DOI: 10.58895/ksp/1000032143-15 erschienen in:
OCM 2013 - Optical Characterization of Materials - conference proceedings
DOI: 10.58895/ksp/1000032143 | https://www.ksp.kit.edu/site/books/m/10.58895/ksp/1000032143/

Image analysis of natural products

Daniel Garten[1], Katharina Anding[2], Steffen Lerm[2], Gerhard Linß[2]
and Peter Brückner[2]

[1] GFE – Society for Production Engeneering and Development Schmalkalden
Naeherstiller Strasse 10, D-98574 Schmalkalden
[2] Ilmenau University of Technology, Department for Quality Assurance and
Industrial Image Processing,
Gustav-Kirchhoff-Platz 2, D-98693 Ilmenau

Abstract Natural products are exposed to various environmental influences resulting in a high phenotypical variability. This makes it very difficult to develop automatic recognition algorithms in contrast to the recognition of manufactured products. Recent developments in the field of computer science, especially the development of powerful classification algorithms like the support vector machine make it possible to face also such complicated tasks. In this paper we present an approach to classify the impurities of a wheat sample to analyse the quality.

1 Introduction

The analysis of a wheat sample for determinating the grain quality is called “Besatz analysis.” The standard for this procedure is the manual sorting and weighting of the impurities of the sample by laboratory assistants or leading millers. This is an expensive, time-consuming and error-prone procedure. Our task was to automate this procedure by image processing in combination with intelligent machine learning algorithms. For the studies nearly 20,000 sample objects were pre-classified by human experts.. The image below shows the high variability on some sample classes. Flawless wheat kernels have a very high phenotypical variability like natural objects in general. Based on optical characteristics the boundaries between the various object classes in feature space are fluid. This leads to a very complex recognition problem.

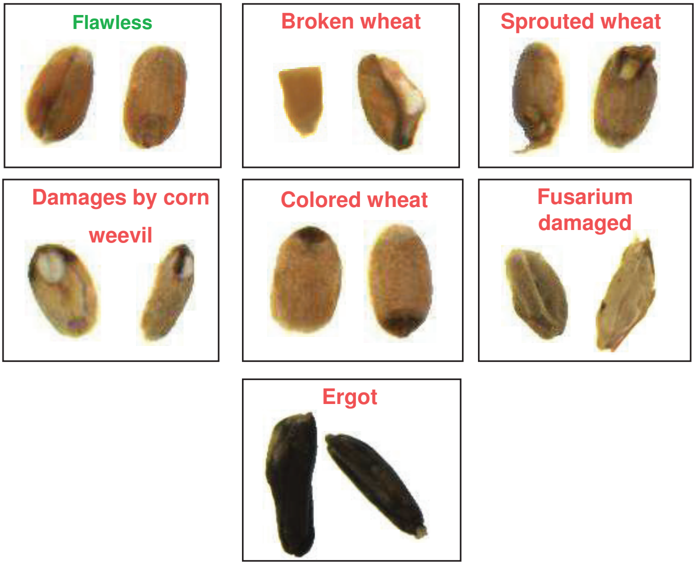


Figure 15.1: Sample objects for the different Besatz classes.

2 Hardware

For this complex optical recognition problem we need a stable image acquisition setup in terms of good single kernel separation and a stable and homogeneous object field illumination. For imaging every single object of a grain sample of nearly 500 g we used a setup consisting of a color line scan camera from the canadian manufacturer JAI with 2048 pixels running at nearly 2000 Hz line frequency and a Zeiss macro lens with a focal length of 50 mm. The object stream needs to be singularized before it passes the image acquisition unit because overlapping objects and occlusions result in recognition errors. Object singularization in the dimension perpendicular to the moving direction of the object stream was achieved by using partition walls. The separation in the dimension parallel to the moving direction was attained by two conveyer belts running at different speed. With this setup a sample with 500 g (nearly 10,000 single objects) can be analysed within 6 minutes. We achieved singularization rates of nearly 99 %. The whole setup is illustrated in the schematic diagram below.

