Optical Filter Selection for Automatic Visual Inspection

Matthias Richter*
*Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
matthias.richter@kit.edu

Jürgen Beyerer*,†
† Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)
Karlsruhe, Germany
juergen.beyerer@iosb.fraunhofer.de

Abstract

The color of a material is one of the most frequently used features in automated visual inspection systems. While this is sufficient for many "easy" tasks, mixed and organic materials usually require more complex features. Spectral signatures, especially in the near infrared range, have been proven useful in many cases. However, hyperspectral imaging devices are still very costly and too slow to use them in practice. As a work-around, off-the-shelve cameras and optical filters are used to extract few characteristic features from the spectra. Often, these filters are selected by a human expert in a time consuming and error prone process; surprisingly few works are concerned with automatic *selection of suitable filters. We approach this problem by stating filter selection as* feature *selection problem. In contrast to existing techniques that are mainly concerned with filter design, our approach explicitly selects the best out of a large set of given filters. Our method becomes most appealing for use in an industrial setting, when this selection represents (physically) available filters. We show the application of our technique by implementing six different selection strategies and applying each to two real-world sorting problems.*

1. Introduction

At the present time, automated visual inspection of bulk material is primarily achieved by utilizing color information. Such solutions ensures high throughput and economic feasibility, but hit a wall when the materials under inspection are of similar color (low *inter-class* variance) or when the materials simultaneously occupy many regions of the color space (large *inter-class* variance). Both is commonly the case with organic materials, like fruit and crop, but also applies to inorganic substances such as minerals and alloys. Often, reliable discrimination is still possible by exploiting reflectance-characteristics outside of the visible spectrum, especially the near infrared, or by utilizing narrow banded, faint fluorescence and luminescence effects. One might be tempted to use the full "spectral fingerprint" of a material for classification by including a hyperspectral imaging device in the inspection pipeline. However, such devices are more expensive, have a low spatial resolution and require brighter illumination or longer exposure times than existing off-the-shelve industrial cameras. Furthermore, these devices produce much higher data volume, which in turn increases the time required for data transfer and processing. These factors make such systems impractical in an industrial setting. A common workaround solution combines off-the-shelve cameras with optical filters. Spectral signatures of the materials under inspection are obtained in the laboratory and analyzed to determine discriminative wavelength bands. Suitable optical filters are acquired or manufactured accordingly. The resulting visual inspection system uses only the reduced, usually one- to four-channel image to perform the sorting task. This approach is all the more attractive, since existing solutions can often be repurposed with minimal effort.

There are two general methods to determine the filters: top-down (design) and bottom-up (selection). In the *design* approach, filter transmission functions are designed based on the results of the analysis and realized using e.g. thin-film optical filters. The resulting solution is optimal for the task at hand, but – depending on the complexity of the transmission function – relatively expensive. *Selection*, on the other hand, chooses the best few from a pool of filters. While this pool may contain arbitrary transmission functions, an interesting case emerges when it is matched with optical filters in a catalogue. This solution is often sub-optimal, but since the filters can be mass-produced, it is generally more cost-effective than the design approach. This work focuses on the second approach, selection, for application in an industrial setting.

1.1. Related Work

The visual inspection community has long since acknowledged the usefulness of filter selection based on hyperspectral imaging. For example, Kleynen et at. selected

a combination of four band-pass filters from a pool of 24 possibilities in order to detect defects in "Jonagold" apple fruits [9]. They rated each combination using the correct classification rate of a quadratic discriminant analysis classifier on the filter responses. In [16], Piron et al. use a similar method to select up to four filters (out of 22) to discriminate weeds from crop. While this exhaustive search works with a small pool of filters, it does not scale well due to combinatory explosion.

