Lappeenranta University of Technology

LUT School of Industrial Engineering and Management

Department of Software Engineering and Information Management

Bachelor's thesis

Wood species identification by visual characterization of fibers

Supervisor and examiner: D.Sc. Tuomas Eerola

Kimmo Kerminen

Korpimetsänkatu 5 D 11

53850 Lappeenranta, Finland

kimmo.kerminen@lut.fi

Telephone: +358407659836

ABSTRACT

Lappeenranta University of Technology

LUT School of Industrial Engineering and Management

Department of Software Engineering and Information Management

Kimmo Kerminen

Wood species identification by visual characterization of fibers

Bachelor's Thesis

2014

26 pages, 6 figures and 6 tables

Supervisor and examiner: D.Sc. Tuomas Eerola

Keywords: machine vision, fiber characterization, classification

The focus of this thesis was to study image-based wood fiber identification, i.e. classification of fiber images and individual fibers to different wood species based on the visual properties of the fibers. This thesis was part of the PulpVision research project and continued from a previous study done for this project, which was a segmentation method for wood fibers in pulp suspension images. The fibers were segmented from the images using this method and selected features were measured from the fibers. With the measurements we used two different classification methods to classify the images. Experiments were done with different combinations of the selected features. The results showed that the developed identification method worked well for fiber images (98.4 % accuracy) but not for individual fibers (44.49 % accuracy).

TIIVISTELMÄ

Lappeenrannan Teknillinen Yliopisto

Tuotantotalouden tiedekunta

Ohjelmistotuotanto ja tiedonhallinta

Kimmo Kerminen

Puulajien tunnistus kuitujen visuaalisen luokittelun avulla

Kandidaatintyö

2014

26 sivua, 6 kuvaa ja 6 taulukkoa

Ohjaaja ja tarkastaja: TkT Tuomas Eerola

Hakusanat: konenäkö, kuitujen karakterisointi, luokittelu

Tämän työn tarkoituksena oli tutkia kuvapohjaista puukuitujen tunnistamista, eli eri puulaatujen kuitukuvien sekä yksittäisten kuitujen luokittelua kuitujen visuaalisten piirteiden avulla. Tämä työ tehtiin osana PulpVision-tutkimusprojektia ja tämä työ oli suoraa jatkoa tämän tutkimusprojektin aiemmalla työlle, jossa kehitettiin puukuitujen segmentointimetodi. Kuidut segmentoitiin kuvista tämän metodin avulla ja segmentoiduille laskettiin erilaisia piirteitä. Näiden piirteiden avulla testattiin kahta eri luokittelumetodia erilaisilla piirteiden kombinaatioilla. Saatujen tulosten perusteella voitiin huomata valittujen luokittelumetodien toimineen hyvin kuitukuvien luokittelussa (98,4 % tarkkuus), mutta heikosti yksittäisen kuitujen luokittelussa (44,49 % tarkkuus).

ABBREVIATIONS

Averages	Avg.
CEMIS	Centre for Measurement and Information Systems
k-NN	K-nearest neighborhood
MVPR	Machine Vision and Pattern Recognition
SD	Standard deviation

LIST OF FIGURES

Figure 1. Fibers from the pinus species.

Figure 2. Fibers from birch.

Figure 3. Fibers from wheat.

Figure 4. Polarity map example.

Figure 5. Representation of two different lengths on a fiber.

Figure 6. Fiber image with correctly segmented fibers highlighted in red.

LIST OF TABLES

Table 1. Average lengths and widths of common non-wood fibers and their uses.

Table 2. Fiber statistics.

Table 3. Classification results using the *k*-NN classifier with statistical features.

Table 4. Classification results using the *k*-NN classifier with individual fibers.

Table 5. Classification results using the naïve Bayes classifier with statistical features.

Table 6. Classification results using the naïve Bayes classifier with individual fibers.

CONTENTS

ABSTRACT II
TIIVISTELMÄ III
ABBREVIATIONS IV
LIST OF FIGURES V
LIST OF TABLES VI
1 INTRODUCTION 2
2 FIBER CHARACTERISTICS 3
2.1 Wood fibers 3
2.1.1 Softwood 3
2.1.1 Hardwood 4
2.2 Non-wood fibers 5
3 MACHINE VISION SYSTEM FOR SPECIES IDENTIFICATION 8
3.1 Fiber segmentation 8
3.1.1 Oriented edge map 8
3.1.2 Tensor voting 8
3.1.3 Curve growing 10
3.2 Fiber feature extraction 10
3.3 Classification 11
3.3.1 *k*-NN 12
3.3.2 Naïve Bayes 12
4 EXPERIMENTS 14
4.1 Data 14
4.2 *k*-NN results with statistical data 16
4.3 k-NN results with measurements of individual fibers 18
4.4 Naïve Bayes results with statistical data 19
4.5 Naïve Bayes results with measurements of individual fibers 21
5 CONCLUSIONS 23
REFERENCES 25

1 INTRODUCTION

This bachelor's thesis was part of the PulpVision ("PulpVision - Image Processing and Analysis Methods for Pulp Process Measurements," n.d.) research project. The PulpVision research project focuses on the development of image-based measurement and characterization methods related to the quality of pulp as a raw material for papermaking. PulpVision develops methods to manage the processing of raw material for fiber-based products. The project develops both entirely new measurements characterizing structures and quality, and aims for transferring research in image-based measurements and off-line methods to on-line measurements and for process quality control.

The purpose of this thesis was to continue from the previous work (Strokina et al., 2013). In this work a method for segmentation of curvilinear structures was developed. This thesis utilized this segmentation method to segment fibers from pulp suspension images and searched and developed different features to characterize the fibers. Based on these features a species identification method using these features was developed. The method was used to classify the species of fibers in the suspension images. The images contain fibers from one non-wood raw material, wheat, and from four different wood species, which are acacia, birch, eucalyptus and pine.

The thesis is organized as follows. Chapter 2 includes a literature overview on wood and non-wood fiber characteristics and their differences. Chapter 3 consists of a short review of the method to segment fibers from suspension images and descriptions on what features and methods were used to characterize and classify the fibers. In chapter 4 the results of the experiments are presented. Finally, in the 5 chapter conclusions are presented.

