Implementation of Open-Source Image Analysis Techniques in Commercial Quality Control Systems

Ross Ockmore\textsuperscript{1,2}, Wei Xu\textsuperscript{1,2,\textcopyright}, and Hong Wei\textsuperscript{1,2}

\textsuperscript{1} University of Surrey, Guildford GU2 7XH, UK
\{w.xu, h.wei\}@surrey.ac.uk
\textsuperscript{2} University of Reading, Reading Berkshire RG6 6AH, UK

Abstract. Using image pattern matching techniques, this paper presents the development of an open-source based quality control program that can distinguish between sample pages in a book printing/production environment, intended to enable automatic rejection of incorrect pages from the binding process. Novel adaptations are made to an ORB (Oriented FAST and Rotated Brief) based image matching system that deals with quantifying the confidence of a match to ensure that even identical pages in the incorrect orientation are rejected during book binding operations. The program is subjected to a variety of quantitative tests to evaluate its performance and from these tests the effects of various parameters used in the program are discovered, allowing tuning to be performed. Potential paths for development of the analysis program are discussed, with the implementation of machine learning, highlighted as a possibility, to provide automated parameter tuning, and suggestions for further optimisation of the software are made, with a view to creating an even more bespoke version that retains performance whilst cutting computational overhead to a minimum.

1 Introduction

Quality control and assurance are an essential part of any manufacturing operation. Preventing the release of defective goods for public consumption is imperative for retention of both company reputation and a healthy customer base. When dealing with end-consumers, a lack of quality assurance leads to dissatisfaction, often resulting in decreased return custom; a publisher receiving a large number of complaints from readers may decide to terminate a contract with a manufacturer/supplier and may pursue reimbursement for lost profits. Losing clients is a major blow to many manufacturers, and it can trigger reviews by remaining clients to ensure that problems with quality are not a companywide issue. To inspire confidence with clients, and to win new business, companies typically aim to achieve ISO 9000 accreditations, which require the implementation of procedures and documentation to carefully control the consistency of output (UKEssays 2018). In the modern era, production techniques have become highly automated in nature, as output speeds and volumes increase to meet demand in the global market. It only makes sense therefore, that the quality assurance process is
also automated to a similar or the same degree, allowing for tight controls and more accurate checking of goods.

Book binding is a process by which leaves of paper are joined together to form a cohesive codex. Dating back thousands of years, this process is still utilised across the world – despite the rise of digital media, traditional paper-based novels and textbooks still account for a large proportion of total book sales (Parker 2018). There is a preference for the physical form and tactile nature of paper-based books in society, and as such it can be assumed that production is likely to continue for many years. Due to this continued demand, it is sensible to develop control measures for book production lines, to bring the industry in line with the latest technological developments, increasing efficiency.

Any device that is intended to be utilised in a production setting for page sorting must be capable of monitoring the orientation and ‘correctness’, making evaluations and determining whether a sheet is suitable for advancement to the stacking/sheaf stage of production. Evaluation of advancement criteria should mimic the human visual inspection process of goods and decision-making abilities, and therefore relies on the capture and processing of images. The prevalence of image processing has risen rapidly in recent years with the advancement of smartphone technology. As digital control of production lines is pervasive within the industry it is reasonable to expect that algorithms/software that can, in combination with a camera sensor, rapidly analyse sample photos of pages and determine whether they should be rejected or accepted are already used by companies. However, it is also likely that companies are paying licensing fees for these capabilities and as such are losing out on potential profit. The creation of a program with similar capabilities was explored, which is instead based upon open-source development, thus removing license fees for the industry.

2 Methodology

Image feature matching is a well-developed field with open-source libraries. Algorithms such as ORB (Oriented FAST and Rotated BRIEF), which was specifically created to offer similar functionality to replace the widely used SIFT algorithm without licensing issues for the end-user (Rublee et al. 2011) are integrated into the library to be used as tools when solving image matching problems.

These tools are designed to extract features that can be used in a matching process from images, firstly by identifying ‘keypoints’ within the image data, and then by producing an image descriptor for the area around the keypoint (Mordvintsev and Abid 2013). This method is analogous to the human process of studying an image, in which different aspects of the image will draw the eye – a building in a natural landscape will likely have a notable difference in composition to its surroundings, with sharp, defined edges and corners.

