

Data mining for image analysis: in-line particle monitoring in polymer extrusion

K. Torabi¹, L. D. Ing¹, S. Sayad² & S. T. Balke¹

¹*Department of Chemical Engineering and Applied Chemistry
University of Toronto, Toronto, Ontario, Canada*

²*iSmartsoft Inc., Toronto, Ontario, Canada*

Abstract

The objective of this study is to demonstrate how data mining techniques are being combined with image analysis methods to develop an automated particle monitoring system. A new optical monitor provides images of contaminant particles in molten plastic flowing through processing equipment. The product, plastic film, is extremely intolerant of particle contaminants: even very low concentrations of particles may cause packaging film defects. Thus, this initial work was aimed at developing the ability to automatically distinguish between images containing detectable contaminant particles from those that do not. A total of approximately six thousand images was obtained from the in-line monitoring of three different lots of low-density polyethylene during processing in a 3.8 cm. diameter single screw extruder. Several different data mining techniques including One-R, Decision Tree, Bayesian and k-Nearest Neighbour classifiers were applied to image datasets and the overall success rates were as high as 99.9% for some models. The Decision Tree method (C4.5) was the most accurate technique for the image classification, and the Bayesian classifier showed the minimum false negative prediction. Selection of threshold grey level was critical. Two new automated image thresholding techniques were developed and assessed.

1 Introduction

A recently developed in-line optical system, called the “Scanning Particle Monitor” or “SPM”, supplies images of particles as small as 10 microns in a flowing polymer melt at any point across the flow stream during extrusion

processing [1]. Extruders are essentially long screw pumps with solid pellets of plastic being melted and pumped through a hole or a slit to form a variety of products. Plastic film is a particularly important product obtained from extruders by stretching the molten plastic exiting the equipment. Particles in the plastic melt cannot be stretched and result in film defects. Even a few particles can ruin large quantities of product. These particles can originate from many sources: high shear forces or high temperatures causing the plastic to react to form "microgels", un-melted plastic and "dirty" feed materials for example.

A fundamental requirement is that when a certain level of particle contamination is reached, some sort of control action is immediately taken to remedy the situation. Currently, many industries use "off-line" manual inspection of the film in analytical laboratories. Large losses due to off-specification products then result. More recently, expensive film scanners have been developed. These devices automatically examine the film continuously as it is produced. They are effective but the particles causing the defects are encased in solid polymer so that little information on the type of particles causing the defect, or the origin of the particles is available. Monitoring particles in the melt using imaging greatly increases the probability of identifying both the type and origin of the particles.

However, a complication is that not all particles in the melt will cause film defects. In particular, sometimes desirable particles are intentionally introduced. Typically these "additives" are used to facilitate processing of the polymer or to improve the properties of the final plastic product. The additives are so fine that they do not cause defects. This initial work has the limited objective of automating selection of images, which contain any particles above 10 microns. Typically this represents about 5 to 12% of the images acquired. Accomplishing this objective required three primary steps: improvement of image quality by selecting specific image processing tools; automating selection of grey level threshold setting and application of data mining methods. These are discussed in turn in the next section.

2 Theory

2.1 Image quality improvement

Today digital images can be manipulated so readily that "seeing" is no longer synonymous with "believing". An extremely wide variety of tools are available for changing images. The SPM software includes more than twenty different tools ranging from contrast enhancement to particle edge correction and each one of them has up to four settings. However, at the same time, the result of data mining can be very dependent upon the quality of an image. Thus, the objective of image pre-processing is to improve the characteristics of the image, which can benefit the data mining method without sacrificing image fidelity. Generally, with so many tools and settings from which to choose, selection of appropriate image processing is done manually and empirically, using typical images. The possibility of applying data mining methods to achieve a more automated and

optimum selection is being investigated and will be the subject of a future publication.

