A new look at feature selection

G. Schwarz[1], D. Espinoza Molina[1], M. Datcu[1, 2]
[1]German Aerospace Center (DLR), Remote Sensing Technology Institute, 82234 Wessling, Germany
[2]Paris Institute of Technology GET/Télécom Paris, 75013 Paris, France

***Abstract*—One of the basic components of image information mining (IIM) systems is feature extraction. Feature extraction delivers a low level "building block" decomposition of the input data. In principle, feature extraction results may depend on the characteristics of the images to be analyzed. In order to avoid a critical dependence on a specific concept, we advocate a general feature finder toolbox approach that handles typical remote sensing images with diverse geometrical and texture characteristics. Our concept considers high resolution optical as well as synthetic aperture radar (SAR) images.**

***Index Terms*—Feature extraction, image information mining, multi-scale, remote sensing, SAR, speckle, TerraSAR-X.**

I. Introduction

A fundamental step of image information mining is feature selection. Feature selection provides us with decomposed "building blocks" of each image that are needed during subsequent data analysis steps [1].

Thus, the choice of a feature selection approach is a strategic decision for any image information mining system. Misconceptions may lead to sub-optimal performance of subsequent image classification and image segmentation. Typical feature extraction algorithms either exploit spatial textures or multispectral characteristics of single pixels. Model-based stochastic approaches often allow multi-variate modeling (e.g. for texture modeling).

As there is a great variety of different instruments and user requirements (e.g., passive or active instruments, wavelength ranges, single or multi-channel designs, needs for the generation of calibrated product data), the feature selection performance may be influenced by the specific characteristics of the images to be handled. On the other hand, what we need is a robust and efficient "generic" implementation of feature extraction in order to reach a good separation of object or target categories [2].

II. Basic Design Aspects

During the design phase of an IIM system, several potential concepts and performance metrics for a feature selection component have to be compared. To this end, we need a stepwise approach.

The first step is theoretical and based on information theory. Information theory tells us about fundamental limits, i.e. what we can reach and what remains unattainable. As a result, we obtain fundamental limits of a selected concept.

The second step needs experimental image processing. Experience with typical image data of the envisaged application fields will lead to typical performance characteristics from a technically oriented perspective. This may result in solutions such as wavelet tools or stochastic modeling.

The third step is driven by the practical needs of image interpreters. Here, the practical needs of human interpreters may call for dedicated solutions. This includes confidence criteria and performance metrics based on validation tests.

III. Feature Extraction: Prerequisites

When we look at typical remote sensing images, we are faced with widely different image characteristics: images of different sensors may deliver data with diverse features; on the other hand, even images of a single instrument may be of different data quality. When feature extraction from a single or from multiple images is used for subsequent classification, then classification results may depend on these instrumental characteristics.

As a consequence, a comparative analysis of image data may call for initial image pre-processing adjusting calibration quality and processing levels, signal-to-noise ratios, brightness levels, variances, histogram shapes and probability density functions, image and object scale, and the removal of degraded sub-scenes (e.g., cloud covered sub-scenes).

In the following, we tacitly assume that all necessary preprocessing and matching steps have already been performed.

IV. Feature Finder: A Toolbox Approach

In our case, we advocate a general *feature finder* toolbox approach with sufficient feature selection algorithms and parameterization capabilities to cope with the diverse geometrical and texture characteristics of typical remote sensing images. These images may be optical images of selected spectral bands as well as SAR images being affected by speckle.

In our concept all ingested images will undergo automated feature extraction by several competing algorithms run in parallel. Typical feature extraction algorithms are texture analyses and color analyses. The analysis of texture features relies on pixel neighborhood relationships and can be performed on multiple scales; in contrast, a color analysis can be confined to single pixel combinations in multiple color bands. The results of each extraction algorithm will be feature vectors as shown in Fig. 1. The feature vectors derived from each pixel or window will be arranged in a way to support spatial backtrack to the original image; the feature vectors will be used later in an interactive feature selection step to pick out those features that are best for a specific task or application (see below).

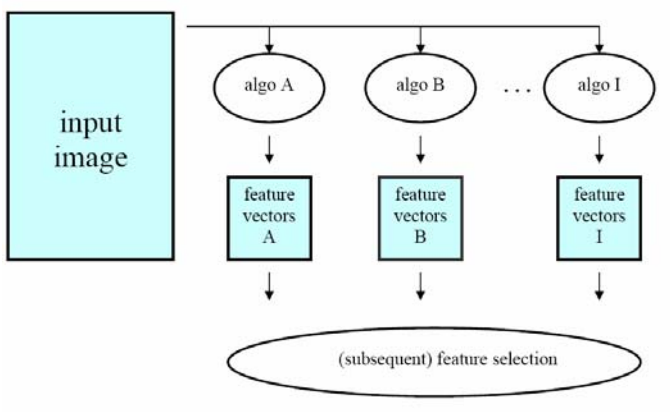

Fig. 1. Extraction of feature vectors.