These tools do require augmentation with further techniques depending on the application and using the Python platform, several additional processing steps were implemented to achieve a good level of functionality for this particular quality control task, as can be seen in the flowchart above in Fig. 1. Firstly, Lowe’s ratio test is applied to reduce the number of false/poor matches passed to the latter stages. As explained in Lowe’s paper, this is implemented by matching each query image feature to its two
closest counterparts from the training dataset and then comparing the likeness of each to determine if a good match has been made (Lowe 2004). Once these matches have been determined, if a sufficient number of matches have been identified the next step of the process is to calculate a confidence score – whilst the ratio test filters matches, it provides no metric of how alike two pages are, and could produce matches between identical letters/words at different locations on a page. ‘Homography’ is the term used to describe the relationship between a deformed image and its master, such that any pixel in the master image transformed according to the homographic relationship will be located in the correct position in the deformed image space (CorrMap 2013). This information is typically employed in photographic manipulation when ‘stitching’ separate images of the same scene, taken from different perspectives, or when attempting to identify an object that is viewed at a different angle from the sample image that produced training features for that object. In the context of this project, homography offers a method by which to create a confidence score for the likeness of two processed pages, which can be used to reject a page from the production line.

The principle behind the calculation of a confidence score stems from the assumption that in a good match scenario both the train and query subjects are identical. As such, it is expected that after a ratio test has filtered subpar pairings, all points remaining should share a common translational and rotational homography between the two images. This is unlikely to be the case as the majority should fit the common homography if the match truly is correct. As such, the proportion of matches that are inliers can be considered against the total number of matches submitted to the algorithm that calculates homography, and this gives a confidence metric. Similarly, to how the ratio can be set for Lowe’s test, this metric has a threshold that is set by the end-user, and so some degree of fine-tuning can be performed depending on testing in different production environments.

Homography can be found using a statistical technique known as Random Sample Consensus (RANSAC), which sorts through data sets generating estimates of fitting parameters from small samples and then testing the application of the fitting parameters to the dataset (Fischler and Bolles 1981). In this case, RANSAC examines pairings and attempts to find the most applicable homography for the majority of pairs. RANSAC is robust against the effects of strong outliers, that is to say, outliers that are very far from the expected result and are prone to skew averaging/horning methods.
As ORB is designed to be accepting of rotational variation, it is ultimately lacking the ability to reject a page in an incorrect orientation for the binding process. This issue is demonstrated above in Fig. 2 where an inverted query subject has been compared to the train image, producing a large number of matches and a high confidence score. As it is can be assumed that both train and query images shall be captured in-plane in the quality control process, then a purely in-plane rotation value can be used to remove false positives from the matching algorithm, and thus reject pages off of the production line. Whilst the homography found previously contains all of the transformation data linking the two images, it can be difficult to extract a purely in-plane rotational value.

Fortunately, the oriented FAST keypoint data used by ORB is retained throughout the calculations made in the analysis process, and within this data, the keypoint orientation is stored. As long as the page can only translate and rotate on one plane, and the assumption can be made that after the ratio and homography filtration that each keypoint pair is correct, this orientation data can be directly compared within pairs and a rotational value computed. An average rotation value can thus be calculated from the set of correctly matched pairs, and whilst this has an amount of variation, it is a reliable indicator for a large rotation. The threshold for rotation can be set at a high value such as 90° so that only an inverted page is rejected.

3 Evaluation

3.1 Testing Setup and Aims

For all testing, a laptop was used with an i5-1035G1 Intel processor clocked at 1.19 GHz, with the ability to boost to 3.6 GHz under load. It is unlikely that the analysis program will use more than one core and so should yield good results in terms of allocating processing power to the program. For image capture during testing, a Microsoft LifeCam HD-3000 USB camera was used. The laptop was using a 64-Bit Windows 10 operating system in all experiments. Python 3.7.4 was used to run the analysis program, and version 4.1.2.30 of OpenCV was used, in conjunction with NumPy version 1.18.1. The latest available driver was installed for the USB camera used and all images captured from the USB camera were 640 × 480 pixels in resolution.
As shown in Fig. 3 the test samples consisted of two with text-heavy and two with a mixture of text and images, henceforth be referred to as Sample 1 (top left in figure), Sample 2 (top right), Sample 3 (bottom left), and Sample 4 (bottom right).

![Sample test media](image)

Fig. 3. Sample test media

Testing was primarily based around the user tuneable parameters involved in the analysis process and external factors affecting the program. This information could then be used to calibrate the program, allowing its performance to be assessed in a quality control capacity. Measurements were taken at least five times for each trial, and averages were established to help control random error. More detailed descriptions of testing procedures are listed in the following sections, with results.

