A Hierarchical Approach to Practical Beverage Package Recognition

Mei-Chen Yeh and Jason Tai
Department of Computer Science and Information Engineering
National Taiwan Normal University, Taipei, Taiwan
myeh@csie.ntnu.edu.tw, whiteorange0617@gmail.com

Abstract. In this paper we study the beverage package recognition problem for mobile applications. Unlike products such as books and CDs that are primarily packaged in rigid forms, the beverage labels may be attached on various forms including cans and bottles. Therefore, query images captured by users may have a wide range or variations in appearance. Furthermore, similar visual patterns may appear on distinct beverage packages that belong to the same series. To address these challenges, we propose a fast, hierarchical approach that can be used to effectively recognize a beverage package in real-time. A weighting scheme is introduced to enhance the recognition accuracy rate when the query beverage is among flavor varieties in a series. We examine the development of a practical system that can achieve a fairly good recognition performance (93% accuracy rate using an evaluation set of 120 images) in real-time.

Keywords: Product recognition, mobile application, sub-image retrieval

1 Introduction

A technology that enables customers get product information in an easy, fast, and intuitive manner is essential for cost-conscious shopping [9]. With developments in handsets that have increased the computing and communication capabilities, content-based product recognition is a complementary approach to existing technologies such as Barcodes [14] and Radio Frequency Identification (RFID) [10], in which a tag is required to be attached to each item for identification. A primary advantage of the content-based approaches over tag-based approaches is the fact that recognition can be performed directly from any part of the content—not necessarily barcodes or RFID tags—with a device that may not be equipped with a tag reader.

Among many products of interest, we study the beverage package recognition problem in this paper as the beverage industry has been continuing to be an economic powerhouse [4]. The beverage package recognition problem can be considered a simplified object recognition problem because the patterns of beverage packages are more structured and rigid comparing to those of other objects such as human faces. However, query photos captured by users can still have a wide range of variations in appearance because beverages can be packed in various forms, e.g. boxes, can, and bottles, and a package can be arranged in any angle to users. Figure 1 (a) illustrates three examples of a Coke can. The most recognizable part (e.g. the brand logo) could be fully or partially captured, or totally invisible on the query image. Moreover, it is

common in package design that distinct beverage products in the same series share similar visual features. In Fig. 1 (b), these images have identical visual patterns, e.g. brand names printed with the same font. However, they should be considered distinct beverages as they are differently flavored, and the price and calorie information may be different. These factors make the beverage package recognition a challenging task.

Motivated by recent successes in sub-image retrieval [11][16], we formulate the beverage package recognition problem as a sub-image retrieval problem where two images are matched even if only *a portion* of them are similar. The query is compared to a collection of panoramic images which are unrolled and scanned beverage labels extracted from various package forms. For example, Fig. 2 (a) shows the panoramic image for a Coke can. By using panoramic images, we need only one reference image for each beverage item. The recognition of query images can then be performed by finding the most similar image in the reference dataset based on a similarity measurement that aggregates patch-to-patch similarities. Thus, two partially similar images can be considered matched.

To solve the problem where two distinct beverage packages share similar visual features, we propose a query-dependent reweighting scheme of local features to cumulate similarities between two images only from critical regions. This is derived from the observation that a keypoint's discriminative power may vary given different contexts. For example, the brand name Fanta in Fig 1 (b) is useful for recognizing Fanta series, but is not discriminative for identifying the beverage among flavor varieties in the same series.

In the remainder of the paper, we first describe related work in this field. Section 3 presents a new dataset for beverage package recognition. We then present the coarse-to-fine filtering approach for recognizing beverage packages in Section 4 and, finally, demonstrate the performance and conclude the paper with a short discussion summarizing our findings.

(a) (b)

Fig. 1. Characteristics of the beverage package recognition problem. (a) Three examples of a Coke can. The most recognizable part (e.g. the brand logo) may be fully or partially captured, or totally invisible on the query image. (b) Beverages in a series: they have common visual elements in the package design.