2.2 Image thresholding

A critical requirement in this work is being able to automatically specify the degree of blackness (i.e. the grey threshold level) that is used by the software to distinguish a particle from its background [2]. This value can range from 0 (black) to 255 (white). As the value is increased from 0 to 255, progressively less dark pixels are considered as particles. A standard way of examining threshold selection is to plot the data as a histogram [3]. If a large number of particles are present, then a histogram such as that in Figure 1a is obtained. There it can be seen that specifying the grey level in the valley between the peak from the particles and the multiple peaks due to the noise, the software should readily be able to distinguish particles from noise in the images. Figure 1b shows the situation in this work: only a small number of particles are present. The single unimodal peak evident there is due to background noise (actually fine particles). No peak is visible for the single “large” particle present in the image.

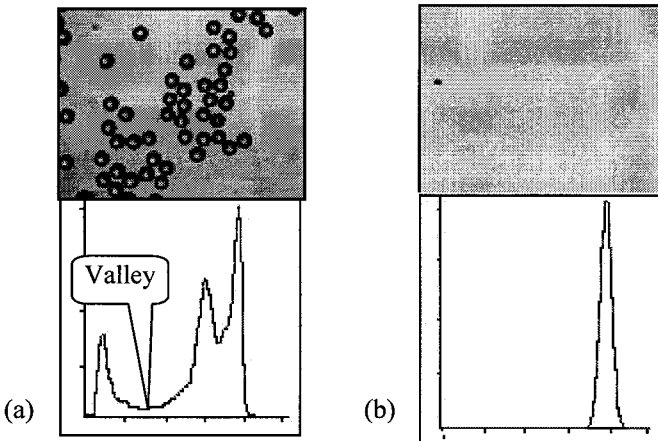


Figure 1: Comparison of histograms of counts versus grey level for threshold selection. (a) Multiple large particles present. (b) Single small particle present.

If the image is cropped (Figure 2), so as to cause the number of counts from the particle to be on a par with those from the background noise, then noise in the histogram becomes prohibitively large. The software used in this work contained yet another method of threshold setting: one that we term “conventional automated threshold selection”. It specified the threshold as the median grey level of the histogram of the whole image. Furthermore, here two new methods of selecting the threshold value were developed: “max min” and “moving average” thresholding. These are discussed in turn in the following sections.

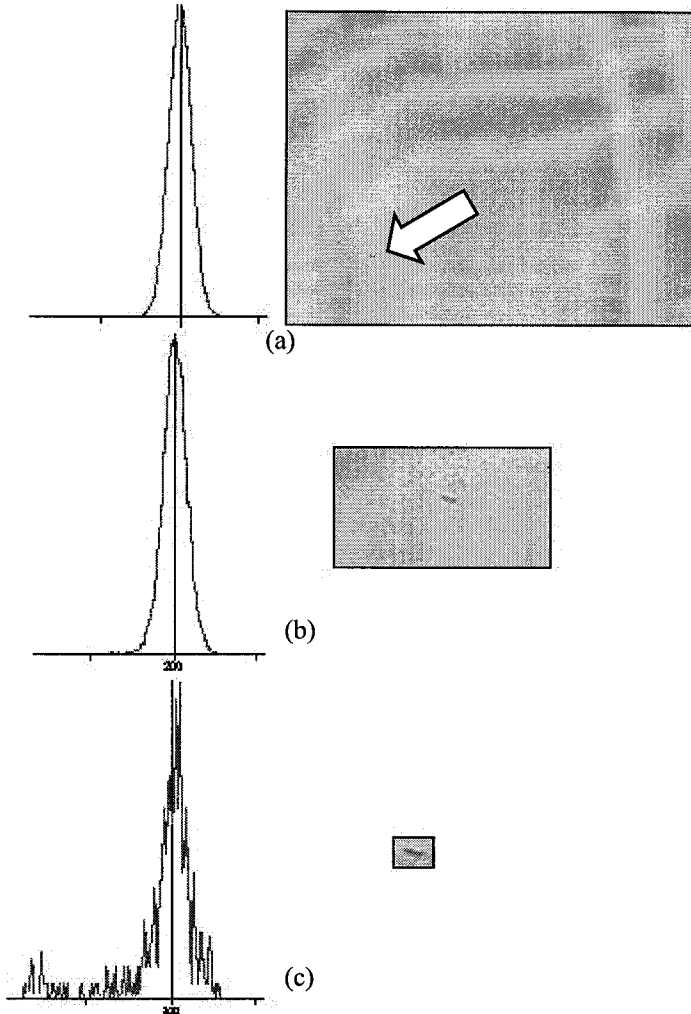


Figure 2: The effect of image cropping on threshold selection. (a) (b) & (c) show progressively more cropped image and the resulting counts versus grey level histograms