This automated feature extraction concept results in a new product type that shall be generated: the *feature map*. A feature map results from the application of the available toolbox routines and contains the image data together with its derived features arranged in so-called class files.

The image data and class files can be accessed via a dedicated viewer (i.e. a *feature browser*) permitting quantitative feature analyses by interactive selection of "best" features taken form the class files. This concept is illustrated in Fig. 2.

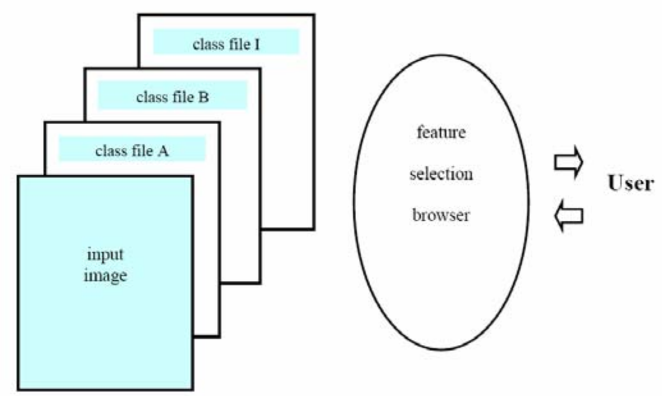

Fig. 2. Feature browser principle.

The feature browser supports the interactive analysis of the total information content of the selected image.

This general approach can be considered as a universal tool that opens the way towards flexible and upgradeable feature selection when additional or improved feature selection algorithms become available, or when specific applications call for dedicated feature selection.

In particular, this is an interesting approach for the detection and analysis so-called "small features" (i.e. infrequently occurring features that are normally lost during conventional clustering), or for time-dependent variations of image contents.

The following example shall give an idea of the capabilities of our toolbox approach. Fig. 3 shows a typical SAR image (TerraSAR-X sub-scene acquired near the Bois de Boulogne in Paris, France) with speckle. For instance, if a user is looking for specific roof top structures of buildings, a conventional single method feature selection approach would probably fail.

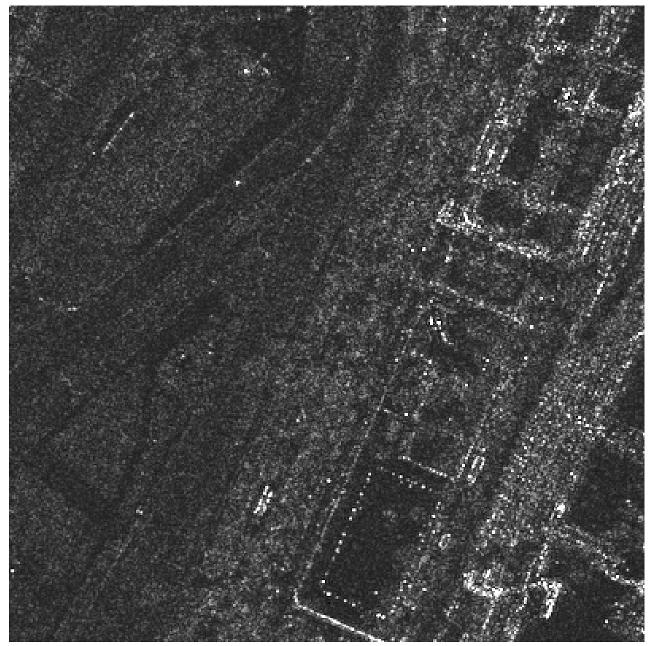

Fig. 3. Speckled TerraSAR-X scene of Paris / Bois de Boulogne.

V. Feature Map Aspects

Our proposal of an explicit feature map as a new remote sensing product type of Earth observation images deserves to be explained in more detail.

Features represent a decisive intermediate stage during information extraction form images. On the one hand, a feature map contains a compact representation of condensed information that can serve as a generic image content survey; on the other hand, advanced information extraction methods such as clustering and classification need some common ground to start with.

A feature map should be the common data set serving both worlds mentioned above. When combined with a generic feature viewer, human users can profit form this standard, too: a user can interpret feature maps like a conventional image and derive quantitative thematic information that would remain obscured otherwise.