Other approaches do not focus on finding the best performing filter combination, but rather identify the most discriminative wavelengths to guide a subsequent (manual) filter selection. Osborne et al. use the regression coefficients obtained in partial least squares analysis as proxy to rank wavelengths [12]. This approach can be used to select both an optimum (with respect to discriminative power), or fixed number of wavelengths. In [5], Feyaerts and van Gool rank wavelengths using the Fisher criterion, i.e. the ratio of variability between, and variability within classes. The highest ranking wavelength is selected automatically, while lower ranking wavelengths are only considered when they are positioned "sufficiently far" from the already selected ones. Similarly, Chao et al. perform a stepwise selection according to the Fisher criterion in a five-class classification problem [3]. However, unlike Feyaerts and van Gool, the ranking in each step is computed with respect to the already selected wavelengths. Similar ideas can be found in the remote sensing field: Pal uses the coefficients of the weight vector of (i) a support vector machine and (ii) sparse multinomial regression to create a ranking. The intuition is that, similar to the approach of Osborne et al., both parameters encode the relative importance of each wavelength [13, 14]. Guo et al. utilize mutual information of each band with a set of key-spectra that they expect to find in the hyperspectral images [7].

Alternative methods lend ideas from filter design: De Backer et al. parameterize a set of band-pass filters by their central wavelength and band-width [4]. The parameters are jointly optimized by adaptive simulated annealing using the Bhattacharya bound (which is an upper bound on the Bayes error) as merit function. Similarly, Nakauchi et al. optimize band-pass parameters – lower and upper wavelength – by a global, random sampling based search followed by local optimization [11]. In both steps the Fisher criterion serves as merit function.

All these band selection and parameter optimization approaches show promising theoretical developments in their respective application areas. However, there is no guarantee that matching *physical* optical filters are available or even realizable in an economically feasible way. Therefore it is worth to take a step back and look at the problem in a different light.

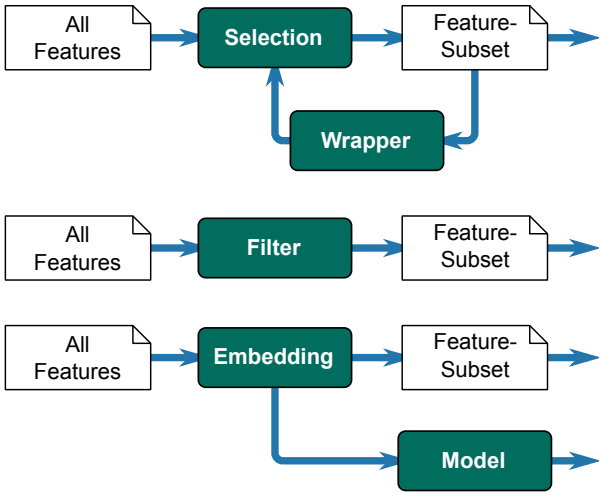


Figure 1. Schema of the different approaches to feature selection. Top: Wrapper methods, Middle: Filter Methods, Bottom: Embedded Methods.

2. Methods

Filter selection can be formalized in the following way: Given a set of filter transmission functions $\mathcal{F}$, a ground truth dataset $\mathcal{T}$, and a merit function γ, the goal is to select an optimal set of filters, i.e. a subset $\mathcal{S} \subseteq \mathcal{F}$ so that $\gamma(\mathcal{S}, \mathcal{T})$ is at a maximum. By simply replacing the words "filter transmission functions" with "features" one arrives at a formal definition of *feature selection* as known from machine learning research. This is an important insight, as it allows to use the numerous methods found in literature. Generally, these methods can be divided into three classes: wrapper, filter and embedded methods (see Fig. 1).

Wrapper methods select a feature-subset according to some selection parameters or feature ranking. A model is trained using the subset, and its prediction performance is used to re-parameterize the feature selection. The process is repeated until some stopping criterion is reached. While straightforward, this method can be very time consuming and tends so over-fit the selection to the chosen predictor. *Filter methods* on the other hand select a subset according to some classifier-independent, objective criterion. Since the election contains the (globally) most relevant features, it is expected to work equally well on different classifiers. However, the selection may be suboptimal when only a specific model is concerned. Finally, *embedded methods* combine wrappers and filters by embedding the feature selection process in a learning algorithm in a fundamental way – hence the name.

