TSB-funded Project 'TADD' - Trainable vision-based anomaly detection and diagnosis Technical Report for November 2013

Ran Song and Tom Duckett

Agri-Food Technology Research Group, Lincoln Centre for Autonomous Systems Research, School of Computer Science, University of Lincoln, UK

Abstract

In November, we investigate the techniques which can be used for label recognition on food tray. The techniques have to be efficient and reliable because we pursue a real-time system capable of handling input food tray images of high variety. We finally ended up with feature-based image registration techniques using SURF and RANSAC. Using such techniques, the system is able to answer three questions related to label recognition: (1) is some label on the tray? (2) if it is on the tray, is it at the right position? (3) if it is not at the right position, how wrong is its position?

1. Introduction

In this report, we give details about the techniques we used for label recognition on food tray images. The basic idea is to utilise SURF (Speeded Up Robust Features) [1] and RANSAC (Random Sample Consensus) [2] to register a test image into the coordinate system of a reference image where everything is assumed to be right. In the current TADD system, we have also moved to SURF from SIFT [3] for feature detection due to efficiency reason.

2. Methods and results

The SURF method is technically complicated for non-expert. Thus here we just give a simple explanation on how it works. SURF contains two

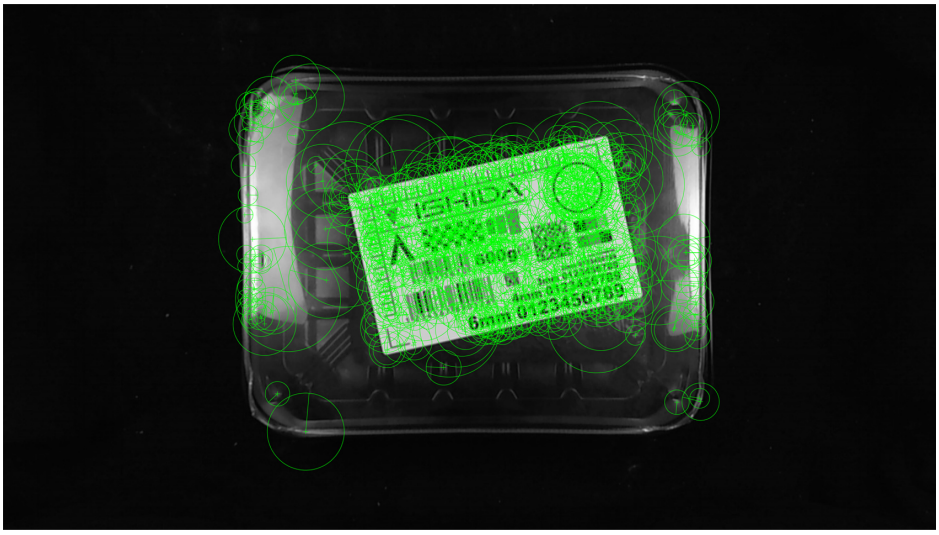

Figure 1: 913 SURF features are detected on a food tray image. The '+' denotes the location of each feature, the size of the circle around each '+' denotes the scale of each feature and the line segment within the circle denotes the orientation of the feature.

Figure 2: The strongest 100 SURF features out of the 913 features shown in Figure. 1

components: a detector and a descriptor. A detector is to find the locations of the features, which thus allows us to visualise them as points in an image. A descriptor is to dig out the (mostly local) properties attached to the points

Figure 3: The raw matches produced by SURF contains some incorrect matches. To visualise the results, we deliberately colour the two images in red and blue respectively.

which can be used to distinguish them and match them over different images in the presence of image rotation, translation, scale change and illumination change, etc.

The detector is based on the Hessian matrix, but uses a very basic approximation. Such an approximation relies on integral images to reduce the computation time. The descriptor, on the other hand, describes a distribution of Haar-wavelet responses within the neighbourhood of the feature point. Again, integral images are exploited for speed. Figure. 1 illustrates our results of applying SURF detector over food tray images and Figure. 2 shows the strongest 100 SURF features.

Although the SURF method offers a descriptor to facilitate the matching of the features and indeed it achieves a high matching rate of 82.6% compared with 78.1% of SIFT, it cannot eliminate mismatches which could lead to wrong registration in the following steps. Therefore, we introduce RANSAC to refine the matches produced by the original SURF and further eliminate all mismatched feature points.

The basic theory of the RANSAC method is simple but ingenious. Firstly, it randomly samples two feature points (from the set of matched feature points produced by SURF) to fix one line (which represents the least-square solution of the motion estimation between two images in our application).

Figure 4: By using RANSAC, the matches are refined. In this case, only the feature points within the label region are preserved.

The underlay (or consensus) of this line is defined as the points whose distance to this line are less than some threshold. After many iterations of this random sampling approach, the final solution is the line which has the largest consensus and the points within the consensus are the inliers that consist of the consensus. The rest of the points are viewed as outliers. To ensure that the random sampling has a good chance of finding a true set of inliers, a sufficient number of samplings must be tried. In this application, we set it to 2000. Figure. 3 shows the result of raw feature matching produced by SURF where we use the yellow lines to illustrate the matches. It can be seen that some features points are not correctly matched. A correct match means that the two matched points should correspond to the same point in the real world although they are from different images. In contrast, Figure. 4 demonstrates that by using the RANSAC method, we can eliminate the incorrect matches where each yellow line correctly links two features which correspond to the same point the real world.

