Beer Label Classification for Mobile Applications

Andrew Weitz
Department of Bioengineering
Stanford University
Email: aweitz@stanford.edu

Akshay Chaudhari
Department of Bioengineering
Stanford University
Email: akshaysc@stanford.edu

***Abstract*—We present an image processing algorithm for the automated identification of beer types using SIFT-based image matching of bottle labels. With a database of 100 beer labels from various breweries, our algorithm correctly matched 100% of corresponding query photographs with an average search time of 11 seconds. To test the sensitivity of our algorithm, we also collected and tested a second database of 30 labels from the same brewery. Remarkably, the algorithm still correctly classified 97% of labels. In addition to these results, we show that the SIFT-based recognition system is highly robust against camera motion and camera-to-bottle distance.**

I. Introduction

The emergence and pervasiveness of smartphones over the last decade has made it possible to search for and keep records of various products and activities on-the-go. One such application of smart-phones is to search for information regarding consumer products and receive instant feedback regarding the product. This process typically involves manually performing an online text search, but this can become cumbersome over time and lacks the "fun" factor of image-based searches. The objective of this study was to evaluate the feasibility and robustness of an automated image processing technique to enable rapid image-based lookups of various beer labels. Such an algorithm would use an input image of a beer bottle and compare the label to a database of beer labels in order to find a match. Indeed, one mobile application called NextGlass already implements such an algorithm to provide beer reviews and create a social network of beer consumption with friends. Therefore, while the eventual goal of this technique would be implementation on a smart-phone, the scope of this project was to develop and characterize the algorithm on a computer first.

II. Database Creation and Pre-Processing

To generate our initial database of beer labels, we collected 100 "clean" images (i.e. not photographs) of various beer labels using Google Image search. The database included a variety of breweries, with no more than 5 labels coming from the same one. Next, for each database image, a corresponding query (test) image of a beer bottle with that label was found. These test images included photographs taken 6 to 12 inches away from the bottle, so that the bottle took up at least a third of the photo. To make these query images similar to those that would be acquired with a camera phone, the image was cropped to a 4:3 aspect ratio. Finally, for computational efficiency, query and database images were downsampled to a matrix size of 400x300 pixels.

To test the sensitivity of our algorithm, we also collected a separate database of 30 labels and 30 matching query images from the same brewery (Samuel Adams). These labels were purposefully chosen to be very similar to the human eye, to allow us to evaluate how well the algorithm could classify similar-looking query images.

To test how well the algorithm could classify test images corrupted by camera motion, we simulated camera motion for each of the 100 query photographs. Motion was simulated using the fspecial command in Matlab ('motion' filter) with motion ranging from 2 to 20 pixels at angles of 0, 45, and 90 degrees.

Finally, we collected (in person) a database of query images for 5 different beer labels, with photographs taken at varying distances from the bottle (6 inches to 5 feet). Images were captured using an iPhone 5 camera. Unlike the images described above, these images were not downsampled for analysis. Rather, we used the full resolution 2448x3264 pixel photographs. This was performed to replicate the conditions for the eventual mobile realization of this algorithm.

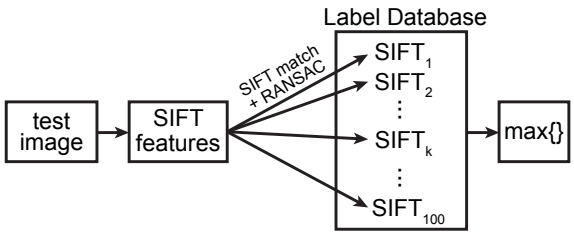

Fig. 1. Image processing strategy for SIFT-based beer label classification.

III. Label Matching Algorithm

The general processing strategy of our SIFT-based classification algorithm is provided in Fig. 1. We chose to implement SIFT, first described by Lowe [1], due to its rotational and scale invariance in image matching. At a high level, our algorithm operates by finding the database image that shares the highest number of post-RANSAC SIFT feature matches with the query image. To accomplish this, SIFT keypoints are first extracted from all the database images. These are pre-computed and stored to save time. These SIFT keypoints are computed by finding the scale-space extrema between differences-of-Gaussians (DoG) pyramids. As first described by Crowley and Stern [2], the DoG pyramids are generated by convolving the image with variable scale Gaussians.