2 Related Work

Recent works have shown some successes of product recognition using local feature based visual searches [15][17][18]. For recognizing products such as books and CD

covers, there are some mobile image recognition systems on the market [1][2][3]. For local-feature-based methods, a query image is represented by a set of local features and a reference image is retrieved if it has sufficient matches of local features with the query image. More specifically, two major components—local features and image matching—are crucial to the product recognition performance.

Robust and invariant local features such as Scale-Invariant Feature Transform (SIFT) [13] and Speeded Up Robust Features (SURF) [5], are applicable for mobile search applications. These descriptors are in general high dimensional feature vectors, e.g. a SIFT feature has 128 elements. Recently, Chandrasekar *et al.* proposed the Compressed Histogram of Gradients (CHoG) [7] which captures gradient statistics from local patches in a histogram and applies tree coding techniques to compress the histograms into low bit-rate feature descriptors. In [6], an experimental study on local feature descriptors for mobile visual search compares MPEG-7 image signatures, CHoG, and SIFT and concludes that SIFT and CHoG outperform MPEG-7 image signatures greatly in terms of feature-level and image-level matching. Since our beverage package recognition system is a client-server based visual search system and the main recognition task—the computational intensive part—is performed on a server, we develop both the SIFT and SURF representations and will compare their effectiveness in the experiments.

To accelerate the matching process, the dataset is usually organized and indexed. When searching for similar instances for a query, only a small fraction of the dataset needs to be examined. Since features have a high dimensionality, classical methods such as KD-trees and its variants [8] often suffer from the "curse of dimensionality". More recently, vector quantization and local-sensitive hashing (LSH) techniques have been popularly adopted to build a visual vocabulary of image features or to partition the feature space [15][12]. In this work, we adopt a LSH-based method [12] because of its simplicity in concept and its effectiveness for speeding up the recognition process.

3 Beverage Image Dataset

We introduce a beverage package image dataset which currently contains 60 reference images. The dataset will continue to grow. Each is a panoramic image by manually unrolled and scanned beverage labels extracted from various package forms. Figure 2 (a) shows a few examples in the dataset. The reason why we built our own dataset in this manner is because unlike other products, the beverage packages have various forms (e.g. bottles, cans, and boxes). Therefore, most of the beverage package images available on the web cannot meet our requirement, i.e. the label must be fully expanded and captured. Although the use of panoramic images compresses information of a 3D object into a 2D image, as we will shown in the experiment, the point correspondences can still be built by using robust local features. Furthermore, unlike the case in the general object recognition tasks where one category has multiple images, we require only one reference image for each beverage.

These beverage package images are essentially different from general images. The following differences are observed from our samples. Firstly, one or multiple text

lines are present on the container which gives an indication related to the content of the package, such as the brand name, the product name, nutrition table, and etc. The texts are usually highlighted with a distinguishing appearance from the background. Secondly, symbolic patterns and cartoon-like figures are commonly used for the graphics design to deliver the look of freshness and delicacy. These observations are useful for identifying the visual elements for beverage package design, and upon which we illustrate why a SIFT-based representation is effective for beverage package recognition.

Fig. 2. (a) Three examples of our reference images. We have one reference image for each beverage; (b) Testing images. A package is captured in three different settings. Please refer to texts for details.

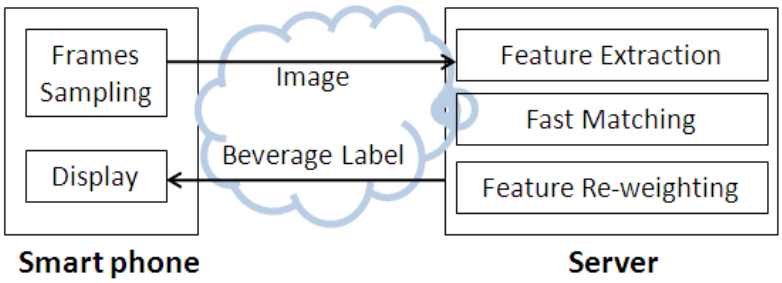

Fig. 3. The framework of our beverage package recognition system. Using a client application on a smartphone, video frames are sampled and sent to a processing server that recognizes the query image.