2 FIBER CHARACTERISTICS

In this chapter we will go through main characteristics of different fibers used in papermaking.

2.1 Wood fibers

In general different wood species can be divided into hardwoods and softwoods, both of which have great differences in fiber morphology, chemical composition. These characteristics affect the qualities of the pulp and the end product. For example, in mechanical pulping, hardwood pulps exhibit good light-scattering power and sheet surface properties. On the other hand, the strength properties are usually poor and therefore hardwood chips are pretreated chemically to balance the strength and optical properties (Lönnberg, 2009).

2.1.1 Softwood

Softwood fibers are generally longer than hardwood fibers, have larger seasonal variations between earlywood and latewood fibers, and the age and growing conditions of the stem. Some of the commonly used softwoods include spruces, firs, pines, hemlock and larches (Borch, 2002). In Figure 1, we can see fibers from the pinus species.

The typical length of a softwood fiber is 100 times its width, and the width is approximately ten times the wall thickness (Lönnberg, 2009). The average lengths of the fibers vary between 2-6 mm and their width between 20-60 µm in commercial pulpwood, cell wall thickness varies between 2-10 µm.

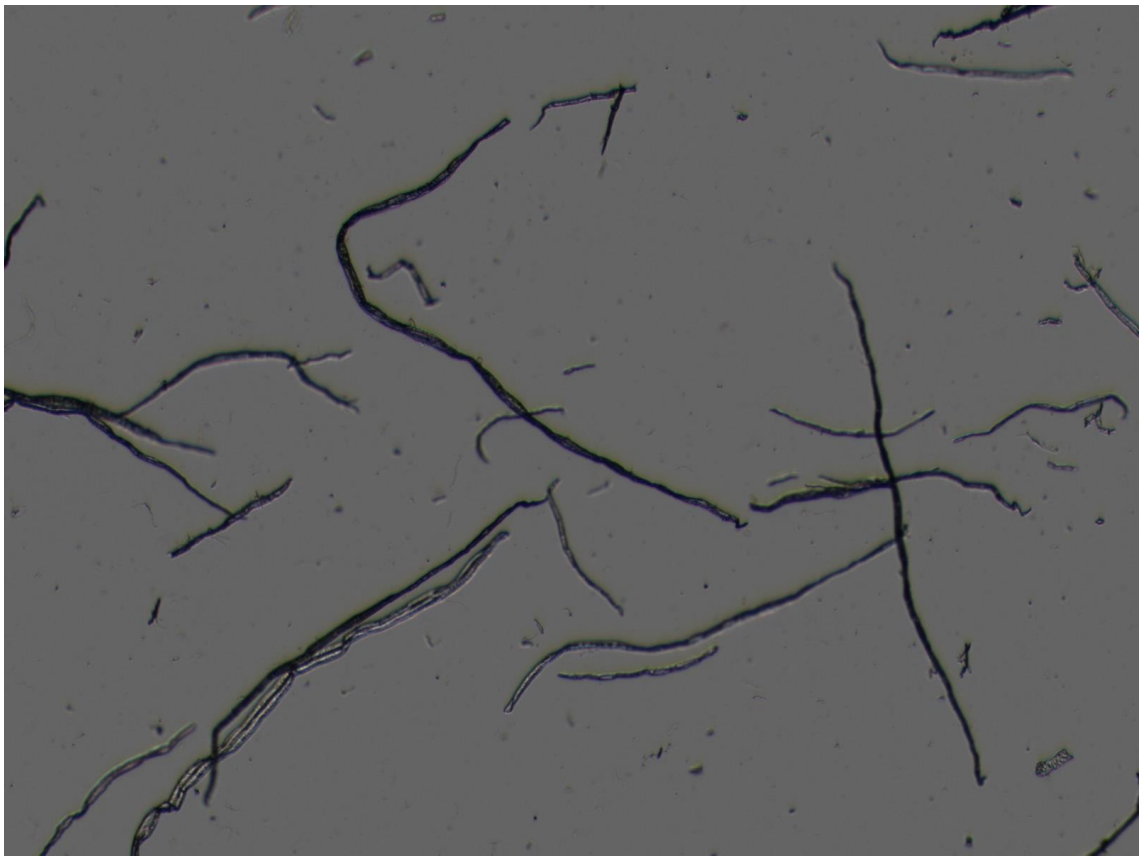


Figure 1. Fibers from the pinus species.

2.1.1 Hardwood

As stated above, hardwood fibers have no clear seasonal variations as the softwood fibers have. Typical dimensions for hardwood fibers are 0.7-1.7 mm for length, 14-40 µm for width and 3-4 µm for cell wall thickness (Borch, 2002). Some common used hardwood species include aspen, poplar and eucalyptus (Lönnberg, 2009). In Figure 2, birch fibers are shown.