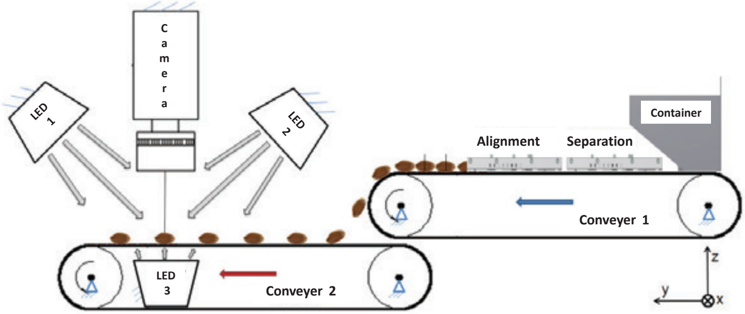


Figure 15.2: Setup for single kernel separation and image acquisition.

For object illumination a combination of transmitted (LED 3) and reflected light (LED 1 and 2) was used. The material of the conveyer belt is semi-transparent and thus a light source located thereunder provided a diffused background illumination. This generates a basis of the simplified object segmentation. A simple thresholding operation in combination with run length encoding makes a realtime line-wise segmentation possible. Right after the last row of the object image is transferred to the evaluation program the image can be classified by the classification module, implemented with a support vector machine [1]. The complete analysis is put into practice in terms of the European Commission Regulation (EC) No 856/2005 [2].

3 Image features

For an image recognition task in general, we have a high amount of possible image features. Detecting relevant features out of these is a very crucial step. In this experiment a bag-of-features with nearly 240 standard operators from the image processing library Halcon [3] as well as self-developed features based on texture information and gray value morphology [4] where used. For detecting relevant features the whole dataset with nearly 20,000 object images was separated randomly into 3 datasets – dataset 1 with 22 % of all available objects for feature scoring, dataset 2 with 45 % for training and dataset 3 with 33 % for the final test. Then the information gain [5] was calculated within dataset 1 which results in a relevance score for every single feature. Thus it is not clear at which threshold a feature can be considered to be irrelevant. This threshold can only be defined in combination with a classifier within a semi-wrapper approach. So we ordered our features in ascending order by the information gain score [5] and iteratively removed the 10 lowest ranked features from the feature vector, trained a classifier on dataset 1 and tested its performance on dataset 2. This results in the curve shown below. From this, the relevance threshold and consequently the final feature set could be estimated.

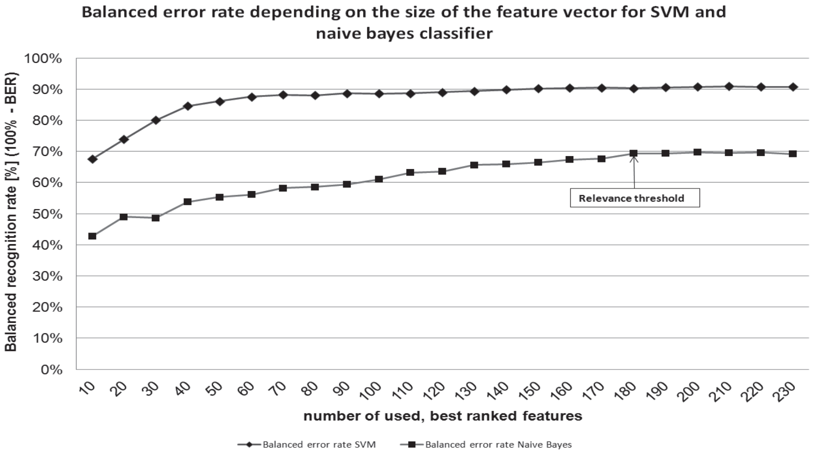
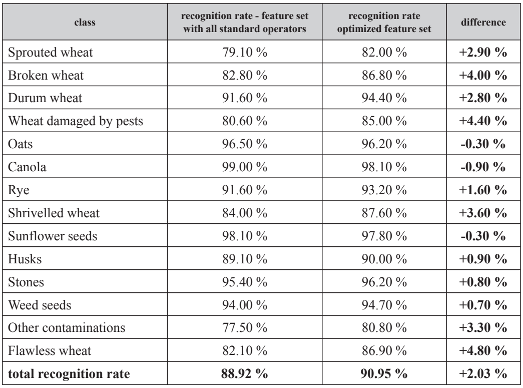


Figure 15.3: Total recognition rate on test set depending on feature set size.