### 3.2 Influence of Magnitude of ORB Feature Generation

As ORB features are the data that is used in comparisons of images, it is reasonable to expect that a change in the number of features generated may have an effect on the number of matches and thus the quality of a comparison. Conversely, generating more features could have an impact on processing, and as such, it is likely that an optimal compromise exists. To test this, the ratio test value was set and held at a nominal value of 0.7, selected arbitrarily from the near optimal region, found in Lowe’s paper.

A comparison was performed by capturing train and query images of Sample 1. The confidence threshold was removed, and the program was set to report the confidence of a true image match, as well as the time taken to complete the analysis, and the number of keypoint matches generated. This data was then recorded.

Figure 4 Below demonstrates the change in both yield and processing time as the number of ORB features generated changes. Yield, in this case, refers to the number of matches accepted as good as compared to the number of features generated. Initially, the yield tends to remain constant as the number of features produced increases, but after a
certain point decreases drastically. The yield value would be expected to remain reasonably constant, as the comparison images are identical, however, the decline behaviour may be explained by considering the nature of a feature.

![Graph showing variation of match yield and processing time due to ORB feature count change](image)

**Fig. 4.** Variation of match yield and processing time due to ORB feature count change

Features are described around key points of interest, and in any image there are only a finite amount of points that a human would consider distinctive. Once all of these points have been found and described, extraneous keypoints will be based on decreasingly distinct features, until minor variations in the mathematical calculations performed by the computer on identical images can result in different key points being identified and described. As such, these extraneous features are unlikely to have matching correspondent points in the opposite dataset, so cannot be matched. Therefore, the number of significantly distinct points (and matches) remains the same, as the number of features increases, thus reducing yield.

The execution time increase exhibited by the program as the number of features increases is logical; the generation of each feature requires computation, as does each match. Therefore, the production of more features will increase the amount of operations the laptop has to execute before the dataset is complete and will increase the amount of points to be compared during matching. The time to compute does, however, remain fairly consistent, with only small increases up to 1000 features.

As a compromise, it is sensible to move forward using 1000 features despite the small increase in delay, as the generation of more features within the train target zone should allow for more tolerance to offset/movement of the query image out of the same relative frame, as there are more points to match in the first place. Using more than 1000 features causes large relative increases in delay and using too few features may affect comparison abilities for offset, identical samples.
3.3 Influence of Ratio Test Value

The ratio test can be either overly discriminate or accepting of keypoint matches depending on the value chosen. Whilst Lowe’s work identified a general range of values that can be applied, it was decided that the effects of varying the ratio value on this specific system and type of media should be observed, so that the software could be tuned to a near optimal ratio value. As such, the optimal number ORB features generated from the previous test was set and held in the program parameters, and again the confidence threshold for matches was removed, with the same values reported and recorded as in the prior test.

![Variation of returned matches and comparison confidence metric due to change in ratio test value](image)

**Fig. 5.** Variation of returned matches and comparison confidence metric due to change in ratio test value

Figure 5 above demonstrates the changes in returned matches and the confidence value achieved in comparison, as ratio test value varies. An increase in returned matches as the ratio test criteria is loosened to be expected, as it is intended to be a filter for bad matches. If the definition of what is a ‘bad match’ is altered to include more matches, then more matches will be classified as ‘good’. In the case of the analysis program, the confidence metric, based on how many matches agree with the homography generated from the datasets, can be viewed as a confirmation of whether a match is ‘good’ or not.

When identical images are compared, if the match passes the ratio test but does not fit the general transform that relates the other matches for the train/query feature set then it is clearly erroneous; the images are directly related by a transform and all key points should have a partner with the same description and transformed key point location. The opposite situation can occur when the ratio test is tightened, as it could wrongly exclude a match that would have otherwise fit the homography perfectly and is therefore wasting a potential ‘good’ match.

Similarly to the ORB feature generation increase, which increases the number of features that can be used to make a positive identification between an offset image and a train from an identical sample, increasing the number of matches allowed through the ratio test should aid in providing a definitive acceptance. Every good match excluded is
one less match from which to generate homography, and if the total number of matches is reduced significantly due to large offset then this one match may have a significant effect on the result of the confidence assessment.