Once we have produced a set of correctly matched feature points, we can use them to do registration. As shown in Figure. 4, all of the inliers are within the label region. Hence, we actually register one label to the other reference label which is supposed to be placed correctly.

Rigid image registration is a fundamental topic in computer vision. In this

Figure 5: The two input images where the bottom one is the reference image.

work, we assume that the motion between two images are affine transform. Such a motion can be formulated as a set of linear equations given a known set of matched points. The unknown parameters, the 6 motion parameters (which describe the rotation and translation between the two labels in this case) can be solved in a least square manner using for example, Levenberg-Marquardt method. Thus by computing the motion parameters, we can know the relative position and orientation of the label to the reference one. In other words, if we assume that the label is at the right position with the right orientation in a reference food tray image, we can know how wrong the position and the orientation of the label in a test image are by doing

Figure 6: Rectified test image where the label is actually at the same position as that in the reference image

registration. Figure. 5 shows one test image and one reference image and Figure. 6 shows the new position and orientation of the test image after the registration. Figure. 7 displays the overlapped feature points in the two images. It can be observed that each pair of corresponding feature points from different images have already been moved to the same position after the registration.

3. Conclusions and future work

The techniques reported above demonstrated the capability of combined 'SURF+RANSAC' solution over the issue of label recognition. According to our preliminary implementation without any code optimisation, the whole process needs about 2.5 seconds. In short, 'SURF+RANSAC' can answer the three questions that we raised at the beginning of this report as long as we know the reference image or simply 'what is the right one like'. And this will guide us in our future work: how can we make the system learn 'what is the right one like'?

The original idea is to first input a tray with blank film and then input another tray with only one label at the right position. Then we need to implement SURF method on both images to output two sets of features.

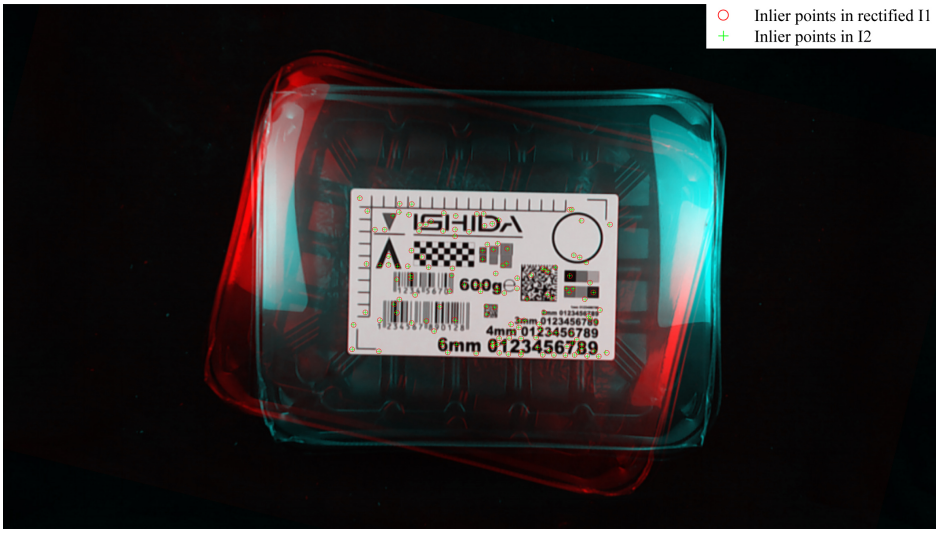

Figure 7: After registration, the feature points from different images are fully overlapped.

The difference set between the two sets is supposed to be the set of SURF features within the label region. However, our tests show that due to the strong reflection, this is not always the case. Reflection can produce new features on a food tray image. In fact, in Figure. 1, most of the detected SURF features outside the label region are not driven by the pattern and the texture of the tray itself but caused by the reflection. The key thing is that there is no guarantee that we can always lay trays at the same position so that the reflections on the surface of the trays always change. And that will lead to different results of SURF feature detection in the same region of a tray. However, because we are currently not quite sure about the eventual lighting condition, we will revisit this idea by using input images captured under different lighting conditions (e.g., using the vision system attached to the Ishida machine).

A potentially good solution to this problem is to do the graph-based segmentation within the tray region. As we mentioned in our previous reports, we used graph-based method for foreground-background segmentation. We may extend this method in order to further partition the foreground region. First we input a reference tray image where all of labels are positioned correctly and apply SURF on it. If the graph-based method can reliably partition its foreground tray region into different components, we can record the SURF features locating in each component. And as long as we can do a

semantic labelling of the components, the system will know what some component is (e.g., a price label or a barcode label). In this way, the system can learn which SURF features correspond to a specific label. Given that a set of SURF features corresponding to the price label are detected in the reference image, we can check whether there are a enough number of matches detected in a test tray image. If the answer is 'No', it means that the price label is missing. If the answer is 'Yes', we can further figure out whether the label is at the right position and how wrong its position is using the aforementioned registration-based methods. And this scheme could be more desirable since we do not need to input trays again and again if one tray contains several labels.

References

[1] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Proc. ECCV, Springer, 2006, pp. 404–417.

[2] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395.

[3] D. Lowe, Distinctive image features from scale-invar-iant keypoints, International Journal of Computer Vision 60 (2004) 91–110.