Once SIFT keypoints are identified, a descriptor is computed for each of them. To create a descriptor that is robust to illumination changes and affine distortions, an 8-bin histogram is created for a 4x4 space around the keypoint at its specific scale. This descriptor has values created by calculating the gradient magnitude and orientation around the keypoint, and rotating it with respect to the most significant orientation determined for that keypoint. All these values are used to generate an orientation histogram of 8 bins, for each of the 16 sub-regions. This generates a descriptor vector of length 128. This vector is subsequently normalized, thresholded, and normalized again in order to mitigate the impacts of non-linear illumination changes.

This process of feature extraction is repeated for each query image by finding its SIFT keypoints and extracting the corresponding descriptor vectors. To identify the matching database image, the SIFT features of the query image are compared to those of each database image. A given pair of SIFT descriptors D1 and D2 is considered to be a match only if the Euclidean distance between them multiplied by some threshold (in this case 1.5) is not greater than the distance between D1 and all other descriptors. Once these potential matches are identified, a homography model is generated and outliers are excluded using RANSAC [3]. The database image with the highest number of feature correspondences post-RANSAC is considered to be the matching image.

In this project, the SIFT keypoints and features were computed and matched using the vl_feat toolbox [4].

IV. Results

A. Algorithm Performance and Sensitivity

Each query classification took around 11 seconds to evaluate. Overall, the algorithm performance was 100%, with the algorithm correctly matching each of the 100 query camera images to the correct label. To our surprise, the algorithm was also robust against similar labels from the same brewery (in this case, Samuel Adams). Out of 30 Samuel Adams labels, 29 of the corresponding query images were correctly classified. All the Samuel Adams labels can be seen in Fig. 2. Four representative examples of correctly classified query images are provided in Fig. 3.

B. Effect of camera motion

Even after query images were filtered to have 16 pixels of simulated camera motion, the label matching algorithm performed with a success rate of at least 50% (Fig. 4). Note that with 100 database labels, random chance is a 1% success rate. The success rate was almost perfect for around 6 pixels of simulated camera motion, but beyond that, it dropped linearly at a rate proportional to the number of pixels of motion. There did not appear to be a strong dependence on the motion angle on the overall success rate as all three angles ($0°, 45°$, and $90°$) exhibited similar success rates.

One example of a query image subjected to motion is shown in Fig. 5. While the number of RANSAC matches decreased from 168 to 13 because of the 20 pixel motion, a correct classification was still made by the algorithm.

Fig. 2. 30 labels from the Samuel Adams brewery were used to generate a sensitivity metric due to their visual similarity.

Fig. 3. Four representative query images and corresponding labels (as identified with our algorithm) show that the algorithm is robust against label from the same brewery. Post-RANSAC correspondences are shown for only one of the label pairs for visual clarity.

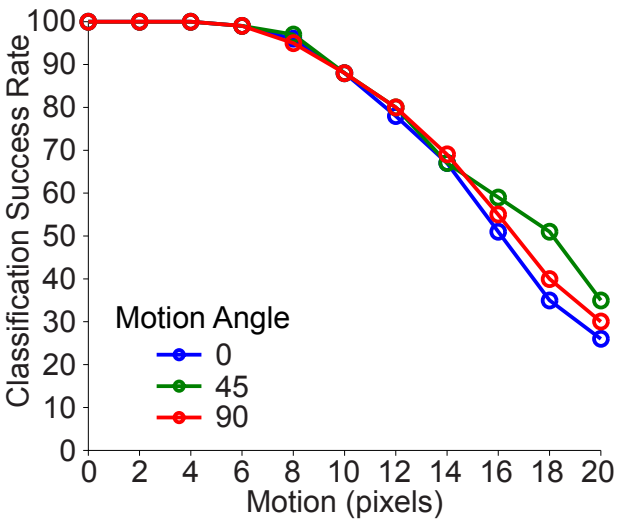

Fig. 4. Classification success rate as a function of simulated motion.