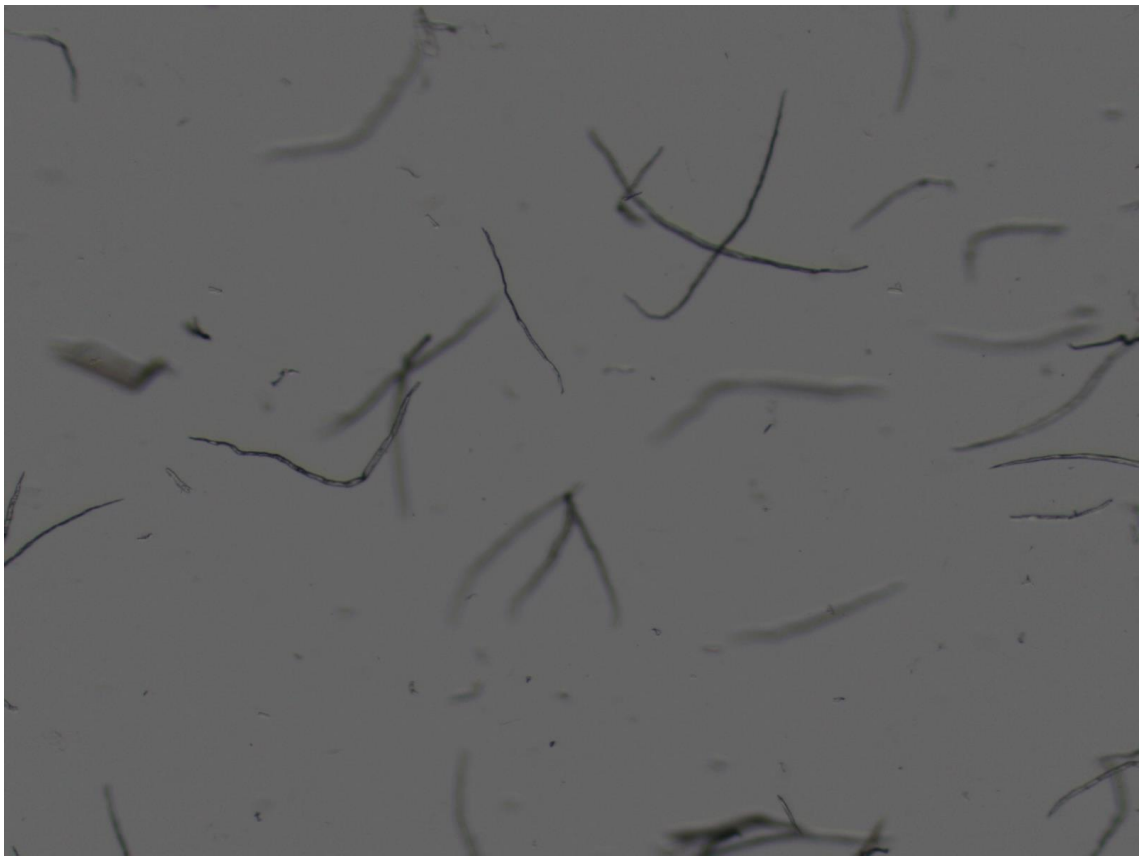


Figure 2. Fibers from birch

2.2 Non-wood fibers

In papermaking today, wood fibers are the most commonly used raw material with a 94% share of all papermaking fiber used. However until the 1880s, the only source of papermaking fibers were non-wood fibers. Also, non-wood fibers are still important raw material in papermaking, especially in regions that suffer from wood shortage (Borch, 2002). For example, in China, 20% of the raw material used for paper and paperboard manufacture come from non-wood plants (Guo et al., 2009). Figure 3 shows fibers from wheat.

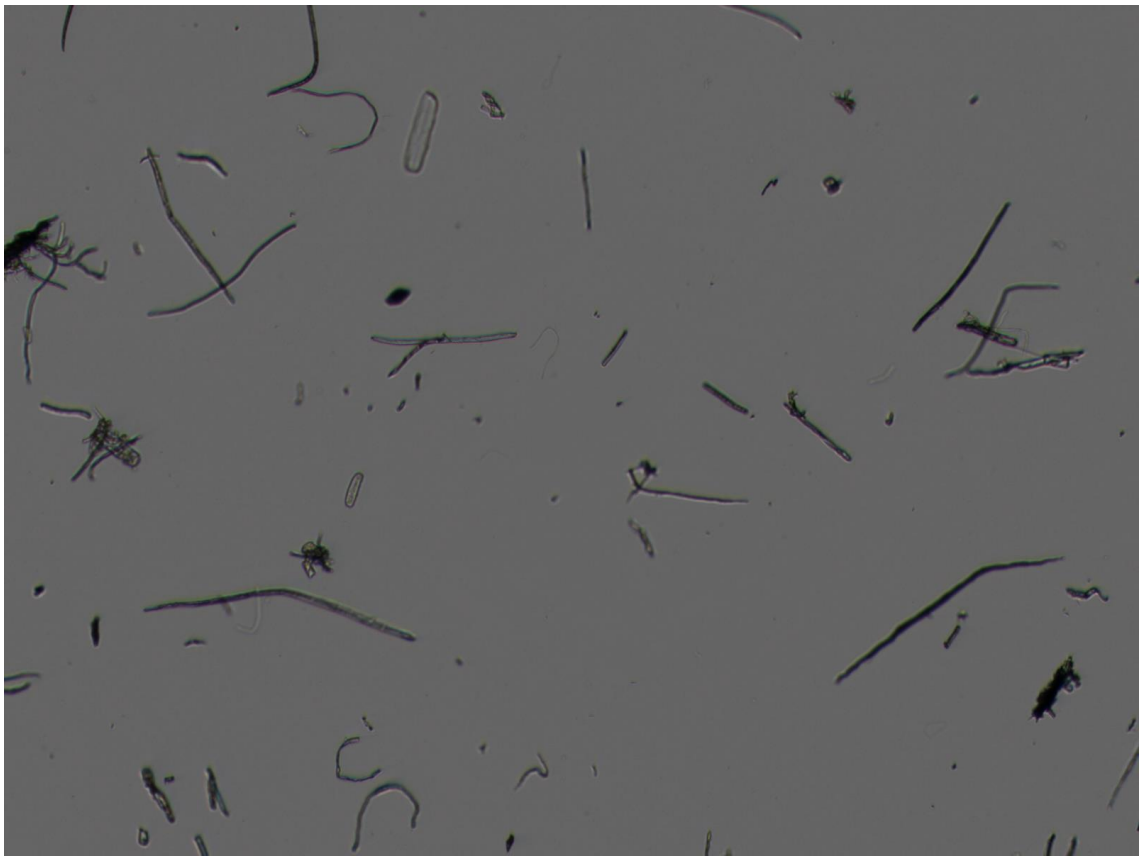


Figure 3. Fibers from wheat.

The following four categories (Swann, 2006) are listed as non-wood fiber sources by the Conservatree (“Conservatree,” n.d.) organization:

1) Dedicated crops grown specifically for paper fiber, e.g. hemp and jute
2) Agricultural food production residues, e.g. wheat straw
3) Industrial residues, e.g. cotton or linen cloth scraps
4) Naturally occurring uncultivated crops, e.g. wild grasses

Different non-wood fibers are used for making different kinds of paper, for example tissue paper, wrapping paper, cigarette paper and newsprint paper. In Table 1 are listed the average fiber length and width of some common non-wood fibers along with the type of paper or paperboard that uses fibers from these sources (Chandra, 1998).

