ilmedia

th

TECHNISCHE UNIVERSITÄT
ILMENAU

Alexander Ihlow and Udo Seiffert

Automating microscope colour image analysis using the Expectation Maximisation algorithm

Automating Microscope Colour Image Analysis Using the Expectation Maximisation Algorithm

Alexander Ihlow and Udo Seiffert

Leibniz Institute of Plant Genetics and
Crop Plant Research (IPK) Gatersleben
Corrensstr. 3, 06466 Gatersleben, Germany
Pattern Recognition Group
{ihlow, seiffert}@ipk-gatersleben.de
http://bic-gh.ipk-gatersleben.de/wgrp/mue

Abstract. Dyed barley cells in microscope colour images of biological experiments are analysed for the occurrence of haustoria of the powdery mildew fungus by a fully automated screening system. The region of interest in the images is found by applying Canny's edge detector to the hue channel of the HSV colour space. Potential haustoria regions are extracted in RGB colour space by an adaptive Gaussian mixture classifier based on the Expectation Maximisation (EM) algorithm. Since the classes *cell* and *haustorium* are at very close quarters, their correct separation is a crucial part and needs a constraining mechanism which ties the EM algorithm to its initialisation data to prevent a too large deviation from it.

1 Introduction

Automating the screening and the analysis of biological experiments is a challenging research area in the field of bioinformatics and engineering. This paper is related to a project where resistance mechanisms of crop plants against the powdery mildew fungus are studied from the genetical point of view. In the experiments, young barley leaves are bombarded with DNA-coated tungsten particles to "switch on or off" desired genes in cells. For analysis purposes, an additional reporter gene[1] is expressed in cells that were hit by a particle. This dyes the affected genetically transformed cells greenish blue and allows their identification by bright field microscopy [8]. The task is to evaluate the susceptibility of the genetically transformed cells to the powdery mildew fungus under the impact of different test genes. A successful penetration of the fungus into the cell is indicated by the development of a haustorium – a dark object with "fingers" that is located between the cell wall and the cell membrane and feeds the fungus by leaching the cell. These objects have to be counted in an automatic analysis procedure.

Since there are many genes to be considered for a potential resistance of the plant against pathogens, a big number of experiments has to be performed to