2.2.1 The Max Min method of threshold selection

In the “Max Min” method the grey level is changed from black (0) to white (255) and both the size of the smallest particle and the total number of identified particles in an image are recorded. The point at which the size of the smallest identified particle reaches a maximum size is considered as the critical threshold value of grey level. The reasoning underlying this is that as grey level is increased, more pixels are considered to represent the identified particle. At some point a new particle appears and becomes the particle of minimum size. This one too will grow. As the noise level is approached a new particle will be

unable to grow significantly because another new smallest particle is immediately identified. No a priori knowledge of the size of a contaminant particle is necessary.

2.2.2 The Moving Average Method of threshold Selection

The “moving average method” is based upon the idea that the rate of change in the total number of identified particles significantly increases at the desired threshold value. The method is similar to that of Weszka [4] who recommended setting the threshold value where the slope of a plot of the total particle area versus threshold setting drops to minimum.

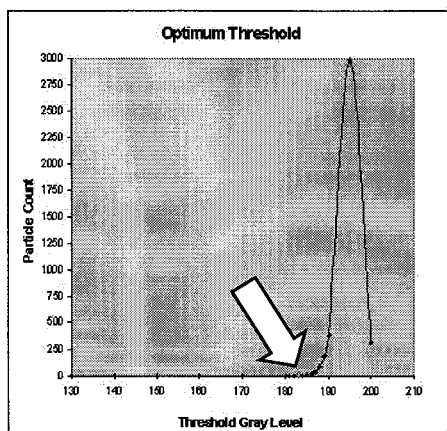


Figure 3: The point where a significant change in the number of total identified particle occurs can be considered as threshold value for the image.

In this study we preferred to work with particle count rather than area because the number of particles significantly changes when we are considering background noise as particles, whereas the total area of particles does not change as rapidly because the noise particles are extremely small. As shown in Figure 3, for each image, the point where a significant shift in the slope of the graph occurs, the threshold value needs to be chosen. The technique was used to identify this point is “moving average. Once the new count exceeds four times the standard deviation of previous results, that point is considered to represent a significant deviation in slope.

2.3 Data Mining methods

Although thresholding can be used to determine whether there is any particle identified in an image, it does not provide reliable results. A major reason is that even if an image does not have any contaminant particle, during the threshold selection at least one particle will be identified although it may only be background noise. Specifying a minimum particle size is possible but results can

then easily be too dependent upon that value. Using image processing software applications for particle size measurement, up to 300 attributes can be measured in each image. These include minimum, maximum, mean, and standard deviation of: count, area, pixel density, margination, integrated optical density, ratio of counted particles area to the total area, fractal dimension, and roundness of object particles. Among all these attributes, the most informative ones need to be identified. Then, the selected attributes can be used in an in-line image classification system in order to distinguish those images, which contain at least one contaminant particle from those that do not.

3 Experimental

Three sets of experiments were carried out at the Department of Chemical Engineering and Applied Chemistry, University of Toronto. A fixed speed single screw extruder (3.8 cm diameter) equipped with sapphire windows, a fibre optic light source and specially designed optical system (SPM) viewed the polymer melt in the extruder near the exit of the equipment. Low density Polyethylene with three different additive mixtures and three different particle contamination levels was processed. Data acquisition speed was ~2 images per second during each extruder run approximately 2000 images were collected. The SPM software was used for data collection and pre-processing. Standard glass micro-spheres were used for image calibration. In this study Visual Basic software was developed that performed automated image pre-processing, thresholding and particle measurements. All images were visually classified into three classes of "With", "Without" contaminant particles, and "Uncertain". Four human experts were chosen to classify the same images in order to improve the accuracy of the classification of some images. Automated image analysis was repeated three times with the different thresholding techniques: conventional auto thresholding, max min and moving average. The datasets were created with fifty-four attributes for each image. Each attribute was the mean value of a specific measurement for all particles in a particular image.