Table 1. Average lengths and widths of common non-wood fibers and their uses

Fiber source	Fiber length (mm)	Fiber width (μm)	Type of paper or paperboard
Jute	2.5	20	Printing, writing, wrapping and bag
Cotton fibers	25	20	High grade bond, ledger, book and writing
Sabai grass	4.1	13	High grade book and printing
Wheat straw	1.4	15	Printing, writing, glassine, greaseproof, duplex, triplex, corrugating medium, strawboard and wrapping

3 MACHINE VISION SYSTEM FOR SPECIES IDENTIFICATION

In this chapter we will go through the fiber segmentation method used in this work, the feature extraction methods and finally the used classification methods.

3.1 Fiber segmentation

The method to segment fibers was developed in the previous work by (Strokina et al., 2013). The method detects curvilinear structures by tensor voting and it was applied to fiber characterization. The method has the following steps:

1) A grayscale image is reduced to an oriented edge map
2) Tensor voting is applied to the edge map
3) Curves are grown from the most salient points

3.1.1 Oriented edge map

To obtain the oriented edge map from a grayscale image, the image is filtered by a second derivative zero mean Gaussian filter in eight directions and the dominant orientation of the edge normal in each pixel is computed as the maximum of the eight filter responses (Lindeberg, 1998). Finally, non-maximum suppression together with hysteresis thresholding (Canny, 1986) is performed in the dominant orientation of the edge normal, thus producing an oriented edge map.

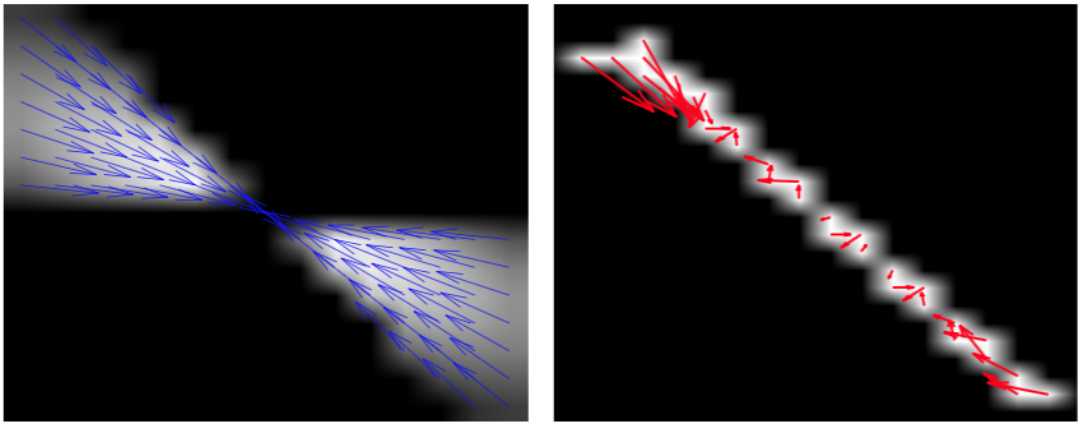
3.1.2 Tensor voting

Once the edge map has been computed, tensor voting (Medioni et al., 2000) is applied to retrieve point saliency, end points, and junction points.

Point saliency indicates how likely a pixel in the image belongs to a curvilinear structure, tensor voting is applied to compute the saliency. A tensor, which encodes the

curve orientation of a pixel in matrix form, is associated with each pixel. Then, each pixel votes for its neighbors in its voting field. The voting field tells us the most likely direction from which we can find pixels that belong to the same curve, the field is oriented along the tangent to the curve in the pixel. A voter's tensor is added to the tensors of the pixels in the voting field multiplied by the field coefficient in the voting process. The field coefficient is computed based on the distance and angle to the voter. The saliency of a pixel can be determined after the voting process by calculating the difference between the bigger and the smaller eigenvalues of the tensor. Also, the smaller eigenvalue indicates how likely that pixel is a junction point (Strokina et al., 2013).

Finally, pixel polarity information (Wai-Shun Tong et al., 2004) can be used to find the end points. Polarity vector of a pixel indicates the direction from which the majority of the votes come from. Therefore, votes coming from only one direction would indicate that the pixel in question is an end point. In Figure 4 we can see an example of a polarity map achieved with this method.



(a) (b)

Figure 4. Polarity map example: (a) Polarity vectors created by one voting pixel; (b) The polarity map

3.1.3 Curve growing

With the information produced from tensor voting, curves can be grown by choosing an unprocessed seed point with high saliency and iteratively growing the curve following the estimated tangent direction (Medioni and Kang, 2004). From there, points are added to the curve if it has the maximum saliency in the tangent direction.

According to (Strokina et al., 2013), information about the junction points and the polarity of the points can be used to separate two intersecting curves. At a region of a junction, the direction of curve growing stays the same as before the junction since there is no certainty of the pixel orientation in the junction region.

3.2 Fiber feature extraction

After the curves are found using the method described in Sec. 3.1. The following features can be calculated on the fibers: length, width, curl index and number of fibers found in the image. These features were selected based on visual inspection of the fibers images of different species and the information in the second chapter.

By using the curve points achieved from the used segmentation method, the length (Strokina et al., 2013), *L*, of the fiber could be computed as the sum of the distances between these points as

$$\mathrm{L} = \sum_{\mathrm{i}=2}^{\mathrm{N}} \sqrt{(\mathrm{x_i} - \mathrm{x_{i-1}})^2 + (\mathrm{y_i} - \mathrm{y_{i-1}})^2}, \tag{1}$$

where $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ represent the coordinates of the pixels.

For the curl index (Page et al., 1985) the projected length of the fiber, *l*, needed to be calculated. This was estimated by computing the distance between the end points of the curve as

$$\mathrm{l} = \sqrt{(\mathrm{x_n} - \mathrm{x_1})^2 + (\mathrm{y_n} - \mathrm{y_1})^2} \tag{2}$$

Figure 5 shows us a representation of the two different lengths.