---

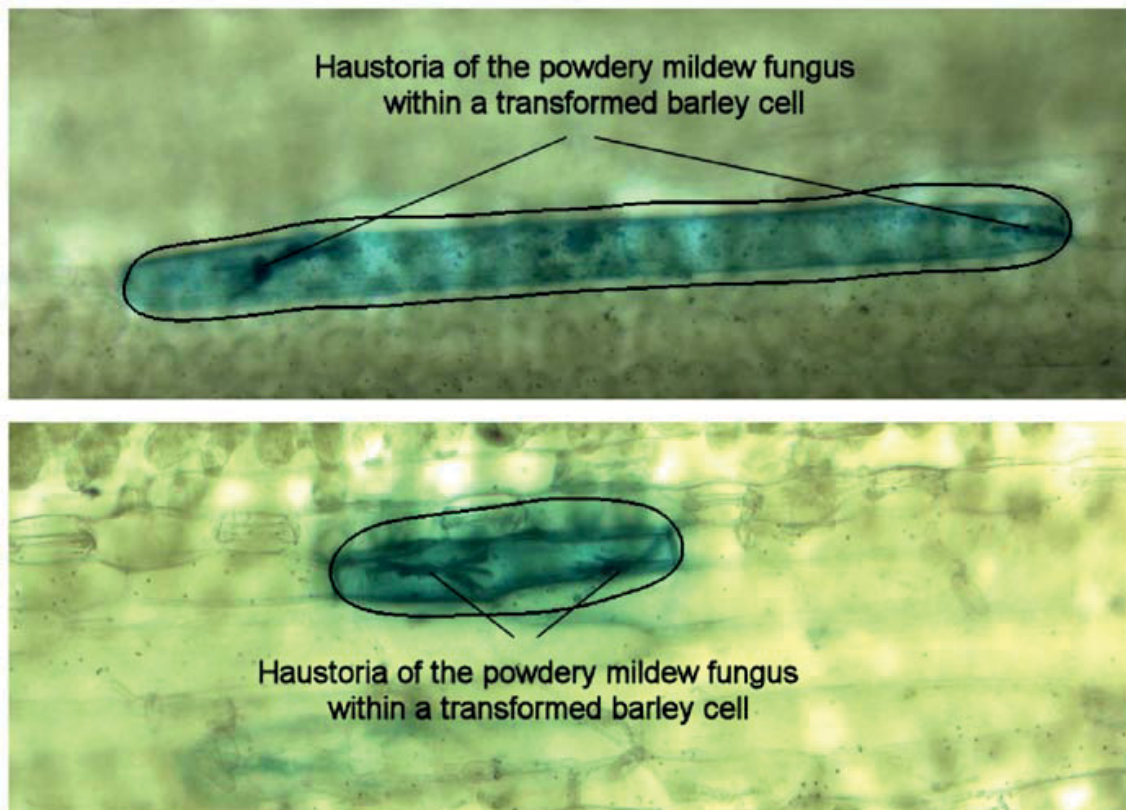[1] β-glucuronidase (GUS) reporter gene

Fig. 1. Cutouts of microscope images of barley cells. The dyed cells are genetically transformed, both cells contain two haustoria of the powdery mildew fungus. At coarse scales, Canny's edge detector marks these cells by a closed boundary.
Color version available via `http://bic-gh.ipk-gatersleben.de/wgrp/mue/prj03.php`

attain a sufficient statistical confidence. Therefore, an automated image acquisition system and an automatic analysis procedure is needed. Manual screening is a tedious, subjective and time-consuming task that cannot be handled by laboratory assistants due to that huge amount of data. For an automatic image acquisition, the microscope slides are mounted on an x-y table which scans a number of preparations fully automatically under the control of a computer, e.g., overnight. Now, finding genetically transformed cells and therein assessing the development status of the haustoria without human interaction is the task and the challenge of the analysis procedure.

This paper describes a method to automatically identify suspicious objects, i.e., parts of genetically transformed cells that may be a haustorium. It is organised as follows: Section 2 introduces the properties of the image material and explains how the regions of interest, i.e., genetically transformed cells, are found in the images. Afterwards, Section 3 describes the identification of potential haustoria via the Expectation Maximisation (EM) algorithm, before Section 4 concludes the paper.

2 Preprocessing of the Image Material

Figure 1 shows two typical cutouts of microscope images, both containing one dyed genetically transformed cell with two haustoria of the powdery mildew fungus inside. By default, the microscope camera produces images of 2600×2060 pixel in 24-bit colour.

In [5] we have shown that the dyed genetically transformed cells can be reliably detected by applying Canny's edge detector [2] to the hue channel of the HSV colour space, rather than performing multi-dimensional edge detection in the RGB colour space or using histogram-based methods. At a coarse scale Canny's algorithm marks the dyed cells by a closed boundary. The bounding box of these closed contours will be the input of the further haustorium detection procedure. Unfortunately, the haustoria stand out scarcely from the dyed cell, and there is no such straightforward colour space transformation to separate them as good as the dyed cells from the remaining cell tissue. Therefore, we stay in the RGB colour space, which contains the entire image information, and show what haustorium detection results can be achieved by pixel classification methods.

3 Cell Image Analysis by Clustering in Colour Space

3.1 Naive Bayes Classification

Suppose a naive Bayes classifier at first. A number of N d-dimensional data vectors $\mathbf{x}_n \in \mathbb{R}^{d \times 1}$ from the entire data set $\mathbf{X} \in \mathbb{R}^{d \times N}$ has to be classified into K classes. If the prior (a priori) probabilities $P(k)$ and the probability density functions $p(\mathbf{x}|k)$ of the $k = 1 \ldots K$ classes are known, then the posterior (a posteriori) probability $P(k|\mathbf{x}_n)$ of a sample vector $\mathbf{x}_n$ to belong to class k can be calculated by Bayes' rule [1] (maximum likelihood decision) according to

$$P(k|\mathbf{x}_n) = \frac{P(k)\ p(\mathbf{x}_n|k)}{\sum\limits_{j=1}^{K} P(j)\ p(\mathbf{x}_n|j)}\ . \tag{1}$$