WEKA [5], an educational data mining software application, developed at the University of Waikato, New Zealand, was used to examine the effectiveness of Decision Tree, Bayesian Classifier, K-Nearest Neighbours, and One-R techniques for identifying which images contained contaminant particles.

4 Results and discussion

4.1 Image data pre-processing

It was rapidly determined that none of the Data Mining methods could succeed with the unprocessed raw images. Pre-processing steps were derived in the usual way (trial and error) and were as simple as possible. The pre-processing included desaturation (to remove color from the image and convert it to grey scale), Hi Gauss filtration and brightness flattening. The pre-processing was conducted automatically using the SPM software.

4.2 Conventional auto thresholding

The first dataset (Dataset 1) consisted of 1726 records of image data with no particle in them (WO indicates “without particles”), and 221 records of images with contaminant particles (WP indicates “with particles”). These images were visually classified into WO and WP classes, and the conventional auto thresholding technique was applied before particle size measurement. Each record had 54 attributes. One-R, Decision tree (C4.5), Three Nearest Neighbors (IB3), and Bayesian classifier were applied to the dataset in order to find the best image classification technique. The results are shown in Table 1. No significant difference between any of the classification techniques and a random selection was evident. The confusion matrix of ten fold cross-validation showed that most of WP images were incorrectly classified as WO.

Table 1: Data mining technique results using Conventional Auto Thresholding

Dataset	Method	One-R	C4.5	IB3	Bayes
1	Median	88%	88%	85%	82%

Table 2: Confusion Matrix for C4.5 and Conventional Auto Thresholding

WP	WO	<= C4.5 Classified as
2	219	WP
20	1706	WO

4.3 Max Min thresholding

After examining more than 50 images, we could demonstrate statistically that the max min thresholding was the best auto thresholding technique and provided results equivalent to individual visual selection of threshold by human experts. We then applied this method to four new datasets:

Dataset 2: 1253 clearly defined WO and 91 WP images. (Images with blurs and shadows were excluded from the dataset).

Dataset 3: 1909 WO and 91 WP images (no images excluded);

Dataset 4: Dataset 3 with only 8 instead of 54 attributes.

Dataset 5: Dataset 3 with a third class defined as “blurry images”.

Decision Tree (C4.5), One-R, 3-Nearest Neighbours (IB3), and Bayesian Classification were applied. As shown in Table 3, the Decision Tree and the One-R models provided the best accuracy rates. Excluding the blurry images, accuracy of class prediction using the first dataset was more than 99.9%. A ten-fold cross validation was used and the confusion matrices are shown in Tables 4 and 5. For the second dataset, only 1 or 2 images out of 1,334 were incorrectly classified by C4.5 and One-R, respectively.

Table 3: Data mining technique results using Max Min Size thresholding

Dataset	Method	One-R	C4.5	IB3	Bayes
2	Max.Min.	99.9%	99.8%	99.8%	95.8%
3	Max.Min.	98.5%	97.8%	97.8%	93.3%
4	Max.Min.	98.5%	98%	98.4%	97%
5	Max.Min.	87%	87%	84%	79%