2.1. Preliminary Considerations

Before developing wrapper, filter and embedded feature selection methods, it is necessary to fill in some de-

tails in the definitions of the preceding paragraphs: The ground truth dataset $\mathcal{T} = \{(\mathbf{s}_i, y_i) | i = 1, \ldots, N\}$ consists of N training samples, where $\mathbf{s}_i \in \mathbb{R}^b$ denotes a measured point spectrum with b spectral bands, and $y_i = \pm 1$ denotes the associated class. The K filter transmission functions (features) $f \in \mathcal{F}$ map a spectrum to a scalar, that is $g = f(\mathbf{s})$ maps $\mathbf{s}$ to a total light intensity g that would be observed by a camera[1]. $\mathcal{S} \subseteq \mathcal{F}$ denotes the set containing the filter selection and the complement $\overline{\mathcal{S}} = \mathcal{F} \setminus \mathcal{S}$ contains the $K - |\mathcal{S}|$ unselected features.

Some selection methods require discrete features to be efficiently computable. The discretization of the feature f_k will be denoted h_k. The method can be chosen arbitrarily, but in this work, features are binarized according to

$$h_k(\mathbf{s}) = \begin{cases} 1 & \text{if } f_k(\mathbf{s}) \leq \tau_k \\ -1 & \text{otherwise,} \end{cases} \tag{1}$$

where τ_k is chosen to minimize training error of Gaussian MAP classification using only f_k as feature.

Note that while this binarization is relatively straightforward, it also discards a great deal of information. For the purpose of this paper this is of little concern, however in practice one should either use finer grained discretization or some technique capable of handling continuous data.

2.2. Wrapper: Linear Discriminant Analysis

A simple wrapper method can be derived from Fisher's linear discriminant analysis (LDA). Briefly, LDA determines a projection direction $\mathbf{w}$ so that class separation of the projected training samples $\mathbf{w}^\top \mathbf{f}_i$ is maximal (here $\mathbf{f}_i$ denotes a feature vector extracted from $\mathbf{s}_i$). The solution to $\mathbf{w}$ is obtained by maximizing the Fisher criterion

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}. \tag{2}$$

Here $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class and within-class covariance matrices of the training samples. By differentiating $J(\mathbf{w})$ with respect to $\mathbf{w}$, it can be shown that $J(\mathbf{w})$ is maximized by

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_{-1}), \tag{3}$$

where $\mathbf{m}_y$ denote the means of training samples in the respective classes [1]. A linear classifier is constructed by choosing a threshold τ to separates the projected features.

Wrapper methods usually utilize cross-validation to evaluate feature subsets by means of classification error. While generally applicable, this process is relatively slow since a classifier has to be trained and evaluated for each fold and subset. LDA, however, allows for an elegant shortcut: Provided that an optimal threshold was chosen, the classification error depends only on $\mathbf{w}$. Therefore $J(\mathbf{w})$ acts as a surrogate of the classifier performance. This motivates the following greedy feature selection method: Starting with an initially empty selection $\mathcal{S}_0 = \emptyset$, features are iteratively added to maximize the Fisher criterion, i.e.

$$f_{t+1} = \arg\max_{f \in \overline{\mathcal{S}}_t} J(\mathbf{w}_{t+1}). \tag{4}$$

The projection direction $\mathbf{w}_{t+1}$ is computed using the feature candidate f and the selection of the last step $\mathcal{S}_t$. After each selection step features may be unselected (to ensure minimality of the selection) if the removal has negligible impact on the selection criterion eq. (2).

2.3. Filter: Conditional Likelihood Maximization

In contrast to wrappers, filter methods evaluate a given subset of features by means of some utility function that does not depend on a specific classifier. Well known methods include RELIEF [10], Correlation-based Feature Selection [8] and measures derived from information theory. Many of the latter methods were unified in the Conditional Likelihood Maximisation (CLM) framework presented by Brown et al. [2]. They derive a 'root' criterion by developing the log-likelihood of a hypothetical parametric model. This root criterion is conditional mutual information (CMI) of a feature candidate f_k with the class labels y_i conditioned on the already selected features $\mathcal{S}$,

$$J_{cmi}(f_k, \mathcal{S}) = I(F_k; Y | F_\mathcal{S}). \tag{5}$$

Here, F_k, $F_\mathcal{S}$, and Y are random variables corresponding to the feature f_k, the selection $\mathcal{S}$, and the class labels y respectively. Similar to the LDA approach above, features are iteratively selected to maximize eq. (5), that is

$$f_{t+1} = \arg\max_{f \in \overline{\mathcal{S}}_t} J_{cmi}(f, \mathcal{S}_t). \tag{6}$$