4 Approach

4.1 The Hierarchical Framework

Our beverage package recognition system is a client-server based visual search system as illustrated in Fig. 3. As described in Section 1, we formulate the beverage package recognition problem as a sub-image retrieval problem given a query image

captured from a cell phone and a set of reference beverage package images. Since the query image may be similar to only a portion of its reference image, global feature based methods is not applicable in our application. Instead, the retrieval can be achieved by matching two images represented by local keypoints and their descriptors. Pairwise comparison among local keypoints can further measure the degree of overlapping between two images. Due to intensive computations of the feature extraction and matching processes, the recognition task is usually performed on a server.

However, not every local keypoint has equal discriminative power. For example, keypoints that capture the recycling symbol are informative, but may not be discriminative as the symbol would appear on various product containers. More importantly, *a keypoint's discriminative power may vary* given different contexts. For example, keypoints that describe the brand name "Fanta" are useful to differentiate Fanta soft drinks from others. However, these keypoints are not discriminative for identifying a particular flavor among the Fanta series. These observations are valid especially in beverage package recognition as products in the same series tend to share some common visual patterns.

Therefore, we propose a hierarchical approach which firstly performs a coarse recognition and determines the context for a refinement search. This can be achieved by using conventional keypoint matching techniques. If the coarse search returns more than on potential matches—it usually happens when the query belongs to a series, we then apply a refinement step that adjusts weights of local features under comparison to refine similarities from those matched keypoints. We now describe the approach in details.

4.2 Coarse Recognition

The first step in the hierarchical approach is designed to identify potential matches and to filter irrelevant images. Conventional image matching approaches for recognizing rigid objects (e.g. books, CD covers) [15][17][18] can be applied. For image representation, we particularly choose SIFT-like descriptors because they are constructed by summarizing the gradient information within a local region. They can capture unique edge patterns and unique local neighborhoods. These characteristics are suitable for describing symbolic patterns and cartoon-like figures which are widely used in beverage packaging design.

We now show an analysis of the SIFT and the SURF descriptor distributions of our reference image dataset using the visualization approach described in [12]. The approach aims to sketch the space of high dimensional local features by using an approximate nearest neighbor probing scheme based on 2-stable locality-sensitive hashing. Each feature is indexed to a bucket by the hashing scheme and the indexing result yields a visualization of the distribution of feature vectors—a distribution that is peaked or has a small entropy value implies that the feature is less descriptive. Figure 4 shows the SIFT and the SURF feature distributions extracted from our beverage package image dataset. The entropy values 5.9 (SIFT) and 4.18 (SURF) of the distributions indicate that both features' descriptiveness is fairly good. For example, the entropy value of the SIFT distribution extracted from the Berkeley natural images

is 4.11, and the entropy value is 2.38 when SIFT features are extracted from noise patches, as reported in [12].

We identify the potential matches by examining if the number of patch correspondences between a query and a reference image exceeds a pre-defined threshold. If more than one reference images are returned in this step—mostly when the query belongs to a beverage series—we proceed to the refinement search step.

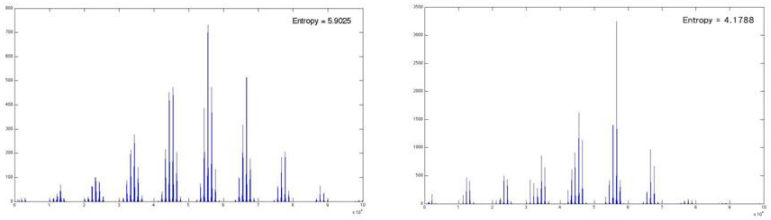

Fig. 4. The feature vector distribution over buckets: (a) SIFT (b) SURF. The entropy values 5.9 and 4.2 imply that SIFT-like local features yield a more informative feature space for beverage package images.