As before it is best to make a compromise between decreasing the confidence by letting bad matches through and restricting the number of good matches passed on by the ratio test. Too many bad matches in the dataset could skew the results of the RANSAC procedure despite its tolerance for outliers, and as stated above, too few good matches leave RANSAC with no data to work from.

It can clearly be seen from the graph that confidence does not notably decrease until the ratio test value reaches 0.7, and all values lower than this result in a reduction of matches with no gain of confidence. After 0.7, increased values return reduced confidence, therefore the matches gained are bad, and offer no positive benefit to the program. Therefore, it is logical to continue using 0.7 for this parameter, as used in previous tests, since a large proportion of matches are already gained by this point compared to the lowest ratio test values.

### 3.4 Influence of Offset

Only a small area of approximately 220 mm × 160 mm on the test media is captured by the overhead camera. The key points are generated and described in this area. When the paper moves through the production line, it is conceivable that the query image could change enough to produce a false positive when compared to the train sample area, and as such could cause a page to be mistakenly rejected.

To quantify the degree to which the analysis program is affected by offset when processing the media it is expected to be used upon, two tests were carried out using Sample 1 and Sample 2 to represent a mixture of page content. As before, each test was carried out comparing train and query image captures from the same page, whilst the ratio value and ORB feature count were held constant at the optimal values found in the previous two tests. Similarly, the confidence threshold was removed, and the same values of confidence and matched pairs were reported as in previous tests as the position of the samples was adjusted to introduce x-axis offset, as well as processing time. Once complete, a confidence threshold can be chosen from a desired offset tolerance (Fig. 6).

Figure 7 shows that offsetting the samples for comparison in the x-axis has a major effect on analysis results. Testing did not consider y-axis variation as it is assumed to be controlled by a production system that feeds paper in a linear fashion.

Both confidence and number of matches returned appear to fall off in a reasonably predictable manner as an offset is increased, although the data is more sparse than desirable for the calculation of an approximate function to describe the behaviour for either attribute. Despite the large reduction in both values, the result from such a test would appear to be fairly reliable, with a fair number of matches generating a reasonable confidence score. It is unlikely in this scenario that an incorrect page would be falsely recognised as acceptable.

This combined score is an attempt to generalise the behaviour of the program, although with a small sample size it is not a particularly robust model. The text-heavy Sample 2 suffers a much more drastic reduction in performance, as shown in Fig. 8.
Figure 6 also demonstrates some minor inconsistency in results that arose during testing. The differing of confidence measurements from the general trend at 20 mm offset has a large influence on the combined result graph. This difference could be attributed to a subtle change in lighting that may not have been noticeable to the naked eye but was significant enough to influence matching performance.
Bearing the results of this test in mind, it is sensible to account for a specific tolerance of offset by design; 5% x-axis offset tolerance on an A3 page allows for around 20 mm of movement, which is 10% of the image capture area width, whilst maintaining 60% confidence. It can be seen in Fig. 8 that this remains true for text-heavy samples, and so one solution to this issue is to design the processing machinery to place a page within this tolerance, or to simply move the camera’s position so that the proportion of the page that is captured is increased. Thus, the 10% tolerance on image capture area offset allows a larger tolerance on overall page position.

3.5 Influence of Illumination

For the testing described above, all work was carried out in consistent lighting condition, however, the changeable conditions could affect the ability of the program to differentiate between samples. As such, a test was devised using a diffused artificial light source, the intensity of which could be varied. An LDR/potential divider sensor circuit was used to measure a relative value of light intensity in the target area. Starting at the brightest value, which was approximately equivalent to good ambient daylight, the light was stepped down between trials, with ORB features and ratio test value held constant as before. Again, measures of confidence, matched pairs and processing time were observed.

![Fig. 8. Variation of returned matches and comparison confidence metric due to change in light intensity](image)

Light intensity was shown to have a large effect on both the confidence metric and on the number of returned matches, as shown in Fig. 9. In spite of this large change, confidence does remain sufficiently high for a positive identification to be made, but this may have been because of the lack of movement of the identical test sample. Whilst the number of matches remains high enough to be considered a good sample size for RANSAC to work upon and produce homography from, by the time a 40% decrease in relative light intensity is reached, the number of samples has dropped significantly.
The drop is such that it is debatable, despite high confidence values, whether it can be concluded that the result of a positive match is statistically sound. It is important to note that if no minimum threshold is set for returned matches, it should theoretically be possible to generate a homography from only two matches, and this is problematic as any two matches can be related by a transform with 100% confidence attached to it. Adding a third match that doesn’t fit the transform in this scenario would only decrease the confidence by 33%.