Table 4: Confusion Matrix for Dataset 2 using One-R & Max Min Thresholding

WP	WO	<=OneR Classified as
90	1	WP
0	1257	WO

Table 5: Confusion Matrix for Dataset 2 using C4.5 & Max Min Thresholding

WP	WO	<= C4.5 Classified as
89	2	WP
0	1257	WO

As shown in Tables 6 and 7, among all different techniques applied to the Dataset 3, the Bayesian had the lowest class-value prediction accuracy, but the best false negative response. The false negative response is very important from a manufacturing point of view, and has to be minimized in order to have an acceptable in-line monitoring system for the polymer film manufacturing. Using decision tree results for Dataset 3, we selected the top 8 attributes, produced Dataset 4, and applied the same data mining techniques. As a result the prediction for Bayesian Classifier was improved from 93.3% to 97%. C4.5 and One-R were the most successful models with 87% accuracy in class value prediction for Dataset 5. However, the confusion matrix, Table 8, for C4.5, shows that it was not able to clearly distinguish the blurry images as a separate class. The most important attributes in these four datasets were Minimum Density, Mean Density and Density Standard Deviation. Particle count and particle size were not important attributes.

Table 6: Confusion Matrix for Dataset 3 using C4.5 & Max Min Thresholding

WP	WO	<= C4.5 Classified as
69	22	WP
23	1886	WO

Table 7: Confusion Matrix for Dataset 3 using Bayesian & Max Min Thresholding

WP	WO	<=Bayes Classified
89	2	WP
132	1777	WO

Table 8: Confusion Matrix for Dataset 5 using C.5 & Max Min Thresholding

WP	WO	U	<=C4.5 Classified as
69	1	21	WP
0	1175	82	WO
20	114	488	U

4.4 Moving average thresholding

Moving Average Thresholding was applied to Datasets 2 and 3. Results (Table 9) were statistically identical to those obtained using the max min thresholding

method. Although accuracy of the Bayesian model was not as good as that for the Decision Tree, its false negative response was the best. As it was mentioned before, the main reason for the lower accuracy of Bayesian classifier was the presence of a great number of irrelevant attributes in the dataset. Among fifty-four attributes only three or four were valuable. After visually examining some "With Particle" images, it was determined that max min thresholding technique provided the best overall threshold setting for this data.

Table 9: Results of data mining techniques using Moving Average thresholding

Dataset	Method	One-R	C4.5	IB3	Bayes
2	Mov.Avg.	99.5%	99.85%	99.5%	98.7%
3	Mov.Avg.	98%	97.7%	98.1%	97.5%

5 Conclusions

For pre-processed images, the accuracy of One-R, Decision Tree, and Three Nearest Neighbours classification techniques was indistinguishable from that of a human expert. This accuracy was very close to 100%. One-R and Decision Tree were the best classification techniques for an automated particle-monitoring system which to be used for blurry unfocused images. The most important attributes were minimum pixel density, mean pixel density and standard deviation of the pixel density. The Bayesian classifier was not as accurate as the other techniques; however, it showed the best false negative response for blurry images. Having no false negative prediction error for our application was very desirable from a process control viewpoint. The main reason for low accuracy of the Bayesian model was the presence of a great number of irrelevant attributes in the datasets. Creating a smaller dataset and selecting the main attributes from the decision tree, significantly improved the accuracy of the Bayesian model.

Acknowledgement

We wish to thank Forouz Farahani and Jan Lutzen (University of Toronto) for the visual examination of some image data utilized in this work. Also, we are very pleased to acknowledge funding provided by the Natural Sciences and Engineering Research Council of Canada.

References

- [1] Ing L.D., S.T. Balke, N. Stoev, H. Tamber, and M. Planeta *In-line Monitoring of Particles in a Polymer Melt During Extrusion Using a Scanning Particle Monitor*, Society of Plastics Engineers ANTEC Conference, Technical Papers, 2001.
- [2] Saha P.K. and K. Udupa, *Optimum Image Thresholding via Class Uncertainty and Region Homogeneity*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, No.7, pp. 689-706, 2001.

- [3] Glassy C.A. *An Analysis of Histogram-Based Thresholding Algorithm*, CVGIP: Graphical Model And Image Processing, Vol.55, No.6, November, pp.532-537, 1993.
- [4] Weszka J. *A Survey of Threshold Selection Techniques*, Comp. Graph. & Image Proc. 7, pp. 259-265, 1978.
- [5] Witten I. and Frank E. *Data Mining*, Morgan Kaufmann Publisher 2000.
(<http://www.cs.waikato.ac.nz/ml/weka>)