covers, [25] extracted distinctive features called top-points from a large image of 10 to 12 overlapping magazine covers laid on a desk and then used these top-points to accurately identify each magazine in the image. For video frames, [27] used an inverted file to index 4,000 movie frames based on about 10,000 small regions (interest points) extracted from these frames, in a manner analogous to how a typical document retrieval system would index millions of text documents based on the keywords extracted from them. For grocery items, [41] applied a technique similar to [27] to match 300 images of individual grocery items against 3,000 images taken from the aisles in a local supermarket. For CD covers, [23] extended [27] by organizing the inverted file into a tree to efficiently index and retrieve 40,000 images of CD covers. For images of buildings, [24] improved upon [23] with a better method for constructing the hierarchy and added a post-verification step for enforcing geometric consistency. It achieved reasonable accuracy in matching 5,000 images of 11 landmarks on the Oxford campus despite the presence of 1 million unrelated background images.

2.2.2 Commercial systems

Several recent startup companies have explored commercial opportunities of image matching technologies on mobile platforms, such as Snaptell [28]. By working with advertisers, these companies enable mobile users to take pictures of posters or billboards, send these pictures to the advertisers, and receive special offers (e.g., coupons, ringtones) for the products or services advertised.

2.3 Question Answering and Image Matching

Photo-based QA aims to couple question answering and image matching in a meaningful and useful manner. Such coupling has been explored a number of times in the literature. [35] describes a system that answers questions about news videos, questions such as *"when will NASA resume shuttle flights?"* The system analyzes the text in the transcripts to find answers. It applies simple computer vision to group shots with similar colors together and uses the grouping information to improve the quality of the answers. [18] proposes a QA-based interface for browsing surveillance videos. This interface supports activity-related questions such as *"did any cars leave the garage?"* by deriving answers based on the activity data captured by a real-time motion tracking system. While these previous works have considered how to use QA system to access multimedia data, our current work on photo-based QA aims to further enable a QA system to understand multimodal input—for example, to enable [18] to handle the question *"did this car leave the garage?"* accompanied by an image of a car.

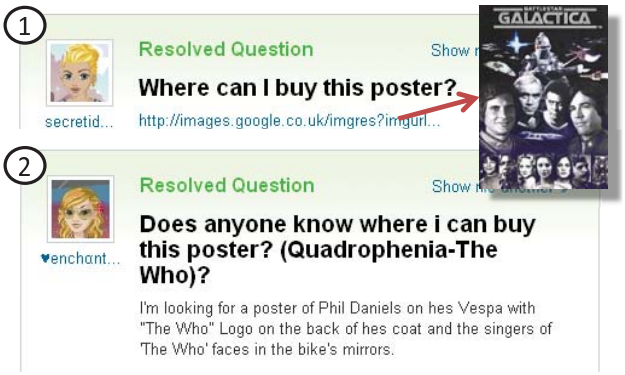

Figure 3: Asking a question about a visual item using (1) and not using (2) a photo (Section 3.1)

3. PHOTO-BASED QA

Text-only QA can be problematic for questions about visual objects (Section 3.1), a deficiency that only highlights the benefits of photo-based QA (Section 3.2). It is now feasible to build systems to handle photo-based QA by exploiting online multimedia data (Section 3.3) and by leveraging the visual knowledge of community human users (Section 3.4),

3.1 Problems with text-only QA

While users can post many questions to current text-based QA systems and find satisfactory answers, they often face challenges in phrasing their questions when it comes to visual objects. Figure 3 shows two questions we observed on Yahoo! Answers that contrasts the restriction of asking questions using only words with the flexibility of asking questions using a picture along with words. Both questions express an intent to find out where to buy a particular poster. Yet, using only text, the user asking the second question needs to provide elaborate visual details.

Human experts reading text-only questions about visual objects may also experience difficulties in interpreting them. Back to the example in Figure 3. In the first case, an expert can simply follow an URL to see the poster and immediately know which poster the question refers to. In the second case, however, an expert must draw a mental picture of a poster according to the visual description given in the question, a task much more challenging and prone to misinterpretation.

Moreover, text-only questions can undermine the usefulness of the search capability of current QA systems regarding visual items. In Figure 2, even though Yahoo! Answers is able to search its database and retrieve a resolved question identical to the input question *"where can I buy this poster?"*, there is no way to tell whether these two questions actually refer to the same poster without seeing the photos.