Inspecting our data in the RGB colour space, we can decompose the mixture distribution of colours into three stretched ellipsoids, representing the three dominant image matters, namely *background*, *cell*, and *haustorium*. Such ellipsoidal distribution can be well modelled by the multivariate Gaussian distribution, which is described by the mean vector $\boldsymbol{\mu}$, specifying the center point of the ellipsoid, and the covariance matrix $\boldsymbol{\Sigma}$, which is responsible for the shape and the orientation of the ellipsoid.

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{\det \boldsymbol{\Sigma}_k (2\pi)^d}}\ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)} \tag{2}$$

See Figure 2 for the segmentation results of this naive Bayes classification where the parameters of the classes were taken from typical samples. In the upper
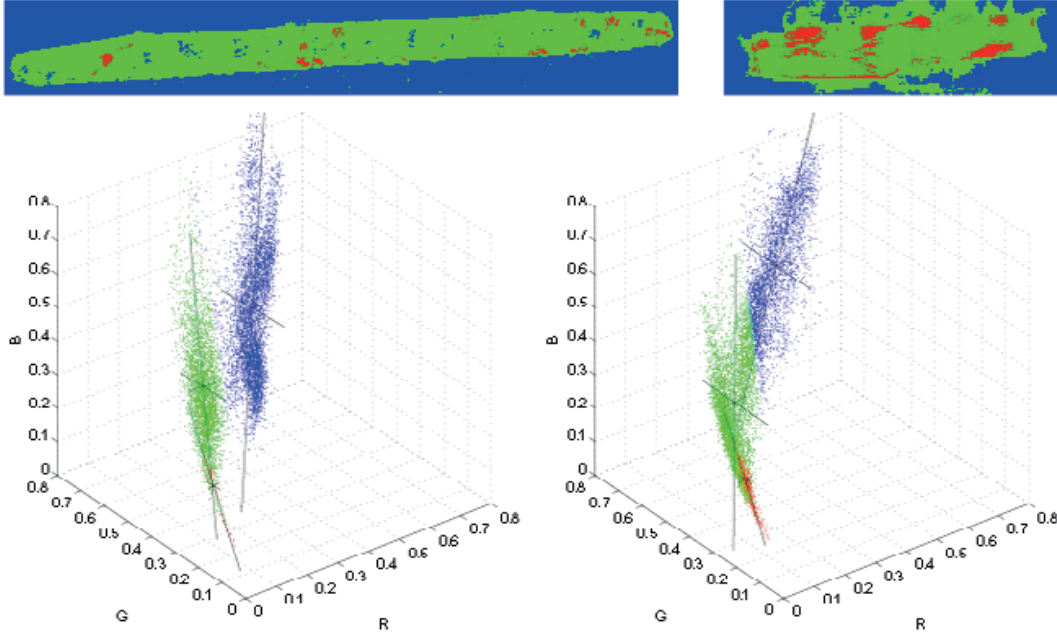
Fig. 2. Segmentation by a naive Bayes pixel-classification in RGB colour space modelling the classes by multivariate Gaussians.

images, the pixel-labels are depicted in a soft-output manner, i.e., the vector of the posterior probabilities $[P(k = 3|\mathbf{x}), P(k = 2|\mathbf{x}), P(k = 1|\mathbf{x})]^T$ is assigned to the RGB value of each pixel, making the saturation of the colour follow the reliability of the estimate. The lower figures show both the clusters in RGB colour space as well as the principal components (eigenvectors) of each cluster. As can be seen, simply assigning parameters from typical images for the three classes and performing a naive Bayes classification does not provide satisfactory results because the parameter set will never match the actual scenario sufficiently due to some inevitable variations in colour and illumination in the image data. Therefore, some "self adaptation" of the classification algorithm to the actual data is needed to improve the classification results.