The total recognition rate of about 91 % on the final test set shows that there is a significant improvement due to the introduction of problem-adapted image features. The reduction of the feature set to the most relevant features resulted into a significant speedup. In the new feature set problem-adapted features based on texture operators and morphological operations were introduced. This leads to the significant improvement. The results in the form of recognition rates accomplished with the feature set contraining of standard image operators in contrary to the optimized feature set are shown in Fig. 15.4.



class	recognition rate - feature set with all standard operators	recognition rate optimized feature set	difference
Sprouted wheat	79.10 %	82.00 %	**+2.90 %**
Broken wheat	82.80 %	86.80 %	**+4.00 %**
Durum wheat	91.60 %	94.40 %	**+2.80 %**
Wheat damaged by pests	80.60 %	85.00 %	**+4.40 %**
Oats	96.50 %	96.20 %	**-0.30 %**
Canola	99.00 %	98.10 %	**-0.90 %**
Rye	91.60 %	93.20 %	**+1.60 %**
Shrivelled wheat	84.00 %	87.60 %	**+3.60 %**
Sunflower seeds	98.10 %	97.80 %	**-0.30 %**
Husks	89.10 %	90.00 %	**+0.90 %**
Stones	95.40 %	96.20 %	**+0.80 %**
Weed seeds	94.00 %	94.70 %	**+0.70 %**
Other contaminations	77.50 %	80.80 %	**+3.30 %**
Flawless wheat	82.10 %	86.90 %	**+4.80 %**
total recognition rate	**88.92 %**	**90.95 %**	**+2.03 %**

Figure 15.4: Improvement of the recognition rate by feature optimization.

4 Classifier optimization

The SVM classifier is considered the most powerful classifier today. Tests indicated that the SVM will be the best classifier for our task also. So we used a SVM classifier with the radial basis function kernel (rbf):

$$k(x, x') = e^{-Gamma|x-x'|^2} \tag{15.1}$$

The kernel parameter Gamma and the regularization parameter Nu for the training of the SVM need to be chosen before the training process very carefully. To find an optimal parameter set a grid search method in combination with 3-fold crossvalidation on the training dataset with all available data was conducted. For the grid search optimization, the position of the grid points in the intervall [0, Max$_{Nu,Gamma}$] are calculated according to the following formula, with Nu as a threshold for the termination of the optimization process and Gamma as the kernel parameter:

$$Pos_{Nu,Gamma} = \frac{Max_{Nu,Gamma}}{2^{i-n+1}} \quad with \quad 0 \leq i \leq 9 \quad n = 10 \tag{15.2}$$

The formula leaves an exponential characteristic for the distribution of the nodes. The interval width grows with increasing values for Nu and Gamma. As expected, the optimal values of both parameters are small. To handle the imbalanced dataset the predictive power is measured in terms of the balanced recognition rate (BRR). This measure is defined as the average of the recognition rates of each classes. For many recognition problems with natural material the influence of Gamma is much higher than of Nu. Nu controls the training set error as well as the number of training vectors which become support vectors and thus affects the decision border to become more complex. With growing Gamma the influence of each of these support vectors on the decision border grows. To cover the high intra-class-variance in the dataset many relevant training vectors are needed with each relatively less but equal distributed influence among them on the decision border. This results in a low value for Gamma and a mid-sized for Nu (see figure below). As we see from the figure for this recognition problem we can also choose a simpler optimization strategy because of the low influence by Nu. We can take a fixed low value for Nu and increase the value for Gamma till the recognition rate starts to decrease.

5 Complexity visualization with Principle Component Analysis (PCA)

The given recognition task and the results so far indicate a very high complexity. One way to understand the complexity of the recognition problem is to establish a suitable visualization of the distribution of the different object clusters in the feature space. The feature space has got 200 or more dimensions and can not be visualised in an easy way. To reduce the dimensions of the feature space a principle component analysis (PCA) [6] could be used. The PCA allows the visualisation of the given high dimensional data in a lower dimensional space with loss of information. The goal of the PCA is the approximation of the n features by a smaller number m of meaningful linear combinations (principal components). On the given dataset a visualisation of the object clusters in feature space was realised with Matlab® for the subclasses of wheat. The result of the PCA confirmed the high complexity and further indicated the need to use a powerful classifier which is able to handle com-