3.2 Benefits of photo-based QA

Problems with text-based QA highlight the potential benefits of photo-based QA. Some conventional wisdom–*a pic-*

ture is worth a thousand words–is especially true for questions. In fact, many users have posted photo-based questions on exiting systems despite the lack of any formal support from these systems. For example, text-based QA systems, such as Yahoo! Answers, have witnessed users adding URLs of photos to their questions, such as the example shown in Figure 3). Photo-centric online albums, such as Flickr [13], are full of examples of users uploading photos and typing questions in the titles or as tags. However, since these systems are unimodal and not designed specifically for this very purpose, they were unable to handle these types of questions adequately. There is clearly a need to design multiomdal systems to realize the benefits of photo-based QA. Later in Sections 6.1 and 6.2 we describe two prototypes to show how a text-based QA system and a photo-centric online album can be transformed into a photo-based QA system.

A user study we performed also affirms the benefits of photo-based QA. In this study, we showed a series of images of 10 buildings to two groups of 10 subjects, who were asked to compose an one-sentence e-mail inquiring about the architect of each building shown. Only one group was told that an image of the building would be attached to the message. Subjects in this group kept their questions simple (such as *"Who is the architect?"*), while subjects in the other group resorted to lengthy descriptions of the appearance of the building, asking questions such as *"Who is the architect of the cartoon-looking building that looks like a combination of many different shapes of the building with silver and orange colors?"* By adding the photo, subjects were able to interact more naturally and formulate their query more easily.

3.3 Multimedia data sources

An effective photo-based QA system requires a rich repository of multimedia data linking photos to some text that might be relevant to user questions. Structured multimedia data can be extracted from online catalogues maintained by merchants (e.g., Amazon.com [2]) or special-interest references (e.g., GreatBuildings.com [14]) regarding a particular category of objects sharing common attributes, such as authors (for books) or architects (for buildings). Other data sources such as Wikipedia [33] and Flickr [13] provide semi-structured multimedia data that can be extracted from a large body of text, comments, or tags surrounding images. In addition to these, we can extract even more data from unstructured sources such as news articles or blogs. In fact, several works have exploited online multimedia data to provide image-based search for relevant text such as category labels [11], product webpages [38], and animal names [7]. The high availability of these multimedia data makes the creation of a knowledge base for automated photo-based QA viable.

3.4 Visual human computation

Photo-based QA systems, like their text-only counterparts, still have problems when faced with obscure questions that have no match in the knowledge base or difficult-to-recognize images. Some text-based QA systems such as Yahoo! Answers have successfully handled these difficult questions by resorting to human computation [31], the idea of using humans to solve intelligent tasks easy for humans but hard for machines. We argue that adding photos to this scheme would not decrease people's willingness to answer questions; many people do not mind looking at photos and sharing their visual knowledge about them. In fact, several commercial systems have exploited human computation to process images, such as HotOrNot.com for rating people's beauty, ImageParsing [36] for annotating web images, or Mechanical Turk [3] for *crowdsourcing* human intelligent tasks including visual recognition. Also, research systems have found ways to get people to contribute their visual knowledge for free, such as the LabelMe [26] tool for segmenting object images and the ESP Game [31] where online players describe images as part of the gameplay. Thus, it is reasonable to expect that the current success of human computation in text-based QA can translate well to photo-based QA, especially in cases when the automated process fails to find satisfactory answers.

4. SYSTEM ARCHITECTURE

We propose a three-layer architecture for photo-based QA (Figure 1), following the QA paradigms described in Section 2.1. The first layer, based on template-based QA, handles common questions that can be answered using images and facts extracted from structured multimedia databases. The second layer, based on IR-based QA, handles harder questions the template-based QA layer fails to answer by searching a large corpus of photo-based questions resolved in the past. The third layer, based on human computation, deals with the most difficult photo-based questions the first two layers are unable to handle automatically. It relies on human computation provided by community users to recognize photos, understand questions, and find answers. Table 4 summarizes the properties of the three system layers.

4.1 First layer: Template-based QA

The goal of the first layer is to provide fast, automated answers using online resources. In a three-step process, it (1) uses the question text to determine the scope of relevant images, (2) performs image matching within that scope, and finally (3) builds a template consisting of the matched image and the original question to retrieve answers from the Web.

4.1.1 Filtering images

A prerequisite of answering a photo-based question is to identify the object in the photo, which can be accomplished by matching the photo to a large database of labeled images. However, image matching can be overwhelming in large-scale applications involving tens of thousands of images. Fortunately, questions often contain contextual information to narrow the scope of relevant images we need to search. Similar ideas have been applied in previous works that use keywords to constrain the search space for location images [40] and web images [10], while the current work explores the use of questions to simplify image matching.