Fig. 9. Daylight equivalent, reference (Left) compared to 50% relative reduction in light intensity (Right)

However, these lighting changes are extreme. As an example of the difference between relative light intensities as shown in Fig. 10, it offers a visual comparison of a capture at 100% relative intensity next to a capture at 50% intensity, the point at which the analysis program returns no reliable matches at all. It is clear that this is caused by an almost complete loss of contrast which is a vital detection feature.

The solution to this issue is to ensure that the production environment has consistent lighting. Some variation will be tolerated, but the loss in performance could compound matching issues for offset samples, and in the case that lighting conditions cannot be guaranteed to remain constant, a camera with a larger sensor and faster aperture may be required to compensate.

### 3.6 Binary Testing

To determine the performance of the system, binary tests were performed, comparing samples pages to one another. In 160 tests, only one incorrect matching judgement was made. A larger sample would be needed to confirm, though this initial testing indicates a basic degree of accuracy and repeatability has been achieved.

### 3.7 Processing Speed

During testing of the analysis program, from all available test data recorded where timing data was noted, an average processing time of just over one second per image comparison task was calculated. This is abnormally high for a system using ORB, as research
Implementation of Open-Source Image Analysis Techniques

suggested that it was an algorithm capable of real time analysis of video feeds. During scripting of the analysis program, testing was carried out and processing times were considerably lower than one second. It was initially believed that the added complexity of homography and inversion checks was responsible for the slow down.

However, upon further investigation of this issue, it appears that the camera capture is responsible for added processing time. It transpires when the camera captures an image, it needs to be initialised each time before a photo is taken.

It is confirmed by tests using a short Python script, the initiation process adds the time as shown in Table 1. The average time to grab a frame from the camera is around one millisecond, as opposed to over one second for the full initialisation and shutdown process.

| Operation                  | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Avg. |
|----------------------------|--------|--------|--------|--------|--------|------|
| Frame Grab (s)             | 0.0010 | 0.0009 | 0.0010 | 0.0033 | 0.0009 | 0.0014|
| Start, Grab, Release (s)   | 1.2990 | 1.2080 | 1.2010 | 1.1476 | 1.2650 | 1.2241|

4 Conclusion

The primary task of the work presented in this paper was to develop open-source based software for replacement of visual inspection in a book production quality control system. It can be concluded from the testing performed that the analysis program is effective, with a low binary test failure rate. Whilst the program is currently slower than the desired speed for real-time on-line inspection, the cause of this issue has been identified as a hardware interaction issue and can be rectified to bring processing time down, facilitating multiple image comparisons per second.

Optimisation of the program can be performed. Employing ORB as it currently does, the program removes the intolerance of angular variation of BRIEF, and then a filter must be applied later on to stop inverted image acceptance. Both of these steps add a small amount of computational time, but the suitability of ORB to this scenario has been proven. The study shows that only the most relevant parts of the algorithm are used in a new process bespoke to the quality control task. It may prove fruitful to move the program to a new programming language that is non-interpreted. This should, again, help to improve processing speed, but may also open up the possibility of combining both the low-level controller and the analysis program on a microprocessor-based system. Furthermore, the tuning performed in this work could be adapted into an automated process, adopting a rudimentary form of machine learning where the end-user trains the program on a large dataset of media similar to that which is intended to be analysed. This may yield more refined tuning parameters than a human input to the system does, and as these parameters have demonstrated to have such a large effect on system performance, a set of better tuned parameters could offer further enhancement to the system.
References

CorrMapp: The Homography transformation, March 2013. https://www.corrmap.com/features/homography_transformation.php. Accessed 14 May 2020

Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)

Lowe, D.: Distinctive image features from scale-invariant keypoints (2004). https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf. Accessed 11 May 2020

Mordvintsev, A., Abid, K.: Understanding Features, August 2013. https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_features_meaning/py_features_meaning.html. Accessed 10 May 2020

Parker, O.: Print books vs. e-books – an update, March 2018. https://graphicartsmag.com/articles/2018/03/print-books-vs-e-books-update/. Accessed 13 Nov 2019

Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2564–2571, November 2011. https://doi.org/10.1109/ICCV.2011.6126544

UKEssays: Importance of ISO 9000, November 2018. https://www.ukessays.com/essays/engineering/importance-of-iso.php?vref=1. Accessed 13 Nov 2019