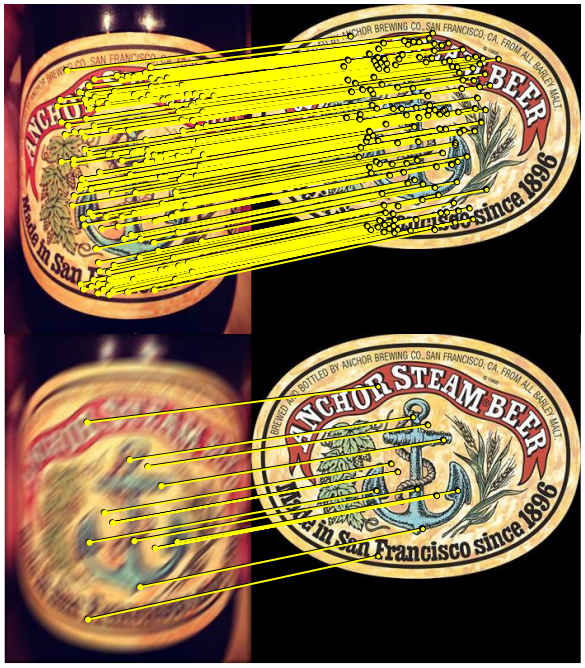

Fig. 5. (Top) Original image pair of RANSAC matches between an Anchor Steam query and label image shows 168 RANSAC matches. (Bottom) Correctly classified query image with 20 pixels of 45 degree motion shows 13 RANSAC matches.

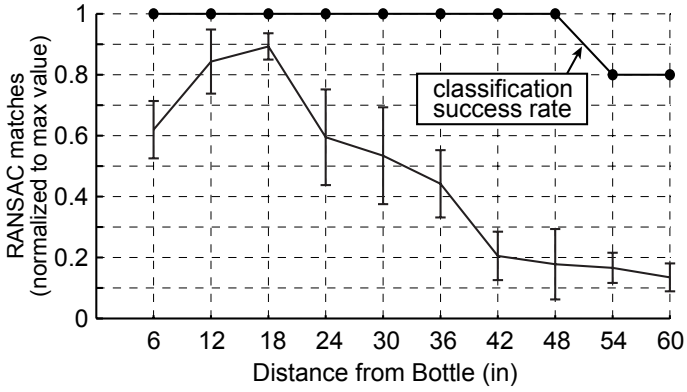

Fig. 6. Classification success rate and the number of RANSAC matches as a function of the distance between the query image bottle and the camera.

C. Dependence on camera-to-bottle distance

As can be seen in Fig. 6, there is no dependence on the distances to bottle and the overall success rate, as long as the bottle is within 4 feet of the camera (n = 5 labels). At distances less than 4 feet, the success rate remains at 100%, while beyond 4 feet, the success rate drops to 80% (i.e. 1 of 5 labels incorrectly classified). The number of RANSAC matches reach a maximum at 18 inches. Fig. 7 provides a representative query image that was correctly matched to its database label at distance of 36 inches, showing that our algorithm is robust to camera-to-bottle distance (as long as the resolution is high enough to discern different keypoints).

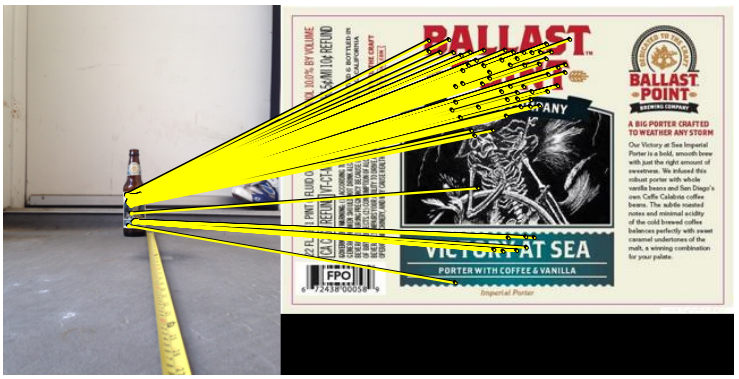

Fig. 7. The query image on the left was correctly matched to its database label (shown on right) even when the bottle was 36 inches away from the camera. This is representative of the algorithm's general robustness against camera-to-bottle distance.