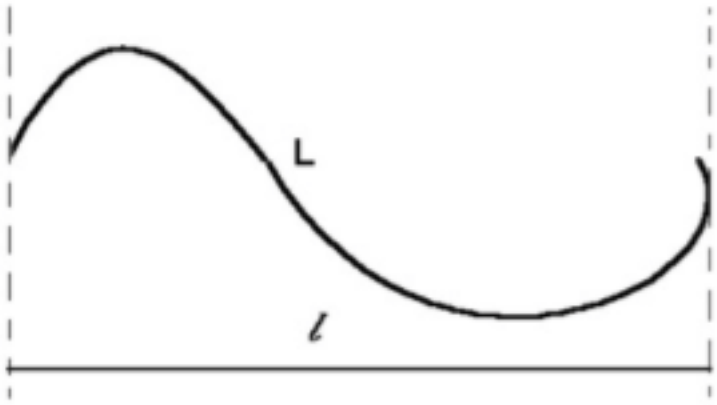


Figure 5. Representation of two different lengths on a fiber.

From here, the curl index can be computed as the ratio between the full length and the projected length as

$$\mathrm{CI} = \frac{\mathrm{L}}{\mathrm{l}} - 1 \tag{3}$$

The width (Kurakina, 2012) of the fiber is calculated based on the coordinates of the curve points and a binary image. The curve points were distributed equally and the width was estimated at each point. The average of these widths was considered to be the width of the fiber.

3.3 Classification

Based on the fiber features described in Sec. 3.2, the wood species can be identified using classification techniques. In this work, the following classification methods were used: *k*-NN (*k*-nearest neighbor) and naïve Bayes classifier.

3.3.1 *k*-NN

The *k*-NN algorithm (Altman, 1992) is a non-parametric method used for classification and regression. In classification, an object is classified based on majority vote of its neighbors, meaning that the object in question is assigned to a class that forms the majority among its *k* nearest neighbors, the parameter *k* is a positive integer

The distance between the object to be classified and its neighbors is calculated using the Euclidean distance. The Euclidean distance can be calculated as

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}\,, \tag{4}$$

where $p = (p_1, p_2, \ldots, p_n)$ and q = $(q_1, q_2, \ldots, q_n)$ are two points in Euclidean n-space.

3.3.2 Naïve Bayes

Naïve Bayes classifier is a simple probabilistic classifier and it is based on applying Bayes' theorem, which can be stated mathematically as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{5}$$

where *A* is a proposition and *B* is the evidence. *P(A)* is the prior probability, i.e. initial degree of belief in *A*, *P(A|B)* is the conditional probability, i.e. the degree of belief in *A* when *B* is known. The quotient *P(B|A)/P(B)* represents the support *B* provides for *A*.

All practical learning algorithms based on Bayes' theorem ("Bayes' Theorem (Stanford Encyclopedia of Philosophy)," n.d.) make some independence assumptions but the naïve Bayes method assumes that the attributes are statistically independent given the class (Bramer et al., 2007, pp. 59–70). Despite the strong assumptions, the naïve Bayes classifier performs well on many classification problems and one of its advantages is its small requirement for training data (Domingos and Pazzani, 1997).

The classification in naïve Bayes is based on estimating P(x|y), the probability or probability density of the features x given the class y. The required probability

distribution can be constructed in various ways. In this work, two approaches were used: Gaussian distribution and kernel distribution.

The normal (Gaussian) distribution estimates a separate normal distributions for each class by computing the mean and standard deviation of the training data in that class. In the kernel distribution, a separate kernel density estimate is computed for each class based on the training data for that class. The kernel density estimation, also called Parzen window density estimation, approximates the probability density function of a given sample by determining the number of observations within a certain area.

4 EXPERIMENTS

In this chapter we will go through the experiments. Firstly, we will go through the data we used in the experiments and then we will go through the results of the experiments. In the first round of experiments, statistical data of the fiber images were used to classify other fiber images. In the second round experiments, measurements of individual fibers were used to classify other individual fibers, in these experiments the training and testing data consisted of 350 fibers from each species. The results are presented in tables that list the overall accuracy of each used feature set.

4.1 Data

As mentioned in Chapter 1, the pulp suspension images used in the experiments contain fibers from acacia, birch, eucalyptus, pine, and wheat. Each image contains fibers from one species only. From each species, 50 images were chosen to construct the training set and 50 images for the testing set. The images were provided by the CEMIS-OULU (Center for Measurement and Information Systems) laboratory of University of Oulu ("CEMIS - CEMIS-OULU," n.d.). The bit depths and resolution of the images were 24 bits and 1600x800 pixels respectively; flat field correction was also applied in the images.

The fibers were segmented from the images with the segmentation method by Strokina et al. (2013). The segmentations were manually checked and only those fibers that were correctly segmented were chosen. In Figure 6 is an example of a fiber image of the acacia species with highlights on correctly segmented fibers. Once the fibers were segmented, features of the fibers were extracted and measured as described in Sec. 3.2. After the measurements were calculated, the average and standard deviation of each feature were calculated. This process was repeated on each image.

For implementing the identification method we have used MATLAB, which was also used to implement the segmentation method developed by Strokina et al. (2013).

Table 2 presents some statistics of the fibers in the test images. These statistics include the average fiber count per image and averages of length, width and curl index.