Iteration is stopped if $J_{cmi}(f_{t+1}, \mathcal{S}_t) < \tau$, that is if adding the feature does not provide sufficient additional information. Similarly, a feature may be removed from the selection if omission does not result in losing too much discriminative information. Since the criterion in eq. (5) is computationally intractable when many features are considered, it is assumed that the selection is independent and class-conditionally independent given the unselected feature candidate f_k. Under these assumptions J_{cmi} can be replaced by an equivalent criterion,

$$\begin{aligned} \hat{J}_{cmi}(f_k, \mathcal{S}) = I(F_k; Y) &- \sum_{f_j \in \mathcal{S}} I(F_j; F_k) \\ &+ \sum_{f_j \in \mathcal{S}} I(F_j; F_k | Y). \end{aligned} \tag{7}$$

[1] Note that the filter functions can be chosen arbitrarily; if only one band is extracted, the resulting method will in fact be a *band selection* technique.

Brown et al. provide an interpretation for each term: The first term encodes *relevance*, that is good features should explain the class labels. The second term encodes *redundancy*, i.e. features that do not add new information about the class labels should not be selected. The third term encodes *conditional redundancy*: Redundant features may still be selected, provided that the correlation inside the classes is stronger than the overall correlation. Using these interpretations, successful methods can be analyzed by reformulating them in the context of CLM.

The *Minimum-Redundancy Maximum-Relevance* (MRMR) criterion [15], for example, can be expressed as

$$J_{mrmr}(f_k, \mathcal{S}) = I(F_k; Y) - \frac{1}{|\mathcal{S}|} \sum_{f_j \in \mathcal{S}} I(F_j; F_k). \quad (8)$$

Eq. (8) can be interpreted such that MRMR assumes class-conditional pairwise independence of selected features (thereby dropping the conditional redundancy term), and with a growing selection gradually adopts the belief of pairwise independence of the selected features [2].

The *Joint Mutual Information* (JMI) criterion [18] can be seen as introducing conditional redundancy to MRMR, i.e.

$$J_{jmi}(f_k, \mathcal{S}) = J_{mrmr}(f_k, \mathcal{S}) + \frac{1}{|\mathcal{S}|} \sum_{f_j \in \mathcal{S}} I(F_j; F_k | Y). \quad (9)$$

However, the effect conditional redundancy is again gradually weakened with growing the feature selection.

In the context of this work, both criteria can be slightly altered to introduce prior knowledge about the features: Two features are likely pairwise and class-conditionally pairwise independent, if the corresponding filters do not overlap. This gives rise to the similarity-MRMR (SMRMR) and similarity-JMI (SJMI) criteria,

$$J_{smrmr}(f_k, \mathcal{S}) = I(F_k; Y) - \sum_{f_j \in \mathcal{S}} s_j^k I(F_j; F_k), \quad \text{and} \quad (10)$$

$$J_{sjmi}(f_k, \mathcal{S}) = J_{smrmr}(f_k, \mathcal{S}) + \sum_{f_j \in \mathcal{S}} s_j^k I(F_j; F_k | Y). \quad (11)$$

Here s_j^k encodes the overlap of the filters f_j and f_k. Specifically $s_j^k = 0$ denotes no overlap, whereas $s_j^k = 1$ means that f_j and f_k are the same filter.

2.4. Embedded: AdaBoost

Embedded methods position themselves between wrappers and filters. Like wrappers, they utilize a model to select features. However, the selection is not based on predictive performance, but rather a direct result of the learning algorithm. In the following, feature selection is embedded into Freund and Schapire's AdaBoost algorithm [6].