We use questions to narrow the scope of image matching in two ways: *category-based filtering* and *keyword-based filtering*. Category-based filtering is applied before the image matching process begins. The database of images is pre-partitioned into categories based on the categorization scheme of the data source. For example, Amazon.com organizes its products into categories such as books, electronics, and grocery. Images extracted from each product category can thus be indexed separately. Consider the question *"is this granola bar delicious?"*, the word *delicious* may suggest that the index of the *grocery* category needs to be checked, and the index of the book category can be safely ignored.

Layer	QA Approach	Agent	Question Type	Data Source	Data Type
1	Template-based	Machine	Common	Online catalogues	Structured
2	IR-based	Machine	Specific	Past resolved questions	Semi-structured
3	Human-based	Human	All	Human knowledge	Unstructured

Table 1: Summary of the properties of the three system layers (Section 4).

Keyword-based filtering is applied after we perform an initial match and obtain a list of candidate photos. We can examine the keywords associated with each candidate photo and decide if the question is relevant to these keywords. In the above example, the candidate photos are those that look like granola bars. The words *granola bar* in the question can be used as a filter and remove those photos that are not called granola bars.

While neither filtering step is mandatory (we may fail to find any useful keywords, for example), we show in Section 5.2 and 5.3 respectively that category-based and keyword-based filtering dramatically improve image matching performance. In Sections 6.1 and 6.2 we show how these two ideas are used in our prototype systems.

4.1.2 Matching images

Having filtered the images down to a manageable scope, we can now match images within this scope. But to enable image matching, we need index the images extracted from online multimedia resources. We have previously developed an adaptive vocabulary-tree method suitable for this task [39]. An overview of this method is given as follows: To index an image, we convert it to a bag-of-features representation by identifying the Maximally Stable Extremal Regions [21] in the image and then for each region applying PCA-SIFT [19] to extract a 12-dimensional feature vector. Each feature vector is assigned to a leaf node in the vocabulary tree and adds a pointer in the node linking back to the image. Each leaf node corresponds to a visual word that represents a cluster of closely related feature vectors. When a node becomes overcrowded with features, the neighborhood containing the node is subject to reclustering, creating new nodes (visual words) to better describe these features. After all the images are indexed, stored in each leaf node will be pointers to images that contain a particular visual word.

To match an image using the vocabulary tree index, we compute the bag-of-features representation as before and for each feature lookup the relevant visual word. Each visual word casts one vote for each of the database images that it points to. Those images that receive the highest votes are considered most similar to the input image. These similar images are presented to the user, who can review them and choose the correct matches.

4.1.3 Answering questions

To answer the question after the image has been matched, we use START [16], a popular template-based QA system, that is pre-populated with structured facts extracted from various online resources of multimedia data. We generate a template based on the question text and the label of the matched photo. For example, suppose the question is "*what is the rating?*" and the user has provided a picture of the DVD cover of the movie *Ratatouille*; we derive the template "*what is the rating of the movie Ratatouille?*" and pass it to START to obtain the desired answer.

4.2 Second layer: IR-based QA

The second layer deals with the questions that can not be answered simply by extracting facts from structured multimedia resources using a template-based approach. Instead, it adopts an IR-based QA approach by maintaining an internal repository of resolved photo-based questions and retrieving from the repository those that are relevant to the current photo-based question.

Retrieving relevant photo-based questions from a repository requires a system to match both photos and questions. To match a new photo, we can apply the adaptive vocabulary tree method [39] as in the first layer. In addition, we need to save the new photo in the repository. Our method is especially attractive for this IR-based QA because of its adaptiveness: it incrementally inserts new images to the tree index and adapts the structure of the tree to better capture the visual properties of these new entries, whereas pervious methods [23, 27] often rely on a separate offline training stage to index images in batch.

To match a new question, we apply the metric described in [30] to compute the similarity score between the new question and every question related to the matched photo and retrieve those with high scores. This metric combines both lexical and semantic similarities to derive an overall similarity score between two questions. For lexical similarity, it counts how many words two questions have in common weighted by the lengths of the questions. For semantic similarity, it uses WordNet [12] to estimate the semantic closeness between every pair of the words in the two questions. While this metric has been shown to perform well on a dataset of computer-related FAQs, in Section 5.4 we evaluate its effectiveness on a dataset pertinent to photo-based QA.