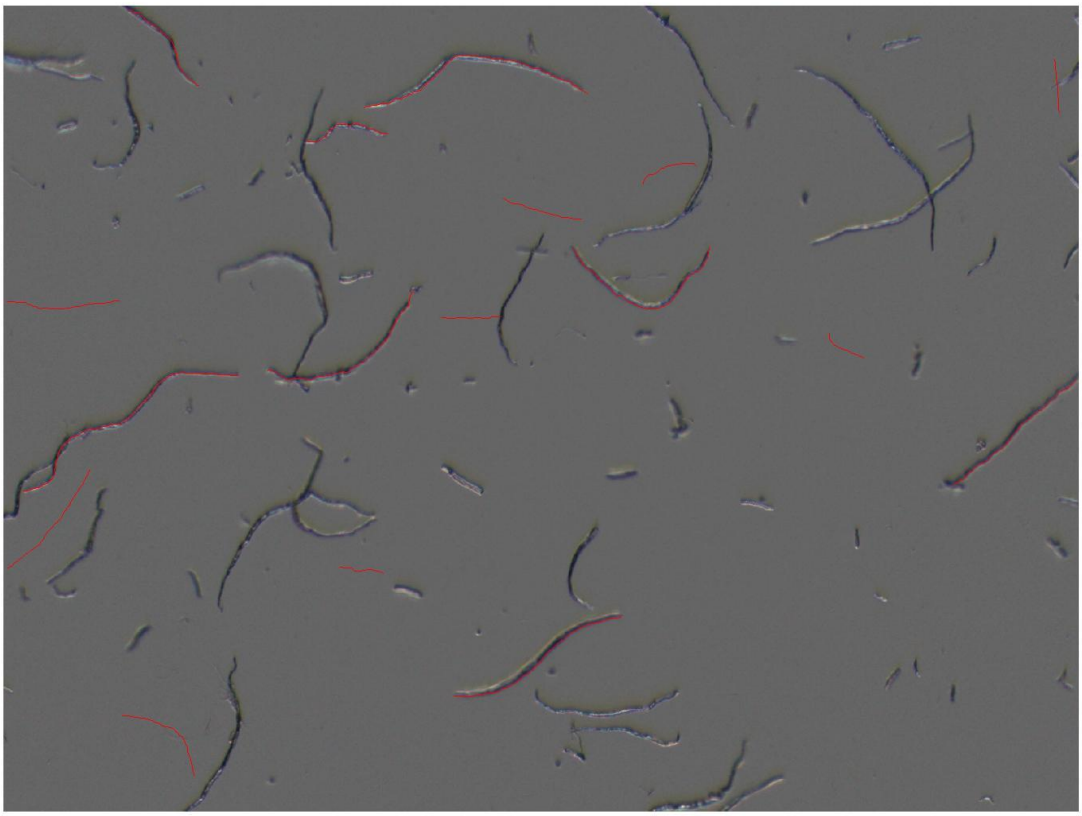


Figure 6. Fiber image with correctly segmented fibers highlighted in red

Table 2. Fiber statistics.

Species	Average fiber count per image	Average length (pixels)	Average width (pixels)	Average curl index
Acacia	22	199.47	8.32	1.59
Birch	18	249.40	10.95	0.72
Eucalyptus	28	226.26	8.68	1.47
Pinus	10	345.06	10.40	5.23
Wheat	18	161.84	8.37	1.14

From this initial data we can make some observations on the differences between each species. The pinus has longer fibers on average than other species. In width the differences are much more subtle. The pinus has a very high curl index compared to

others. This is due to the fact that the pinus has longer fibers in general, so when these long fibers are curved it increases the curl index. We can also observe a noticeable difference in the average number of fibers per image between different species. However, this is more a feature of the data we have used rather than a feature of an individual species.

Next we will go through the classification results using the k-NN and the Naïve Bayes classifiers.

4.2 *k*-NN results with statistical data

With the *k*-NN classifier we used different values for *k* (3, 5, and 7) to test its effect on the accuracy of the classification. In Table 3, we can see the overall accuracies of the classification results using different statistical features and different values of *k*.

First off, we can observe that using the averages of length alone we achieved 98 % accuracy ($k = 7$). This is also one of the highest accuracies achieved in the experiments, which would indicate that it is the best feature to depict the fiber images. However using the standard deviations of length the results were less impressive with an accuracy of 46.8 % ($k = 3$). This was also the case with other single features experiments, i.e. the classifications were less accurate using the standard deviations. This suggests that the standard deviations are quite similar for each species making it difficult to classify them based on that data only.

In the other experiments, using a single feature only, averages of width performed quite well with the accuracy of 65.2 % ($k = 5$). The least accurate feature was the curl index, the best accuracy being 39.2 % ($k = 7$) using the averages of curl index and the least accurate being 22 % ($k = 5$) using the standard deviations of curl index. This suggests that the curl index is very similar between different species so there is not enough variation to make accurate classifications.

Table 3. Classification results using the *k*-NN classifier with statistical features.

Features used	Accuracy (%)		
	k = 3	k = 5	k = 7
Averages (Avg.) of length	**96**	**96**	**98**
Avg. of width	63.2	65.2	64.4
Avg. of curl index	35.6	37.2	39.2
SD of length	46.8	45.6	45.2
SD of width	36	42	44.8
SD of curl index	23.2	22	22.8
Avg. of length and width	**96**	**96**	**98**
Avg. of length and curl index	**96**	**96.8**	**98**
Avg. of width and curl index	62	59.6	55.2
SD of length and width	48.4	51.2	50
SD of length and curl index	52	49.2	48
SD of width and curl index	42.4	40.4	32.8
Avg. of length, width and curl index	**96.4**	**96.8**	**98**
SD of length, width and curl index	49.6	48	48
Avg. and SD of length and width	**87.2**	**89.6**	**94**
Avg. and SD of length and curl index	**86.4**	**86.8**	**92**
Avg. and SD of width and curl index	52	52.8	53.6
Avg. and SD of length, width and curl index	**86.4**	**86.8**	**92.4**

The best result for classification using *k*-NN was the accuracy of 98 %, which was achieved using the following feature sets:

- Set 1:
    - Averages of length
    - $k = 7$
- Set 2:
    - Averages of length and width
    - $k = 7$
- Set 3:
    - Averages of length and curl index
    - $k = 7$
- Set 4:
    - Averages of length, width and curl index
    - $k = 7$

From the results, we can also observe that in the case of multiple features the overall accuracy was over 90 % whenever averages of length was one of features used in the set, barring three sets where the accuracy was under 90 %.

We can also observe that using standard deviations alone as features yields less accurate results, but combined with the data of averages the accuracy increases.

When comparing the results using different values for *k* we can see that the differences in accuracy are small. The biggest change in accuracy was 9.6 percentage points when the value of *k* was increased from three to seven and standard deviations of width and curl index were used as features. In most cases the accuracy increased slightly when the value of *k* was increased, but the increase was mostly around 2 percentage points.