The goal of AdaBoost is to combine of several weak classifiers h_t, that may perform barely better than chance, to a strong classifier H. This classifier predicts class labels according to a weighted sum of the votes of each h_t,

$$H(\mathbf{s}) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{s}) \right). \quad (12)$$

The weak classifiers and corresponding weights α_t are selected in an iterative process: A weight distribution W_t, i.e. $W_t(i) \geq 0$ and $\sum_i W_t(i) = 1$, encodes the importance of each training sample, where initially each training sample is equally important. In the t-th iteration, h_t is selected to minimize the weighted error rate on the training samples,

$$h_t = \arg\max_h \left| \frac{1}{2} - \varepsilon(h) \right|, \quad \text{where} \quad (13)$$

$$\varepsilon(h) = \sum_{i=1}^{N} W_t(i) \big[h(\mathbf{s}_i) \neq y_i \big]. \quad (14)$$

The weight α_t is computed from the weighted training error $\varepsilon(h_t)$, typically as log-odds of the (weighted) correct classification rate,

$$\alpha_t = \log \frac{1 - \varepsilon(h_t)}{\varepsilon(h_t)}. \quad (15)$$

Finally, the weight distribution is updated so that the training samples that h_t classified incorrectly will be more important in the next round:

$$W_{t+1}(i) = \frac{W_t(i) \exp\left(\alpha_t \big[h_t(\mathbf{s}_i) \neq y_i \big] \right)}{\sum_{i=1}^{N} W_{t+1}(i)}. \quad (16)$$

Iteration is stopped if either a maximum number of classifiers is selected, or if $\varepsilon(h_t)$ is not significantly different from random choice.

By recalling the feature discretization in Section 2.1 it is apparent how AdaBoost can be used for feature selection: Each discretized feature is itself a weak learner. The classifier ensemble then represents the feature selection, where $|\alpha_t|$ encodes the importance of the feature f_t.

3. Application

We now turn our attention to the application of the proposed methods to two real-world sorting problems: (a) discriminating tobacco from cotton string, feathers, and grass and (b) discriminating maize polluted with the carcinogenic mycotoxin Aflatoxine B_1 from uncontaminated kernels.

Spectral measurements of tobacco, cotton string, feathers and grass were obtained using a hyperspectral line-scan camera sensible in the short-wave infrared spectrum with a spectral resolution of 256 bands from $905nm$ to $2513nm$. 215 band-pass filter candidates with a central wavelength λ_c

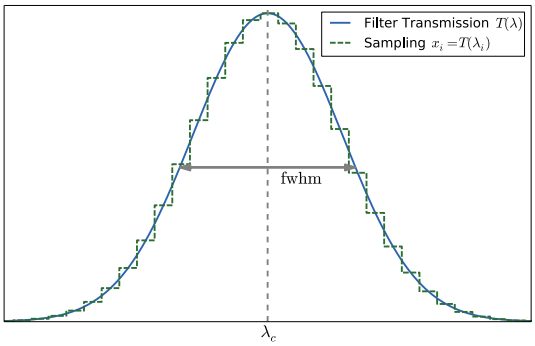


Figure 2. Filter model and sampling for feature computation.

between $1050nm$ and $2380nm$ and full width at half maximum (fwhm) of $10nm, 25nm$ and $50nm$ were chosen according to an optics catalogue.

In the maize sorting problem, the grains were illuminated with ultraviolet light in order to cause fluorescence in the contaminated grains. Since the emitted light falls in the visible spectrum, images were recorded in the spectral range of $315nm$ to $1170nm$ with a spectral resolution of 270 bands. 113 band-pass filter candidates with λ_c from $337nm$ to $1064nm$ with fwhm of $10nm, 25nm$ and $50nm$ were again sampled from an optics catalogue.