4.3 Third layer: Human-based QA

The third layer handles cases when a photo-based question contains an image or text that is too difficult for the first and second layers to deal with automatically. It resorts to human computation by adopting a community-based QA model, such as Yahoo! Answers, where it awaits an expert to recognize the photo and provide the answer. When an answer becomes available, the system delivers it to the user. Moreover, the system adds the photo-based question along with the answer into the repository so that the second layer can serve similar requests in the future.

5. EVALUATION

In this section, we evaluate core components of our photo-based QA system described above based on a pilot multimedia dataset of more than 30,000 images (Section 5.1) [1]. We have identified the primary technical challenges to be in accurately matching query images to large databases of images using both categories and keywords, (first layer) and in retrieving questions in the database that match the

[1]To obtain this dataset, please contact the authors.

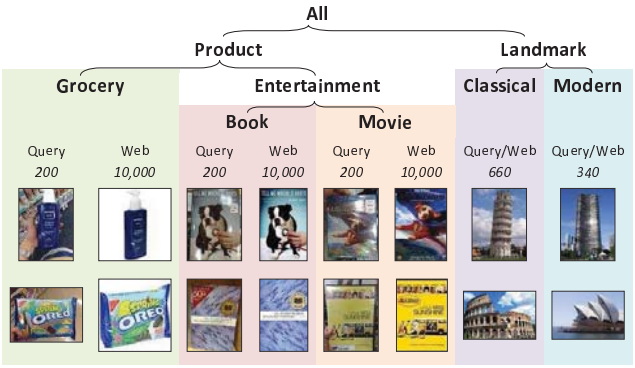

Figure 4: Images in our pilot dataset (Section 5.1)

Category	Sample Questions
All	What is it?
Product	Is it in stock? How much is this on Amazon? Where can I buy it?
Entertainment	What is its rating? What is its review? Is there a sequel?
Book	Is it a fiction? Is there a paperback edition? Who is the author?
Movie	Is there a blue-ray edition? What is its boxoffice? Who is the director?
Landmark	Where is it? Who is the architect? When was it built?

Table 2: Sample questions answerable by the template-based QA layer. (Section 5.1). Questions of a category also apply to its subcategories.

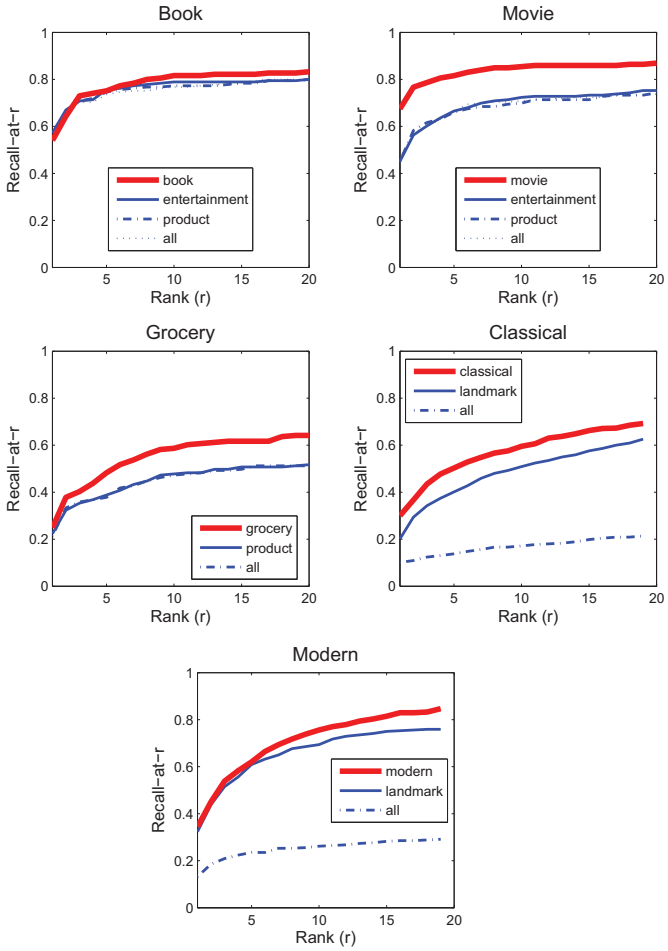

Figure 5: Image matching performance on our pilot dataset and the effects of category-based filtering (Section 5.2).

user's question (second layer). For each component, we describe evaluation methodology and report on the image retrieval performance under category-based filtering (Section 5.2) and keyword-based filtering (Section 5.3), and question matching performance (Section 5.4).