3.2 EM Classification Using the Complete Data Set

The Expectation Maximisation (EM) algorithm [4,7] is known to be a powerful clustering technique for mixture distributions where the parameters of the underlying probability density functions are adapted in an iterative way, trying to yield the best recovery of the mixture components. Its clustering performance depends on two major conditions: the precision the actual data is represented by the data model, as well as the initialisation parameters, because it can converge to local extrema instead of finding the global optimum. Such clustering methods are used for many different applications in image processing, e.g., skin detection [6]. In [3] an advanced image querying system is described which applies the EM algorithm to an eight-dimensional space of colour, texture, and position features, where the number of mixture components is chosen following the Minimum Description Length (MDL) principle. Fortunately, we know the number
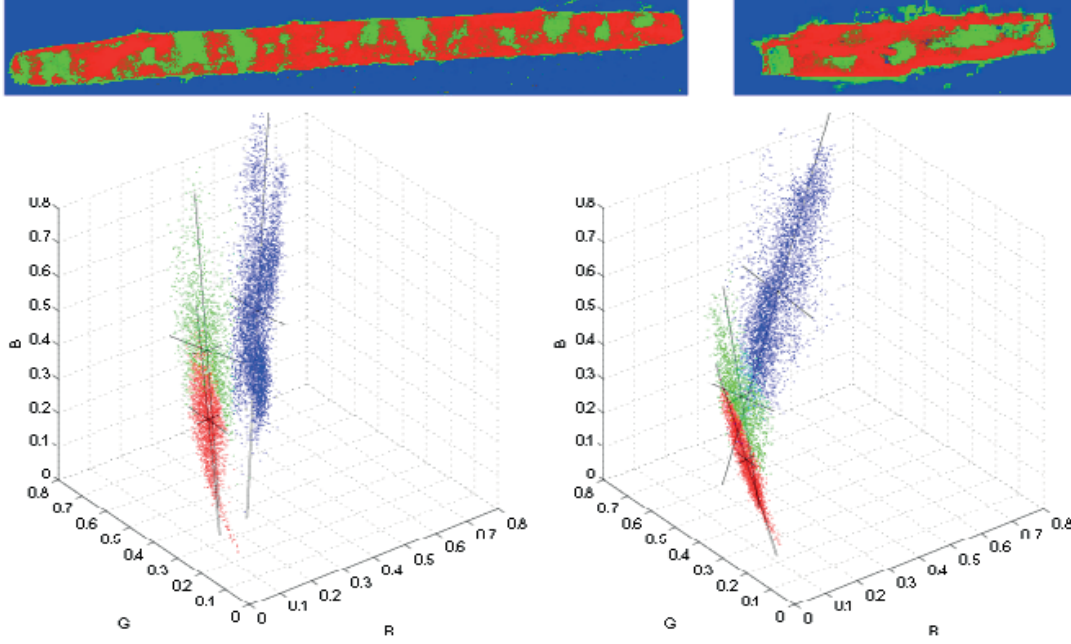
Fig. 3. Segmentation results of the EM algorithm when iterating on the entire data set of RGB colour vectors.

of mixture components in feature space very well due to the speciality of our image material. Furthermore, colour appears as the dominant feature, therefore we can ignore texture and position features and use the RGB colour information as the only feature.

We initialise the a priori probability of the classes with $P(k) = 1/K = 1/3$ (since we do not know $P(k)$ in advance) and perform a data-driven initialisation of the mean vectors and covariance matrices of the classes from exemplary, hand-segmented image parts, as already done for the naive Bayes classification. Then, the iteration of the EM algorithm is run in the following manner:

The probability (at iteration step t) of each data vector $\mathbf{x}_n$ to belong to class k is calculated (expectation step) by

$$P^t(k|\mathbf{x}_n) = \frac{P^t(k)\ p(\mathbf{x}_n|\boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t)}{\sum\limits_{j=1}^{K} P^t(j)\ p(\mathbf{x}_n|\boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t)}\ . \tag{3}$$