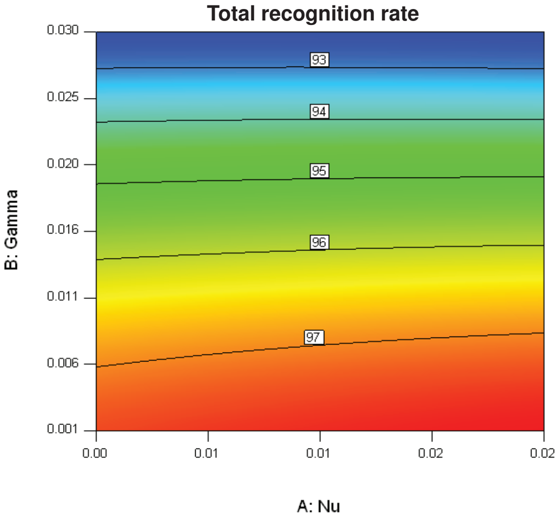


Figure 15.5: Total recognition rate in dependency of Nu and Gamma.

plex borders between the classes in feature space. The PCA can be seen as a quick and simple way to get a first impression on the complexity of the recognition problem, especially in line with a feasibility study.

It can be clearly seen that the subclasses of wheat in Fig. 15.6 show far more overlap in feature space after PCA due to a higher intra-class-variability in combination with less inter-class-variability. The clusters of the different grain types (Fig. 15.7) can be visualy separated also in the 3-dimensional space after PCA. For Further results about complexity visualisation you can refer to [7].

The task to separate foreign grain and also weed seeds from flawless grain can be implemented with more simple color-based image features like done within optical sorting machines. But in practice this also comes along with a higher false-positve rate for objects not belonging to the class of flawless wheat. For the purposes of sorting this is tolerable; for the purposes of analysis, the rate of false-positives is considered to high to be reliable.

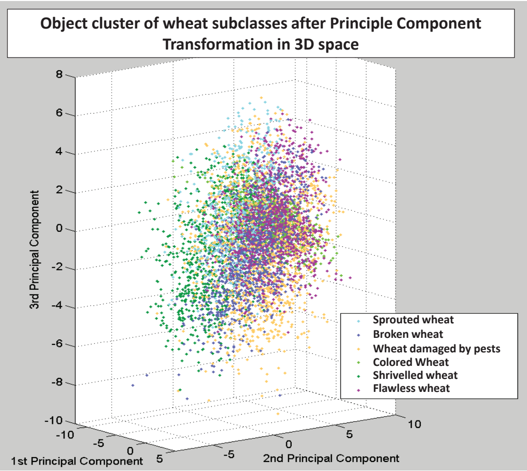


Figure 15.6: Cluster of the wheat subclasses after PCA in 3-dimensional space.

6 Results

After this optimisation the dataset has been expanded with toxic ergot and fusarium-damaged wheat and further samples for every class. With optimized features and optimized SVM parameters (regularization parameter Nu and parameter Gamma of the radial basis function kernel) we created a classifier for practical testing. It could be demonstrated that the whole system consisting of automatic sample separation, single kernel imaging and classification is able to achieve a high accuracy of recognition. Therefore, sample material with known composition from different crop years, also from later years than the material used for training has been analysed in a comprehensive test. It turned out that recognition rates of 90 % could be reached.

The whole system was also tested within a field test in two grain mills. The results confirmed the findings from the presented laboratory test and validated the applicability of the system.

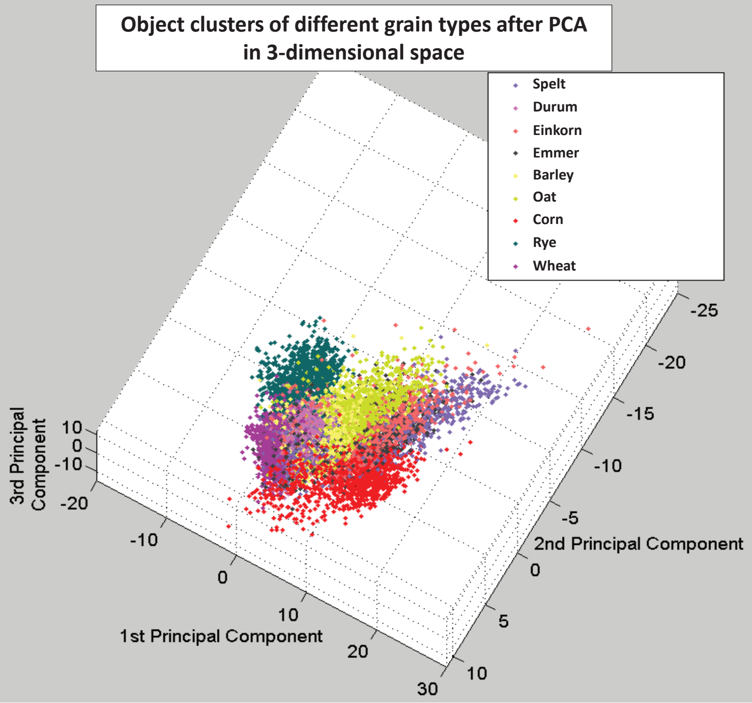


Figure 15.7: Cluster of the different grain types after PCA in 3-dimensional space.