5.1 Databases

To populate our database, we collected five categories of images: books, movies, groceries, modern landmarks and classical landmarks, from sources that contain multimedia documents (text with images) with structured information. The book and movie data were collected from Amazon [2], where we retrieved 10,000 images for each category. We also collected 10,000 images of grocery items from Amazon [2] and Koamart [20] and 1,000 images of 100 famous landmarks (1/3 modern, 2/3 classical), 10 per landmark, from GreatBuildings [14] and Flickr [13].

While the image database consists mostly of stock photos of products, we populated the test set by visiting local supermarkets, bookstores, libraries, movie rental stores, and dorm rooms, using a camera-phone to take pictures of particular items. We collected about 200 images each of books, DVD covers of movies, and grocery items in this manner. Because the landmark photos in the database themselves were regular photos (rather than stock product images), we used those images to query the database, ignoring self-matches. The test set was labeled by hand, matching each test image to the stock photos in the database. Because the database images often contained multiple instances of the same object, for example, different editions of the same book, some test images had multiple labels.

We organized these categories into a taxonomy shown in Figure 4. For each category in the taxonomy, we built a set of question templates targeting the structured information that can be extracted from pertinent multimedia data sources. Table 5.1 lists a few sample questions for each category the template-based QA is expected to handle.

Books	Movies	Groceries	L-Classical	L-Modern
75.1%	81.6%	48.3%	47.7%	58.2%

Table 3: Recall-at-5 for each category (Section 5.2).

5.2 Category-based filtering

In this experiment, we evaluated the performance of the image matching algorithm described in Section 4.1.2 on the

Figure 6: Sample image matching results (Section 5.2).

Keyword	# Matches	No filtering	With filtering
Grocery items			
chocolate	828	66.6%	100%
rice	602	61.5%	84.6%
cereal	308	58.3%	100%
granola	198	71.4%	81.0%
kellogg	122	50.0%	100%
Books			
novel	445	72.2%	86.1%
story	423	81.8%	100%
Landmarks (classical only)			
cathedral	70	40.0%	68.6%
tower	40	60.0%	92.5%
Landmarks (modern only)			
tower	70	50.0%	87.1%
hall	40	65.0%	95.0%
Landmarks (all)			
tower	110	40.9%	80.0%
hall	60	70.0%	98.3%
house	60	53.3%	95.0%

Table 4: Recall-at-5 for several groups of query images when the set of database images is narrowed via keyword-base filtering (Section 5.3).

pilot dataset and measured the effects of category-based filtering. We constructed an adaptive vocabulary tree to index the images in the dataset. For each test image, we retrieved the top 20 matches. We measured performance by the recall-at-r, which is the rate at which at least one correct match appears among the top r matches. This metric was chosen because the image matching technique we used is designed to efficiently return the top r matches, instead of giving the exact ranking of all images in the database. For each leaf category (e.g., book), we used its test images to query a vocabulary tree trained on all the training images in the dataset. To examine the effect of category-based filtering, we repeated this process with smaller trees trained on subcategories down the hierarchy. For example, for the grocery category, we tested it against trees trained on images of all, product, and grocery categories respectively.

Table 3 shows the recall-at-5 for the five leaf categories in the pilot dataset, where the vocabulary tree was constructed using only database images from the same category. The vocabulary trees were able to recognize some categories more easily than others; for example, recognition rate of movie covers was much higher than that of classical landmarks (81.6% vs. 47.7%) because of the presence of more distinctive features and regularity in visual layouts, suggesting that this system would be better tuned to applications involving these kinds of products than for buildings.

Figure 5 displays, for each leaf category, the recall performance as a function of r (the number of top matches shown in the result) under different filtering conditions. The recall performance was the lowest without any filtering, when the system had to match images against the largest vocabulary tree containing all the images (lowest curves). With increased levels of category-based filtering, the recall performance went up as a result of matching against smaller vocabulary trees (higher curves). The best performance occurred when category-filtering narrowed down to the leaf category itself (highest, red curves). Performance improvements provided by category-filtering illustrate the benefits of integrating a QA system with an image matching system: on the one hand, the QA system uses questions to identify relevant categories to ease the burden of image matching. On the other hand, even if the QA system fails to find relevant categories, image matching still achieves reasonable accuracy, which can in turn benefit the QA process.

Figure 6 shows some sample results of matching query images against all categories of images. Some cases can clearly benefit from category-filtering. For example, for the *mini-wheats* test image (row 4), database images of landmarks and books would not have appeared in the result were the system able to determine that the question was about a grocery item and focus only on grocery items.