V. Discussion

The perfect success rate of the algorithm is a testament to the robustness of the SIFT keypoint detector and description technique. This is especially true considering that 29 out of 30 Samuel Adams labels which were very similar in appearance could be correctly classified, with the SIFT algorithm teasing out minor differences. It is interesting to note that in this sensitivity analysis, the correct classification was made possible by the additional correspondences detected in the subtle background behind the Samuel Adams text. In addition, several matches were also made in the actual text of the name of the beer. Thus, despite a very similar macroscopic appearance, the subtle background and the name of the beer were used to perform accurate classifications.

The robustness of the algorithm to motion was not entirely unexpected either. Since the SIFT keypoint detector relies on blurring with Gaussians of variable scales, the net effect is similar to that of a motion blur. With the robustness of SIFT, even though the number of RANSAC matches decreased with motion, accurate matches were still possible. Based on the severely degraded image quality of the motion image in Fig. 5 (which still produced an accurate classification), it might be safe to claim that this algorithm is immune to typical blurs seen in pictures created with mobile phones. Furthermore, the lack of sensitivity to the specific angle of motion may be due to beer labels generally not having a predominant angle in their gradients.

The effect of distance between the camera and the beer bottle was shown to be relatively mild since perfect classifications could still be performed when beer bottles were 4 feet away from the camera. Fig. 6 seems to suggest that it might be best to have the bottle 18 inches away in order to maximize the number of RANSAC matches. When the bottle was 6 inches and 12 inches away from the camera, it was challenging to get the entire label in the picture which results in lost information that could have been used for keypoint matching. However, the interesting aspect to note from Fig. 6 is that while the number of RANSAC matches kept decreasing with the distance, the classification success rate stayed relatively constant. This may suggest that the absolute quantity of matches may not be as important as the uniqueness of the detected keypoint. It is also worth reiterating that the camera images were not

downsampled for the distance experiment. If the images were to be downsampled, there would be very little fine detail available in images that are far away from the camera. This would suggest that there is a need to evaluate a dynamic depth-based downsampling algorithm.

Each query classification took around 10 seconds to evaluate on 8 parallel processors in MATLAB. Implementing a parallel algorithm for mobile phones is quite reasonable since most new smart-phones are indeed octa-core processors. The SIFT keypoints and descriptors for the labels were precomputed and cached in order to increase computational efficiency. While 10 seconds is reasonably efficient, a faster algorithm that could perform the detection in 1-2 seconds would certainly be preferable. This would especially be true if the database of labels would be more than the 100 labels used in this study. Implementing this algorithm in C or Java could lead to increased efficiency. Thresholding the SIFT keypoint detection (which was not done in these experiments) would also dramatically reduce the computation overhead. Together, these points suggest that the algorithm developed here could be readily deployed on a mobile-based platform.

VI. Conclusions

In the project, we developed and characterized a digital image processing algorithm for the automated detection of beer labels from photographs of 100 different beer bottles. The algorithm achieved a high (100%) success rate, was sensitive to subtle differences between distinct labels, and displayed robust classification against simulated camera motion and large camera-to-bottle distances. This tool would be appropriate for various mobile phone applications, including resources for consumer product information and even social networks.

Acknowledgement

We would like to thank Professors Bernd Girod and Gordon Wetzstein, TAs Huizhong Chen and Jean-Baptiste Boin, and project mentor Jason Chaves for valuable advice on this project, as well as broader insight into digitial image processing stategies throughout the quarter.

Work Breakdown

Andrew Weitz: Collected original database of 100 test and database images, took photographs of bottles at various distances, and developed code to perform general SIFT matching between query photographs and database images. Contributed to poster and paper.

Akshay Chaudhari: Collected database for sensitivity analysis, and developed code to test robustness against motion and camera-to-bottle distance. Contributed to poster and paper.

References

[1] D. Lowe. (1999). Object recognition from local scale-invariant features. *Proc. 7th International Conference on Computer Vision* (Corfu, Greece): 1150-1157.

[2] J. Crowley and R. Stern. (1984). Fast computation of the difference of low pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(2):212-222.

[3] M.A. Fishler and R.C. Bolles. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*. 24(6):381-395.

[4] A. Vedaldi and B. Fulkerson. (2010). VLFeat: An open and portable library of computer vision algorithms." *Proceedings of the international conference on Multimedia. ACM.*