7 Summary

The automatic recognition of natural material like grain is always a challenging task. Recent advantages in the development of new classification algorithms like support vector machines, random forest classifiers and neural networks now make it possible to solve many of these problems. With our setup for objects separation and image acquisition it is possible to acquire the images of 10,000 objects of a wheat sample in less than 10 minutes. With problem-adapted feature extraction and a highly optimized classifier it is possible to reach an accuracy of nearly 90 % under practice conditions. This seems to be a very good result. We as-

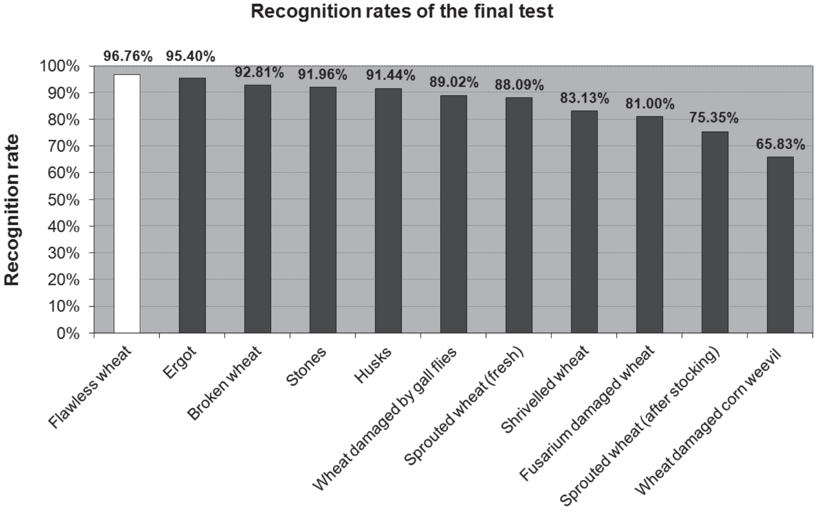


Figure 15.8: Recognition rates per class.

sume that further improvements come up by using a higher resolution, e. g. a colour line scan camera with 4096 pixels. With a decreasing price hyperspectral cameras can also be considered in the future for this problem. Especially ultra-violet light can deliver information for the recognition of broken wheat and other whear subclasses. Also the detection of fusarium seems to be improvable by using spectral information. Recent research, for example [8] strongly indicates this. Discussions with future users of the system indicated that the time for an analysis is sufficient but the precision could be improved. Accordingly the main focus in future research will be the increase of the recognition rate.

Acknowledgements

A project, funded by the Federal Ministry for Economic Affairs and Technology under the promotional reference 16INO496, forms the basis of this paper. The responsibility for the content of this paper lies with the author.

References

1. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2000.

2. N.U., "Commission regulation (ec) no 856/2005," *Official Journal of the European Union*, vol. 143, no. 3, 2005.

3. ——, *Halcon 8.0.3 Reference Manual*. Munich: MVTec Software GmbH, 2009.

4. D. Garten, *Einfluss von Bildaufnahme und Bildmerkmalen auf die Erkennungsguete bei der automatischen Besatzanalyse von Brotweizen (Influence of image acquisition and image features on the recognition rate for the automatical Besatz analysis of wheat)*. Ilmenau: ISLE Verlag Ilmenau, 2011.

5. T. M. Mitchell, *Machine Learning*. Pittsburgh: The Mc-Graw-Hill Companies, Inc, 1997.

6. I. T. Jolliffe, *Principal Component Analysis, Series: Springer Series in Statistics*. New York: Springer, 2002.

7. K. Anding, *Automatisierte Qualitätssicherung von Getreide mit überwachten Lernverfahren in der Bildverarbeitung (Automated quality assurance with supervised machinge learning in image processing)*. Ilmenau: ISLE Verlag Ilmenau, 2010.

8. G. Polder, "Detection of fusarium in single wheat kernels using spectral," *Seed Science and Technology*, vol. 33, pp. 655–668, 2004.