4.3 k-NN results with measurements of individual fibers

In these experiments we used measurements from individual fibers to classify other fibers. In Table 4, we can see the overall accuracies of the classification results using different statistical features and different values of *k*.

Overall, in these experiments the accuracies of the classifications were very low. The overall accuracies were mostly around 30 %. The best result was achieved with the *k* value of 7 and using a feature set consisting of the length, width and curl index of the fibers. The least accurate feature was the ratio between length and width with an accuracy of 23.03 % ($k = 3$). These results would suggest that there is not enough variation between the individual fibers of different species using these features.

From individual features, the length of the fiber gave a slightly better accuracy in the experiments, accuracy of 31.89 % ($k = 5$), compared to the width and curl index of the fiber, indicating that it is the best feature of these three to classify an individual fiber.

Table 4. Classification results using the *k*-NN classifier with individual fibers.

Features used	Accuracy (%)		
	k = 3	k = 5	k = 7
Length	**31.83**	**31.89**	**31.43**
Width	26.63	26.51	28.00
Curl index	29.09	28.06	26.11
Length and width	**32.29**	**33.14**	**33.71**
Length and curl index	**31.83**	**31.09**	**32.23**
Width and curl index	**30.80**	**32.06**	**32.46**
Length, width and curl index	**32.86**	**33.77**	**33.83**
Ratio between length and width	23.03	24.80	24.46
Ratio between length and curl	29.31	28.69	28.40
Ratio between width and curl	26.40	27.14	27.83
Ratios between length and width, and length and curl	29.89	**31.71**	**31.94**
Ratios between length and width, and width and curl	29.89	**31.71**	**31.94**
Ratios between length and curl, and width and curl	27.71	**30.00**	**30.34**
Ratios between length and width, length and curl, and width and curl	27.71	**30.63**	**31.09**

Similarly with the results in Sec 4.2, changing the value of *k* did not have a significant impact on the results. In most cases the accuracy increased slightly when the value of *k* was increased. Overall the change in accuracy was around 2 percentage points and the biggest change was 3.38 percentage points when increasing *k* from three to seven and the following feature set was used:

- Ratio between length and width
- Ratio between length and curl
- Ratio between width and curl

4.4 Naïve Bayes results with statistical data

Here we used the naïve Bayes classifier with two different methods to compute the distribution. The overall accuracies of the results can be seen in Table 5.

As in the previous experiments with the *k*-NN classifier, using only the averages of length we can achieve a very high accuracy of 98.4 %, using the Gaussian distribution.

Also from the single features, averages of width performed adequately with an accuracy of 64.8 % using the Gaussian distribution.

Table 5. Classification results using the naïve Bayes classifier with statistical features.

Features used	Accuracy (%)	
	Gaussian distribution	Kernel distribution
Avg. of length	**98.4**	**94**
Avg. of width	64.8	57.2
Avg. of curl index	35.6	34.8
SD of length	48.8	45.2
SD of width	50	43.2
SD of curl index	24	12
Avg. of length and width	**97.6**	**90**
Avg. of length and curl index	**98**	89.6
Avg. of width and curl index	66.4	57.2
SD of length and width	59.6	59.2
SD of length and curl index	50.4	34.8
SD of width and curl index	49.6	34.8
Avg. of length, width and curl index	**98.4**	87.2
SD of length, width and curl index	57.2	50
Avg. and SD of length and width	**93.6**	88.8
Avg. and SD of length and curl index	**95.2**	74
Avg. and SD of width and curl index	70.4	56.4
Avg. and SD of length, width and curl index	**95.6**	77.6

The least accurate single feature was using the curl index, with accuracies of 35.6 % (Gaussian distribution) using the averages of curl index and a very low 12 % (kernel distribution) using the standard deviations of curl index.

The best accuracy in these experiments was 98.4 % which was achieved with the single feature experiment of averages of length and with the following feature set:

- Averages of length, width and curl index
- Gaussian distribution

In multiple feature experiments, the highest accuracies were achieved when averages of length was one of the used features, similarly as in the experiments with the *k*-NN

classifier. Also, using the standard deviations alone as a feature yields the lowest accuracies.

When we compare the results achieved with Gaussian and kernel distributions, we can observe that using the Gaussian distribution the results were better in all of the experiments. On average the difference in accuracy between the two distributions was approximately 9.3 percentage points, with the largest difference being 21.2 percentage points when averages and standard deviations of length and curl index were used as features. This suggests that the proposed features were approximately normally distributed.

4.5 Naïve Bayes results with measurements of individual fibers

Finally, the classification results using the naïve Bayes classifier with features of individual fibers can be seen in Table 6.

Compared to the results using the *k*-NN classifier in Sec. 4.3, the naïve Bayes performed slightly better with individual fibers. With naïve Bayes the highest accuracy was 44.49 %, which was achieved using the kernel distribution and the feature set of length, width and curl index. The least accurate feature was the ratio between length and curl with the accuracy of 19.83 % (Gaussian distribution).

With length, width or curl index used as a feature, the results show that the length was the most accurate single feature. It achieved an accuracy of 36.23 % using the Gaussian distribution, while the curl index achieved an accuracy of 27.84 % using the kernel distribution. This would suggest that the length of the fiber is the best feature compared to the other two features.

When comparing the two different methods for computing the distribution, the kernel distribution performed slightly better than the Gaussian distribution in all cases except one, where the length of the fiber was used as a feature. On average the difference in accuracy was 7.6 percentage points in experiments where the kernel distribution performed better.

Table 6. Classification results using the naïve Bayes classifier with individual fibers.

Features used	Accuracy (%)	
	Gaussian distribution	Kernel distribution
Length	**36.23**	**35.14**
Width	30.00	31.77
Curl index	22.11	27.84
Length and width	**41.83**	**42.63**
Length and curl index	28.17	**37.94**
Width and curl index	27.31	**36.62**
Length, width and curl index	31.66	**44.49**
Ratio between length and width	29.49	30.29
Ratio between length and curl	19.83	30.85
Ratio between width and curl	26.29	30.09
Ratios between length and width, and length and curl	25.37	**35.09**
Ratios between length and width, and width and curl	25.37	**35.09**
Ratios between length and curl, and width and curl	22.51	**34.42**
Ratios between length and width, length and curl, and width and curl	24.17	**36.14**

5 CONCLUSIONS

The purpose of this thesis was to search and develop different features to characterize fibers segmented from suspension images, and to develop a species identification method that utilizes the selected features.