5.3 Keyword-based filtering

In this experiment, we evaluated the benefits of filtering with a particular keyword on the performance of image matching. As described in Section 4.1.1, the difference between category-based and keyword-based filtering is that the former performs the filtering before querying the vocabulary tree (in essence, querying a different database), while the latter performs the filtering after the results are retrieved. We tested with fourteen keywords people might include in their questions about certain categories, for example, *granola* for the grocery category. For each keyword, we selected all test and training images relevant to that keyword. Then, we used each test image to query the relevant database containing all images of that category and removed the retrieval results that did not match the given keyword.

Table 4 shows the recall-at-5 for several keywords, with and without the keyword filtering step. It shows, for each

keyword, the number of database images that matched that keyword. The number of matches was generally much smaller (less than 1000) than the total number of database images per category (10,000); thus, many incorrect matches were easily removed from the results. The effect this had on the retrieval performance was significant: for many of the keywords, the recall-at-5 was 100% after filtering.

5.4 Question matching

In this experiment, we evaluated the effectiveness of the question matching algorithm described in Section 4.2 applied to photo-based questions. Question matching is the basis of the IR-layer and takes place after a user has selected a correctly matched image. The success rate of question matching directly determines the performance of the IR-layer in retrieving relevant past questions, when the template-layer fails to produce an automated answer.

The experiment setup was as follows: We began with a seed set of test questions consisting of 25 unique *base* questions in English about photos, based on the actual questions contributed by the subjects who participated in our previous study [37]. In addition, we recruited 10 bilingual subjects, each of whom provided a variant for each of the 25 base questions. Thus, we obtained a total of 250 variants, 10 variants per base question. These variants were collected in a way to minimize the bias that would have been introduced if we simply presented the base question to subjects and told them to ask the question differently. Instead, we asked a subject to translate a base question to a foreign language. Then, we presented the translated version of the question to another subject and told the subject to ask the question in English, which was then shown to yet another subject to translate to a foreign language, and so on. This methodology ensured that no subject was exposed to the original question in English when asked to provide its variant.

We performed 100 simulated runs of 25 unique questions being asked by different users about the same photo. At each step of a run, a variant v of a base question b was randomly drawn and inserted into the database D, as if someone had asked v about the photo. After the insertion, we tested whether our system could retrieve v if someone else asked the same question but in a different way, namely, another variant $\bar{v}$ of b was used to query D. We repeated this test on all the variants of the questions already in D up to the point, and measured the average rank of the correct variant v in the result and the success rate of v showing up as the top match (recall-at-1) or among the top five matches (recall-at-5).

Figure 7 (left) shows as more unique questions are asked about the same photo (x-axis), it becomes more difficult to retrieve a matching question, as indicated by higher average ranks of the correct matches (y-axis). However, we note that in reality, the number of unique questions asked about a single photo would generally stay small; if there are 25 questions, for example, the correctly-matched question still usually appears within the top 3. Figure 7 (right) shows the recall-at-1 and recall-at-5 retrieval performance. While showing the top 5 results gives a significant performance boost to the retrieval, the correct match still appears as the first result most of the time.

6. APPLICATIONS

Photo-based QA can potentially impact a wide-range of existing applications. We have developed three prototypes

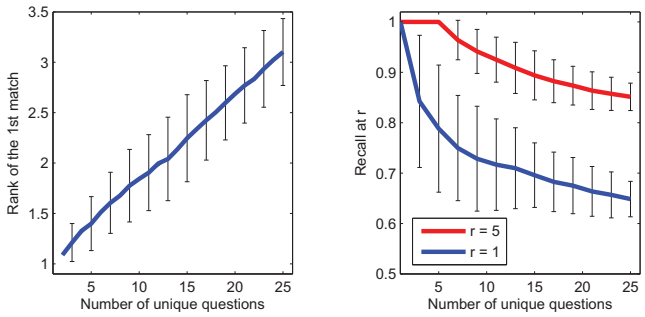

Figure 7: Question matching performance (Section 5.4)

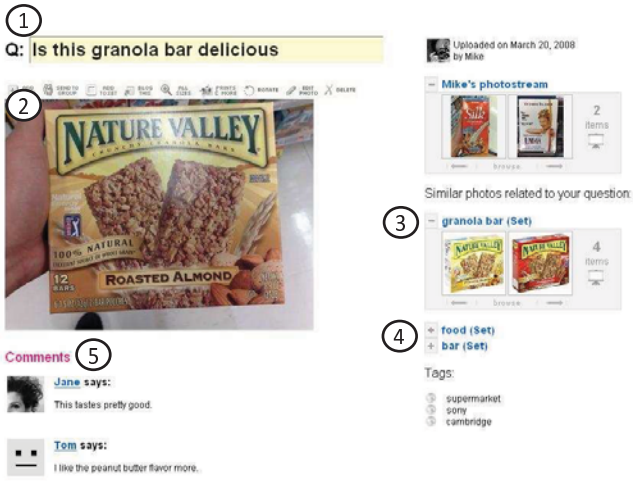

Figure 8: QA-enabled Flickr albums (Section 6.1)

to demonstrate (1) how a photo-centric service (e.g., Flickr) can introduce a QA component to support photo-based QA, (2) how a text-centric QA system (e.g., Yahoo! Answers) can incorporate a photo modality to strengthen its service, and (3) how a mobile platform can support photo-based QA.