A new parameter set for the iteration step $t+1$ containing the prior probabilities, mean vectors and covariance matrices for each class is calculated according to (maximisation step)

$$P^{t+1}(k) = \frac{1}{N} \sum_{n=1}^{N} P^t(k|\mathbf{x}_n) \tag{4}$$

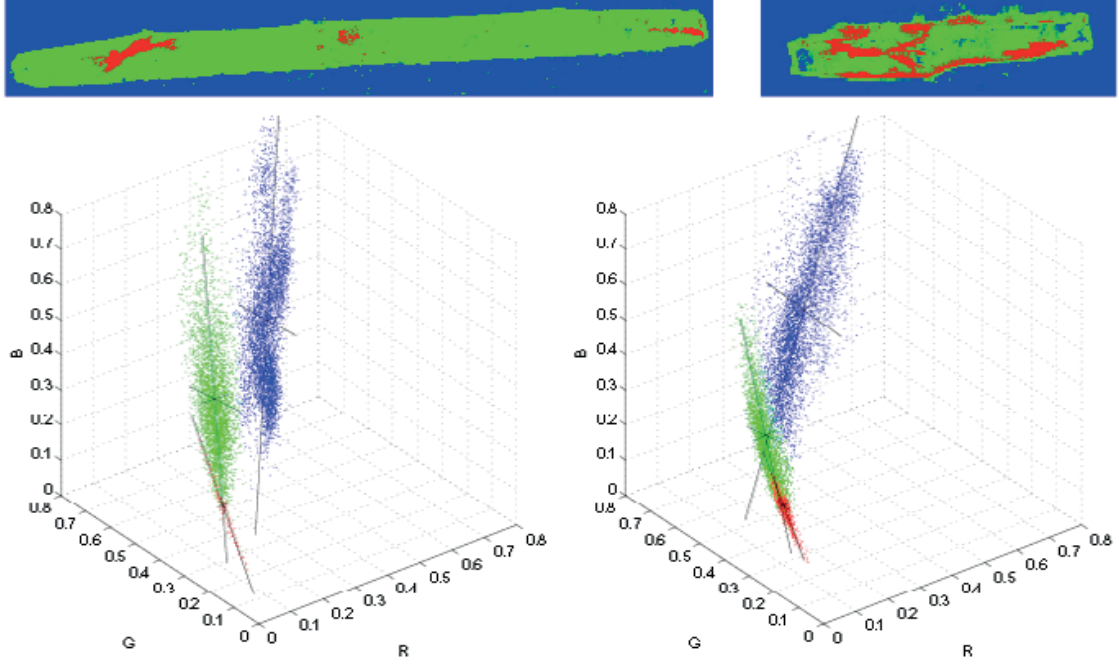$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{N P^{t+1}(k)} \sum_{n=1}^{N} P^t(k|\mathbf{x}_n)\ \mathbf{x}_n \tag{5}$$

Fig. 4. Segmentation results of the EM algorithm iterating data vectors that were estimated by a reliability of at least $R_{min} = 0.65$.

$$\boldsymbol{\Sigma}_k^{t+1} = \frac{1}{NP^{t+1}(k)} \sum_{n=1}^{N} P^t(k|\mathbf{x}_n) \, (\mathbf{x}_n - \boldsymbol{\mu}_k^{t+1}) \, (\mathbf{x}_n - \boldsymbol{\mu}_k^{t+1})^T. \qquad (6)$$
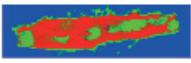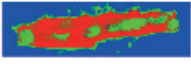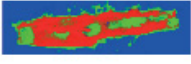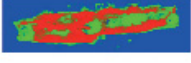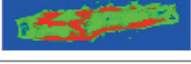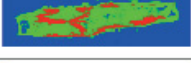
The algorithm is terminated when the labelling in the segmented image does not change anymore.