4.3 Refinement Search

As beverages in a series usually share common visual elements in packaging design, the similarity between the query and the candidates indexed by k returned in the previous step should be re-calculated from only the discriminative regions. We now assign a weight $w(p_i)$ to each keypoint p_i that estimates the likelihood of belonging to a particular beverage package:

$$w(p_i) = \frac{N - \sum_k t_{i,k}}{N - 1}, \quad (1)$$

where N is the number of candidates, and $t_{i,k}$ is a binary variable that represents the presence of absence of a keypoint in the k-th candidate image that matches p_i. Intuitively, $w(p_i)$ relates to the occurrence frequency of p_i in candidate images. For example, $w(p_i) = 1$ if p_i is matched to keypoints that exactly appear on only one candidate image, and $w(p_i) = 0$ if all candidate images have matched keypoints. Figure 5 shows three query examples with distinct flavors. The weights of each keypoint are coded in red dots. Light color indicates more discriminative power. It is interesting to observe that most discriminative keypoints are located at their unique regions (fruit patterns in the example) while the weights of common parts, e.g. the beverage name, are assigned with a smaller value (represented with dark dots).

The weights are assigned in a similar manner to the term frequency–inverse document frequency (tf-idf) approach. However, we did not use a visual vocabulary, and, more importantly, the weights were not pre-computed from the whole reference image dataset. To save the computation, we applied a LSH based fast matching approach [12] to build the keypoint correspondences, and identified the beverage if the ratio of the top two weighted similarities is above a threshold.

Fig. 5. Keypoints and their weights. Light color indicates more discriminative power.

5 Experiments

We present two experiments to evaluate the effectiveness of the proposed approach in recognizing beverage packages. The first experiment evaluates the overall recognition performance. To mimic the scenario how the approach will be used in practice, we captured the query images in a popular chain of convenience stores using an iPhone. Note that these query images were collected in a very different way than that of the reference images. We took three images for 40 randomly selected beverages, resulting in a testing set of 120 images. If a beverage is packed in a can or a bottle, the images were snapped with the brand name (or logo) fully, partial visible or totally invisible. For boxes, we adjusted the zoon-in factor and obtained one that contains only a portion of the box, one that captures the full box, and one that contains the box under recognition and parts of its nearby beverage packages. Figure 2 (b) illustrates a few examples.

Each image in the testing set is used as a query, and at most one beverage is retrieved. Table 1 summarizes the recognition performance. Our recipe that combines a keypoint matching method with a query-dependent weighting scheme achieves promising performance in both accuracy rate and computational speed. In particular, SIFT features achieve 15.8% higher accuracy than SURF features. We believe this is due to the fact that the manner how SURF integrates the gradient information within a patch loses some discriminative power. This leads to a worse matching result where a patch may be matched to dissimilar ones. Table 1 also lists the average runtime[1] (in seconds) for recognizing a query image. By using the LSH technique, the proposed method has a runtime of about 0.1s, and, thus, is a viable solution to applications that require real-time processing.

Figure 6 shows the images that failed in the experiment using SIFT—they are rejected by the system. As the images on the top row have a very simple design—mainly texts and color blocks—very few keypoints are available on those images. Furthermore, the camera flash creates an unnatural shininess on drink can images that may deteriorate the matching results. The number of matched keypoints thus does not exceed the threshold in the coarse recognition step. For the images on the bottom row,

[1] The reported runtime includes all processing time between snapping a picture and the showing the relevant information on screen.

the system cannot differentiate the green tea and the black tea as their package designs in the "try-it" series are very similar. They differ only on portions of the background color. Since the SIFT-based representation does not take color information into account, this difference unfortunately cannot be identified in the refinement search process.

We conducted the second experiment and examined only the beverage packages that belong to a series in order to evaluate the proposed weighting scheme. We collected additional beverage package images from the web—24 images among 9 series in which there is a corresponding reference image in our dataset, and 8 images in which there isn't. The images have 33 distinct flavors. We try to mimic the situation when a new flavor variety is launched in market while the dataset is not yet updated to include the new example. The system should have the ability to reject these queries.