6.1 Flickr

This prototype demonstrates how a photo-centric service can extend its capability to support photo-based QA. The interface of this prototype is modeled after a typical screen for viewing a photo on Flickr [13]. In Figure 8, a user named Mike wants to ask *"is this granola bar delicious?"* (1) using a photo he just uploaded (2). The system applies *keyword-based filtering* based on the question text to find photos similar to Mike's photo and arranges the search results as Flickr photo sets. Mike clicks on the set *granola bar* (3) and sees the images of the top two matches. If he can not find a good match, he can scroll to the right or launch a slideshow to view the other matches in the same set. Alternatively, he can check the matches in the other sets (i.e., food and bar) (4). When Mike moves the mouse cursor over a matched photo, a popup window appears to offer him the best automatic template-based answer (first layer). Unsatisfied, Mike can instead click on the matched photo to view a list of relevant questions resolved in the past (second layer). Still

Figure 9: Photo-based Yahoo! QA (Section 6.2)

unhappy about what he has found, the comment section (5) provides an opportunity for him to receive human-based answers from the community (third layer).

6.2 Yahoo! Answers

This prototype demonstrates how a text-centric QA system can be extended to support photo-based QA. The interface of this prototype is modeled after Yahoo! Answers' current interface for entering questions. In Figure 9, a user selects an image of a book (1) and uploads it to the server. Then, she starts to type her question *"what is the rating of this book?"* (2). As she types, the system analyzes the typed text and applies *category-based filtering* to suggest a number of categories related to the text (3). She selects the category *Entertainment*, which in our pilot dataset includes the movie and book categories. The image matching engine then devotes its effort to only images of movie and book covers, ignoring the images of other categories. It finds a list of matched photos (4) and the correct match appears at the second position in the list. By clicking on the second match, she is presented with a template-based answer that tells her the rating of this book is 5.0 on Amazon.com (5). She still has the option to view a list of resolved questions relevant to her current question found by the IR-based QA layer (6) or resort to human-based QA by posting her question to the forum and waiting for an answer (7).

6.3 Mobile application

This prototype demonstrates how photo-based QA can be provided on a mobile platform. In Figure 10, a user takes a picture of a building (1) and types a question *"who is the architect?"*. She sends the request to a server. Delivered back to her is a list of similar building images (3) and other questions similar to hers (4). She spots a question relevant to her question (5) and clicks into it to view the actual photo someone else has submitted and asked a question similar to hers. Looking at the photo, she is intrigued by the presence of a police car and decides to ask a new question *"what is the police doing there?"*. Luckily for her, the system has found a similar question asked by someone else and already resolved; it immediately shows her the answer (7).

7. CONCLUSION

We presented *photo-based question answering*, a multimedia system that combines the recent technical achievements in question answering (Section 2.1) and image matching (Section 2.2). We motivated the development of photo-based QA systems by highlighting the problems with text-based QA systems (Section 3.1) and demonstrating the usability benefits for these systems to understand images (Section 3.2). We argued that building such photo-based QA systems is feasible because not only can we harness online multimedia data to handle common photo-based questions automatically (Section 3.3) but also it is possible to exploit human computation to deal with harder cases (Section 3.4). We proposed a three-layer system architecture based on template-based, IR-based, and human-based QA (Section 4). To evaluate the technical feasibility of this architecture, we constructed a dataset (Section 5.1), and used this dataset to demonstrate the effectiveness of category-filtering (Section 5.2), keyword-filtering (Section 5.3), and question matching (Section 5.4). We concluded with three prototypes to show the practicality of photo-based QA in real-world applications (Section 6). As future works, we will evaluate the usability of each prototype system. On the technical side, we will explore newer image matching techniques and test them with datasets with more varieties of object.

8. REFERENCES

[1] 199QUERY: text any question and get the answer by SMS! http://www.199query.com/.

[2] Amazon.com: Online shopping. http://www.amazon.com/.