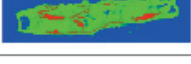As can be seen in Figure 3, the clustering separates the *background* and *cell* class very well but it suffers from an overestimation of the *haustorium* class. This solution is optimal from the EM point of view, but it is not our desired result for an appropriate segmentation. In spite of different initial parameters, the EM algorithm tends towards bad results of the same manner. Incrementing the model order, i.e., providing more classes generally does not yield more solid results, especially for the right hand image.

3.3 Constraining the EM by Reliability Information

A straightforward solution to achieve appropriate segmentation results is found in constraining the algorithm to the initial parameter set, which is known quite well in our particular case. Using the complete data set (all image pixels) makes a large number of *cell* labels to turn over into *haustorium* labels during the iterations. Iterating on reliably estimated data vectors only (instead on the entire data set) prevents the algorithm from deviating too much from its initial parameters. The classification reliability of each sample is given by $R = \max_k\{P^t(k|\mathbf{x}_n)\} \in [1/K \dots 1]$ and is inherently calculated in each iteration.

Table 1. Tapering the subset of data samples which the EM uses for iteration by a stepwise variation of the reliability parameter R_{min}.

In the following, we restrict the data set, which the EM operates on, to data samples that were classified with a reliability of at least R_{min}. Note that the lower bound of this parameter depends on the number of classes and that an appropriate parameter value has to be found empirically by a visual inspection of the segmentation results. See Table 1 for a test series of our particular segmentation problem. It can be observed that there is a significant changeover between $R_{min} = 0.50 \ldots 0.60$. Choosing R_{min} larger than 0.75, we observed convergence problems of the algorithm for the right hand image, where the algorithm oscillated harmonically between two states instead of terminating. This can be explained by the recurrent changing of the considered data set parts during the iterations and needs further attention.

Figure 4 shows the detailed segmentation results for $R_{min} = 0.65$. Despite some misclassified objects in the *haustorium* class it shows the haustoria quite good — with this method we are able to automatically identify suspicious objects, i.e., potential haustoria. Now, further analysis on the detected objects is needed to distinguish haustoria from discolourations or other parts inside the cell that have a similar colour, e.g., the cell nucleus. As a next step, therefore these image parts have to be further evaluated, taking form parameters of the detected objects into account, e.g., by detecting the "fingers" of the haustoria. This will be examined in the near future and is out of the scope of this paper.

This paper is accompanied by a continuative web site of the presented results. Visit `http://bic-gh.ipk-gatersleben.de/wgrp/mue/prj03.php` for a more detailed compilation of exemplary cell images and their clustering results.

4   Conclusions

The Expectation Maximisation (EM) algorithm is applied in the RGB colour space to perform a segmentation of microscope colour images for the identification of small objects which stand out scarcely from the region of interest. To provide satisfactory results, it is shown that this special problem needs a constraint mechanism which ties the EM algorithm to its initialisation parameters and forbids a too large deviation from it. This constraint mechanism is realised by dynamically restricting the data set the algorithm operates on to a reliably estimated part only. The mechanism is parametrised by a reliability threshold parameter which has to be determined empirically. This technique prevents a defection of the desired segmentation and provides good retrieval results of suspicious objects via an automatic analysis procedure.

References

1. Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1753. (available via http://www.stat.ucla.edu/history/essay.pdf).
2. John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 8(6):679–698, November 1986.
3. Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems*. Springer, 1999.
4. Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
5. Alexander Ihlow and Udo Seiffert. Microscope color image segmentation for resistance analysis of barley cells against powdery mildew. In *9. Workshop "Farbbildverarbeitung"*, ZBS Zentrum für Bild- und Signalverarbeitung e.V. Ilmenau, Report Nr. 3/2003, pages 59–66, Ostfildern-Nellingen, Germany, October 2003.
6. Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. In *IEEE Conference on Computer Vision and Pattern Recognition '99*, pages 274–280, June 1999.
7. Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, 26:195–239, 1984.
8. Patrick Schweizer, Jana Pokorny, Olaf Abderhalden, and Robert Dudler. A transient assay system for the functional assessment of defense-related genes in wheat. *Molecular Plant-Microbe Interactions*, 12(8):647–654, 1999.