Since in both cases the exact characteristics of the filters were not known, the transmission spectra were approximated according to

$$T(\lambda) = \exp\left\{-4\log 2 \cdot \left(\frac{\lambda - \lambda_c}{\text{fhwm}}\right)^2\right\}. \qquad (17)$$

Features were then computed as dot product of the characteristic filter vector $\mathbf{x}$, whose components $x_i = T(\lambda_i)$ were sampled on the central wavelengths λ_i of the corresponding components of $\mathbf{s}$, and the spectrum $\mathbf{s}$

$$f(\mathbf{s}) = \mathbf{s}^\top \mathbf{x}. \qquad (18)$$

Figure 2 illustrates the relationship between the filter model $T(\lambda)$ and sampling x_i. In case of filter selection by SMRMR and SJMI, similarity was computed using the cosine measure:

$$s_j^k = \frac{\mathbf{x}_j^\top \mathbf{x}_k}{\|\mathbf{x}_j\|_2 \, \|\mathbf{x}_k\|_2}. \qquad (19)$$

3.1. Results

Table 1 shows each method's suggestion of three filters for the maize sorting problem. All methods but AdaBoost selection agree on the filter centered on $\lambda_c = 775nm$ with fwhm $= 50nm$, but differ in suggestions of complimentary filters. This shows that, as hinted in Section 2, each method bases the selection on different underlying assumptions of the data.

Method	**Three best filters** ($\lambda_c[nm], \text{fwhm}[nm]$)		
AdaBoost	$(800, 50)$,	$(515, 10)$,	$(766, 10)$
LDA	$(575, 50)$,	$(775, 50)$,	$(1050, 50)$
MRMR	$(575, 50)$,	$(775, 50)$,	$(676, 10)$
SMRMR	$(775, 50)$,	$(850, 25)$,	$(730, 10)$
JMI	$(750, 50)$,	$(775, 50)$,	$(800, 50)$
SJMI	$(775, 50)$,	$(825, 25)$,	$(730, 10)$

Table 1. Characteristics of the first three filters selected by the different methods on the maize sorting problem.

To evaluate which assumptions best reflect the given problem, a random forest classifier was trained using up to ten selected features and the classification error was estimated in a five-fold cross-validation. To provide a lower bound on the achievable error rate the classification error using all available features was also determined. Results on both sorting problems are shown in Figure 3.

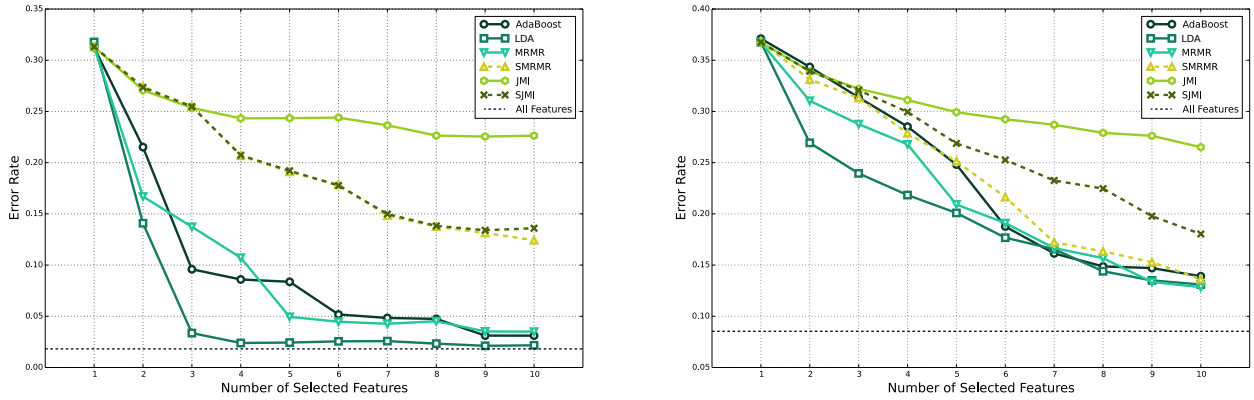
In both sorting problems, feature selection by wrapping Fisher LDA provides the best selection, while MRMR also tends to deliver reasonable suggestions. In case of the tobacco sorting problem, AdaBoost also selects good features. Contrary to expectation, encoding prior knowledge in form of feature similarity, i.e. SMRMR and SJMI, did not improve on the selection results.

