Feature extraction techniques to use in cereal classification

Ole Mathis Kruse, IMT

Department of Mathematical Sciences and Technology
Problem

- Is it possible to discriminate between different species- or varieties of cereal grains - using image analysis?

![Barley](image1)

![Oat](image2)

![Wheat](image3)

![Wheat - Mjølner](image4)

![Wheat - Bjørke](image5)
The images

- Images are taken with a high resolution (1280 x 1024) digital camera.
- Three images (rotated 120°) of each sample.
- Each image splitted in four
  - Larger data material.
  - Easy to implement cross-validation.
- 204 images for species analysis.
- 84 images for variety analysis.
Texture analysis

- How to discriminate the samples?
- Single grain analysis
  - Measure size, roundness, colour etc. for each grain.
  - Extraction of single grain is difficult.

- Texture analysis
  - Finds features for the texture of the image surface.
  - Easy to implement.
  - Several different features available.
Feature detectors

- Some examples of feature detectors
  - Angle Measure Technique (AMT)
    Images are folded out to an intensity vector. A circle is placed at random points on the vector, and the angle between the intersections are calculated. Parameters are calculated from these angles.
  - Histogram statistics
    Several (~10) statistical parameters are calculated for each image. Some of these parameters are based on the intensity distribution of the images.

From: Esbensen, Hjelmen & Kvaal (1996)
Feature detectors continued

- Gray Level Co-occurrence Matrix (GLCM)
  Builds a new matrix that counts the number of different neighbouring relations. Four statistical parameters are calculated from the GLCM matrix.

- Singular Value Decomposition (SVD)
  From Linear Algebra, a matrix M (e.g., an image) can be factorized to the form $M = USV^*$. $U$, $S$, and $V^*$ are matrices which capture different characteristics of the image. We only use the $S$ matrix, which has nonzero values only on the main diagonal. These nonzero values are the singular values of $M$ and can be thought of as scalar “gain controls.”

- There are lots of other methods, and also variants of the methods shown here.
Output from feature detectors

- All methods give a matrix that ranges from 4 to 1000’s of columns with one row for each image

<table>
<thead>
<tr>
<th></th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$V_3$</th>
<th>$V_4$</th>
<th>$V_5$</th>
<th>$\ldots$</th>
<th>$\ldots$</th>
<th>$\ldots$</th>
<th>$\ldots$</th>
<th>$\ldots$</th>
<th>$V_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ldots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- A complicated dataset with many variables
- Difficult to analyse with univariate statistics (e.g. ANOVA)
- Must use multivariate techniques
Multivariate statistics

- Multivariate statistics analyses all variables at the same time and finds patterns that describe the variability
- Covariance between variables is taken into account

- Several different multivariate statistical methods
  - Principal Component Analysis (PCA)
  - Partial Least Square Discriminant Analysis (PLSDA)

Fra: http://en.wikipedia.org/wiki/Principal_component_analysis
Principal Component Analysis (PCA)

- Typical result from PCA analysis

![Typical result from PCA analysis](image)
Partial Least Square Discriminant Analysis (PLSDA)

- PLS is a multivariate regression technique
- PLSDA allows discrimination/classification of the data
Classification results from different feature detectors

- Compare the PLSDA classification with the known species/variety.
- Calculate % correct classification for each species/variety for the different feature detectors.

<table>
<thead>
<tr>
<th></th>
<th>Barley</th>
<th>Oat</th>
<th>Wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMT</td>
<td>90</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Histogram statistics</td>
<td>90</td>
<td>92</td>
<td>83</td>
</tr>
<tr>
<td>GLCM</td>
<td>85</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>SVD</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Bjørke</th>
<th>Magnifik</th>
<th>Mjølner</th>
<th>Olivin</th>
<th>Polka</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMT</td>
<td>71</td>
<td>25</td>
<td>58</td>
<td>33</td>
<td>83</td>
</tr>
<tr>
<td>Histogram statistics</td>
<td>100</td>
<td>0</td>
<td>54</td>
<td>67</td>
<td>50</td>
</tr>
<tr>
<td>GLCM</td>
<td>88</td>
<td>0</td>
<td>71</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>SVD</td>
<td>38</td>
<td>67</td>
<td>67</td>
<td>0</td>
<td>83</td>
</tr>
</tbody>
</table>
Summary

- Using 12 images of each sample gives better data quality.
- Feature detectors are used to extract texture information from the images.
- Multivariate statistics are used to analyse the feature data.

- The cereal grain species are classified quite well.
- SVD is the best detector.
- Classification of wheat is strongly dependent on the varieties and the different feature detectors.