The features we selected for the identification method were the lengths, widths and curl indexes of the fibers in the suspension images. These features were then measured from the images and from the results we formed data that was used for training and testing our classifier. These features were used to classify the fiber images and individual fibers.

For the classifier we tested two different classification methods, which were the *k*-NN and naïve Bayes classification methods. When classifying fiber images, both of the used classification methods performed well in the experiments. However, when classifying individual fibers neither of the two methods were able to classify the fibers with a high enough degree of accuracy.

With the *k*-NN algorithm, we discovered that when the value of *k* was increased from three to seven, we could slightly increase the accuracy of the classification in both classifying the fiber images and individual fibers.

Using the naïve Bayes method we discovered that for classifying the fiber images the Gaussian distribution was more suitable compared to the kernel distribution. When classifying individual fibers, the kernel distribution was more suitable compared to the Gaussian distribution

When classifying fiber images, we can observe that the naïve Bayes classifier with Gaussian distribution achieved better results in 60 % of the experiments compared to *k*-NN with a *k* value of seven, and 68 % and 62 % with *k* values of three and five respectively. When we compare the features used in the experiments we can see that by using just the lengths of the fibers, we can achieve a classification accuracy of over 90 % with both classification methods. This indicates that with fiber images from the species used in these experiments, we could only measure the fiber lengths and be able to classify the images quite accurately.

With individual fiber classification the results were less impressive. In the best case we were able to identify the fibers with an accuracy of 44.49 % when using the naïve Bayes classifier with kernel distribution. When comparing the two classification methods, the naïve Bayes classifier with kernel distribution achieved the best results. It performed better in all of the experiments when compared to *k*-NN with a *k* value of seven, and 93 % of the experiments with *k* values of three and five.

As an overall conclusion, we were able to develop a classification method that is able to identify the species of fiber images with a high degree of accuracy but not able to identify individual fibers accurately enough. Both classification methods performed well in the experiments with the naïve Bayes classifier being the most accurate.

REFERENCES

Altman, N.S., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am. Stat. 46(3), 175–185. doi:10.1080/00031305.1992.10475879

Bayes' Theorem (Stanford Encyclopedia of Philosophy) [WWW Document], n.d. URL http://plato.stanford.edu/entries/bayes-theorem/ (accessed 8.17.14).

Borch, J. (Ed.), 2002. Handbook of physical testing of paper, 2nd ed., rev. and expanded. ed. Marcel Dekker, New York.

Bramer, M., Coenen, F., Tuson, A. (Eds.), 2007. Research and Development in Intelligent Systems XXIII. Springer London, London.

Canny, J., 1986. A Computational Approach to Edge Detection. IEEE Trans. Pattern Anal. Mach. Intell. 8(6), 679–698. doi:10.1109/TPAMI.1986.4767851

CEMIS - CEMIS-OULU [WWW Document], n.d. URL http://www.cemis.fi/in-english/expertise/cemis-oulu-2 (accessed 8.17.14).

Chandra, M., 1998. Use of Nonwood Plant Fibers for Pulp and Paper Industry in Asia: Potential in China. Masters Thesis, Va. Polytech. Inst. State Univ.

Conservatree [WWW Document], n.d. URL http://www.conservatree.org/about/About.shtml (accessed 8.17.14).

Domingos, P., Pazzani, M., 1997. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Mach. Learn. 29(2-3), 103–130. doi:10.1023/A:1007413511361

Guo, S., Zhan, H., Zhang, C., Fu, S., Heijnesson-Hultén, A., Basta, J., Greschik, T., 2009. Pulp and fiber characterization of wheat straw and eucalyptus pulps–A comparison. BioResources 4(3), 1006–1016.

Kurakina, T., 2012. Characterization of fiber and vessel elements in pulp suspension images. Masters Thesis Lappeenranta, Univ. of Tech.

Lindeberg, T., 1998. Edge Detection and Ridge Detection with Automatic Scale Selection. Int. J. Comput. Vis. 30(2), 117–156. doi:10.1023/A:1008097225773

Lönnberg, B., 2009. Mechanical pulping. Paperi ja Puu Oy, Helsinki, Finland.

Medioni, G., Kang, S.B., 2004. Emerging Topics in Computer Vision. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Medioni, G., Lee, M.-S., Tang, C.-K., 2000. A computational framework for segmentation and grouping. Elsevier, Amsterdam ; New York.

Page, D.H., Seth, R.S., Jordan, B.D., Barbe, M.C., 1985. Curl, crimps, kinks and microcompressions in pulp fibres their origin, measurement and significance, in: Papermaking Raw Materials : Their Interactions with the Production Process and Their Effect on Paper Properties : Transactions of the Eighth Fundamental Research Symposium Held at Oxford: September 1985. Bury, Lancashire : FRC, Pulp and Paper Fundamental Research Society ; [United Kingdom] : Immediate Proceedings, 2003, pp. 183–227.

PulpVision - Image Processing and Analysis Methods for Pulp Process Measurements [WWW Document], n.d. URL http://www2.it.lut.fi/project/pulpvision/ (accessed 8.16.14).

Strokina, N., Kurakina, T., Eerola, T., Lensu, L., Kälviäinen, H., 2013. Detection of Curvilinear Structures by Tensor Voting Applied to Fiber Characterization in the proceedings of the 18th Scandinavian Conference of Image Analysis (SCIA), pp. 22-33, Espoo, Finland

Swann, C.E., 2006. The Future of Fiber. Pap. 360 1, 10–12.

Wai-Shun Tong, Chi-Keung Tang, Mordohai, P., Medioni, G., 2004. First order augmentation to tensor voting for boundary inference and multiscale analysis in 3d. IEEE Trans. Pattern Anal. Mach. Intell. 26(5), 594–611. doi:10.1109/TPAMI.2004.1273934