4. Conclusion

Numerous works have shown the benefit of using optical filters derived from spectral analysis for visual inspection tasks. There are two general approaches to automate the search for suitable filters: design of a specialized transmission function and selection from a pool of possibilities. The design approach generally produces filters optimally suited for the task at hand, but the high manufacturing costs hamper application in an industrial setting. Selection, on the other hand, results in low costs due to the usage of off-the-shelve filters, although the solution may be suboptimal. While methods suitable for filter design, especially from the field of remote sensing, are available, surprisingly few works are considered with automatic filter selection.

In a comprehensive approach, we explicitly reduced filters selection to *feature selection* as known in the machine learning literature. We then exemplified our approach by implementing a wrapper method based on LDA, filter methods using information theoretic measures, and by embedding the selection into the AdaBoost algorithm. Although targeted at visual inspection at an industrial scale, the presented approach is flexible enough to also be used for band selection in remote sensing applications.

Evaluation on two different real world sorting problems shows the characteristics of each approach. In both cases, the wrapper method produced the best selection, followed



(a) Tobacco sorting.

(b) Maize sorting.

Figure 3. Error rate of a random forest classifier trained using up to ten features selected using the different methods presented in Section 2. The dashed black line marks the error rate when all available features are used.

by MRMR and AdaBoost. JMI, SJMI and SMRMR did not provide robust selections. One reason that the LDA method appears to fare better may be explained by the feature discretization step required by the other methods, but not LDA – it simply has more information at its disposal. This suggests that CLM methods may be enhanced by finer grained estimation of the mutual information terms. Likewise, the embedded selection might be improved by using e.g. Real AdaBoost [17] instead of AdaBoost.

References

[1] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. 3

[2] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012. 3, 4

[3] K. Chao, Y. Chen, W. Hruschka, and B. Park. Chicken Heart Disease Characterization by Multi-spectral Imaging. *Applied engineering in agriculture*, 17(1):99–106, 2001. 2

[4] S. De Backer, P. Kempeneers, W. Debruyn, and P. Scheunders. A Band Selection Technique for Spectral Classification. *Geoscience and Remote Sensing Letters, IEEE*, 2(3):319–323, 2005. 2

[5] F. Feyaerts and L. Van Gool. Multi-spectral vision system for weed detection. *Pattern Recognition Letters*, 22(6):667–674, 2001. 2

[6] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In *Computational learning theory*, pages 23–37. Springer, 1995. 4

[7] B. Guo, S. R. Gunn, R. Damper, and J. Nelson. Band Selection for Hyperspectral Image Classification Using Mutual Information. *Geoscience and Remote Sensing Letters, IEEE*, 3(4):522–526, 2006. 2

[8] M. A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999. 3

[9] O. Kleynen, V. Leemans, and M.-F. Destain. Selection of the most efficient wavelength bands for 'Jonagold' apple sorting. *Postharvest Biology and Technology*, 30(3):221–232, 2003. 2

[10] I. Kononenko. Estimating Attributes: Analysis and Extensions of RELIEF. In *Proceedings of the European Conference on Machine Learning*, pages 171–182. Springer, 1994. 3

[11] S. Nakauchi, K. Nishino, and T. Yamashita. Selection of optimal combinations of band-pass filters for ice detection by hyperspectral imaging. *Optics Express*, 20(2):986–1000, 2012. 2

[12] S. D. Osborne, R. Künnemeyer, and R. B. Jordan. Method of Wavelength Selection for Partial Least Squares. *Analyst*, 122(12):1531–1537, 1997. 2

[13] M. Pal. Margin-based feature selection for hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, 11(3):212–220, 2009. 2

[14] M. Pal. Multinomial logistic regression-based feature selection for hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, 14(1):214–220, 2012. 2

[15] H. Peng, F. Long, and C. Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005. 4

[16] A. Piron, V. Leemans, O. Kleynen, F. Lebeau, and M.-F. Destain. Selection of the most efficient wavelength bands for discriminating weeds from crop. *Computers and Electronics in Agriculture*, 62(2):141–148, 2008. 2

[17] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999. 6

[18] H. Yang and J. Moody. Feature Selection Based on Joint Mutual Information. In *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pages 22–25. Citeseer, 1999. 4