[3] Amazon Mechanical Turk. http://www.mturk.com/.

[4] A. Andrenucci and E. Sneiders. Automated question answering: review of the main approaches. In *ICITA '05*, volume 1, pages 514–519, 2005.

[5] Ask.com search engine - better web search. http://www.ask.com/.

[6] AskMeNow - get answers with search designed for mobile. http://www.askmenow.com/.

[7] T. L. Berg and D. A. Forsyth. Animals on the Web. In *Proc. of CVPR '06*, pages 1463–1470, 2006.

[8] ChaCha: Good answer. http://www.chacha.com/.

[9] C. Y. Chen, T. Kurozumi, and J. Yamato. Poster image matching by color scheme and layout information. In *Proc. of ICME '06*, pages 345–348, 2006.

Figure 10: Mobile photo-based question answering (Section 6.3)

[10] X. Fan, X. Xie, Z. W. Li, M. J. Li, and W. Y. Ma. Photo-to-search: using multimodal queries to search the web from mobile devices. In *Proc. of MIR '05*, pages 143–150, 2005.

[11] L. Fei-fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proc. of CVPR Workshop '04*, pages 178–178, 2004.

[12] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[13] Flickr: Photo sharing. http://www.flickr.com/.

[14] ArchitectureWeek great buildings collection. http://www.greatbuildings.com/.

[15] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, NJ, USA, 2000.

[16] B. Katz. Annotating the world wide web using natural language. In *Proc. of RIAO '97*, 1997.

[17] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, and B. Temelkuran. Omnibase: Uniform access to heterogeneous data for question answering. *Natural Language Processing and Information Systems*, pages 230–234, 2002.

[18] B. Katz, J. Lin, C. Stauffer, and E. Grimson. Answering questions about moving objects in surveillance videos. In *Proc. of AAAI Spring Symposium on New Directions in QA*, 2003.

[19] Y. Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. In *Proc. of CVPR '04*, volume 2, pages 506–513, 2004.

[20] Asian online market and groceries superstore. http://www.koamart.com/.

[21] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. of BMVC*, volume 1, pages 384–393, 2002.

[22] Naver knowledge search. http://kin.naver.com/.

[23] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. of CVPR '06*, volume 2, pages 2161–2168, 2006.

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of CVPR '07*, pages 1–8, 2007.

[25] B. Platel, E. Balmachnova, L. M. J. Florack, and Ter. Top-points as interest points for image matching. In *Proc. of ECCV '06*, pages 418–429, 2006.

[26] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025*, 2005.

[27] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of ICCV '03*, volume 2, page 1470, 2003.

[28] SnapTell: Image recognition based mobile marketing. http://www.snaptell.com/.

[29] E. Sneiders. Automated question answering using question templates that cover the conceptual model of the database. In *Proc. of NLDB '02*, pages 235–239, 2002.

[30] W. Song, M. Feng, N. Gu, and L. Wenyin. Question similarity calculation for faq answering. In *Proc. of SKG '07*, pages 298–301, 2007.

[31] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of CHI '04*, pages 319–326, 2004.

[32] E. Voorhees. The TREC-8 question answering track report, 1999.

[33] Wikipedia: the free encyclopedia. http://www.wikipedia.com/.

[34] Yahoo! Answers. http://answers.yahoo.com/.

[35] H. Yang, L. Chaisorn, Y. Zhao, S. Y. Neo, and T. S. Chua. VideoQA: question answering on news video. In *Proc. of ACM MM '03*, pages 632–641, 2003.

[36] B. Yao, X. Yang, and S. C. Zhu. Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *Proc. of EMMCVPR '04*, pages 169–183, 2007.

[37] T. Yeh and T. Darrell. Multimodal question answering for mobile devices. In *Proc. of IUI '08*, 2008.

[38] T. Yeh, K. Grauman, K. Tollmar, and T. Darrell. A picture is worth a thousand keywords: image-based object search on a mobile platform. In *Proc. of CHI '05*, pages 2025–2028, 2005.

[39] T. Yeh, J. J. Lee, and T. Darrell. Adaptive vocabulary forests br dynamic indexing and category learning. In *Proc. of ICCV '07*, pages 1–8, 2007.

[40] T. Yeh, K. Tollmar, and T. Darrell. Searching the web with mobile images for location recognition. In *Proc. of CVPR '04*, volume 2, pages 76–81, 2004.

[41] Y. Zhang, L. Wang, R. Hartley, and H. Li. Where's the weet-bix? In *Proc. of ACCV '07*, pages 800–810, 2007.