Except the black tea and the green tea packages in the "try-it" line, others are correctly recognized. We examined the matching results and observed that the similarities between a query and its reference image are neatly accumulated from the discriminative regions. Furthermore, the system can successfully reject the 8 images that represent new flavor varieties. Note that the 8 images could be retrieved as a false positive by conventional systems because they have similar patterns with those packages in the same line of products. The refinement search step is essential to identify the existence of critical regions that differentiate the query from others in the same series.

The proposed system has a graphical user interface as shown in Fig. 7. It streams a video and displays frames when the application is activated. It then samples frames, performs recognition and shows relevant information if the beverage package is recognized. The phone would be used as a "scanner" for checking out product information and the usage should be easy and intuitive.

Table 1. Comparison of SIFT and SURF features for beverage package recognition. Runtime is reported in seconds.

Feature	Recognition Accuracy	Runtime (Exhaustive)	Runtime (LSH)	Speedup
SIFT (128-d)	92.5%	17.8905 (5.5385)	0.1289 (0.0368)	167.73x
SURF (64-d)	76.7%	11.0793 (2.3183)	0.1086 (0.0372)	150.50x

Fig. 6. Testing images for which our method failed.

Fig. 7. Example outcome of our system. The beverage is automatically recognized and annotated with product information such as price and calorie count.

6 Conclusions

In this paper we propose an approach for practical beverage package recognition for mobile application. We examine the challenges faced in the design and the development of a practical system that can achieve a fairly good recognition performance. There are a few directions we may explore to further enhance the approach. For example, the current representation is based on SIFT descriptors that describe the gray-level images alone. However, as we observed from our reference image dataset, the color design of beverage packages seems to follow certain rules. For example, the similar, contrast, or complementary hues are commonly used in the same serious of products. A representation that encodes both the shape and color information should be more effective. Furthermore, once an image is described by more than one type of descriptors, an indexing approach that can enable fast retrieval of visual instances described by multiple cues would be desired.

References

[1] Google Goggles, http://www.google.com/mobile/googles/.
[2] SnapTell, http://www.snaptell.com.
[3] Kooaba, http://kooaba.com.
[4] http://www.foodprocessing.com/wp_downloads/gt_appetite.html
[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346-359, 2008.
[6] V. Chandrasekhar, D. Chen, A. Lin, G. Takacs, S. Tsai, N. –M. Cheung, Y. Reznik, R. Grzeszczuk, and B. Girod. Comparison of local feature descriptors for mobile visual search. In *Proceedings of IEEE International Conference on Image Processing*, Hong Kong, September 2010.
[7] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. CHoG: compressed histogram of gradients. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Miami, Florida, June 2009.
[8] V. Gaede and O. Gunther. Multidimensional access methods. *ACM Computer Survey*, 30(2):170-231, 1998.
[9] H. J. Gam. Employment of fashion orientation, shopping orientation and environmental variables to determine consumers' purchase intention of environmentally friendly clothing. In *International Textile and Apparel Association*, 2009.
[10] B. Glover and H. Bhatt. *RFID Essentials*. O'Reilly Media, 2006.
[11] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM International Conference on Multimedia*, New York, NY, October 2004.
[12] W. –T. Lee and H. –T. Chen. Probing the local-feature space of interest points. In *Proceedings of IEEE International Conference on Image Processing*, Hong Kong, September 2010.
[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91-110, 2004.
[14] R. C. Palmer. *The Bar Code Book: Reading, Printing, and Specification of Bar Code Symbols, 3rd Edition*. Helmers Publishing, 1995.
[15] D. Nister, and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, New York, NY, June 2006.
[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, June 2007.
[17] S. Tsai, D. Chen, J. Singh, and B. Girod. Rate-efficieent, real-time CD cover recognition on a camera-phone. In *Proceedings of ACM International Conference on Multimedia*, Vancouver, Canada, October 2008.
[18] S. Tsai, D. Chen, V. Chandrasekhar, G. Takacs, N. -M. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod. Mobile product recognition. In *Proceedings of ACM International Conference on Multimedia*, Florence, Italy, October 2010.