If the same feature map approach is applied to images of various instruments, then a quantitative inter-comparison of pre-processed feature data becomes a realistic scenario. This opens the way towards standardized *feature level* information extraction without the tedious work of comparing original image data available only in various incompatible formats.

Of course, we must be aware of some basic feature extraction caveats. In practice, most feature extraction routines are not invariant to object scale, shift, rotation, or illumination.

This means that object features extracted from one selected image may differ from object features extracted from another image containing a physically similar object. This issue has to be assessed by representative test cases using alternative feature extraction routines and the selection of processing parameters.

Depending on the application field, the corresponding requirements may be more or less stringent. In practice, preprocessing of the input data may offer a solution in many practical situations: Adaptation of the image brightness histograms is a very pragmatic approach to obtain uniform conditions for comparisons (see Section III).

VI. Multi-Scale And Despeckling Aspects

If we want to characterize textures, a multi-scale approach is a good way to proceed; however, in the case of SAR images, we are faced with the peculiar signal-to-noise characteristics of the multi-scale images. The final "visibility" and classification of objects may hinge on the available multi-scale levels and the applied despeckling.

For many applications, the speckle noise contained in SAR images prevents a robust land cover or urban scene classification. On the other hand, if we apply excessive despeckling, one will notice artifacts appearing in particular around small scale structures. This must be avoided by accurate determination of the speckle characteristics of the input data.

When we apply appropriate despeckling [3], automated feature detection and subsequent classification of urban scenes can be performed (see Fig. 4). Hence, analysis tools for the determination of the characteristics of the input data and a state-of-the-art despeckling routine are important components of our toolbox.

Besides speckle parameters, the basic multi-scale characteristics of our remote sensing images have to be determined prior to their use during feature extraction, and for subsequent clustering and classification. Independent from the imaging characteristics of specific sensors and the processing steps, the results of feature selection, clustering, and classification shall not differ too much. This necessitates the prior selection of appropriate multi-scale levels and processing parameters leading to a stable information extraction [12].

VII. Feature Browsing Methods

The performance provided by the feature browser hinges on the algorithms contained in it. These algorithms shall use compact and robust representations of the available feature vectors.

In the following, we will outline four typical methods that have shown promising results. Depending on the selected feature vector algorithms illustrated in Fig. 1, the feature browsing methods described below are foreseen to be selectable, too.

A. Cross-Validation

The idea of cross-validation is to split all available data into a number of subsets and to compare the analysis results of each subset [4]. In the simplest case, when one uses two subsets, one can use the first subset for training and the second one for validation. Thus, one can verify the performance of a selected algorithm, test its dependence on input data characteristics, and make error estimates.

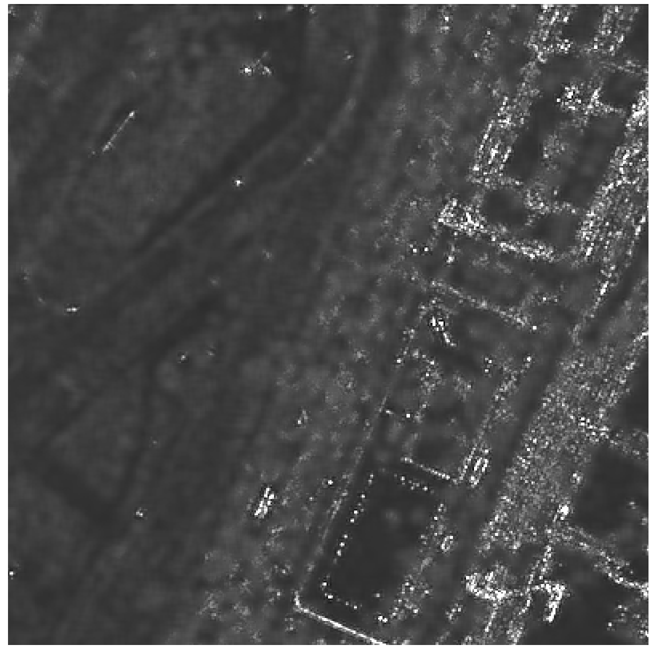

Fig. 4. Despeckled version of Fig. 3.

Cross-validation is a straightforward and robust approach for accuracy estimation and model selection [5]. In our case, the cross validation can be used to optimize the results of the available feature extraction algorithms, their discrimination capability, their sensitivity to the characteristics of the input data, and their overall stability.

The critical point in cross-validation is the selection of representative data for each subset. As a rule, larger number of samples will be less critical than too few samples per subset.

B. Mutual Information

The mutual information measures the mutual dependence of two random variables. In the continuous case, we can define it as [6]:

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right)\,dx\,dy,$$

where $p(x,y)$ is the joint probability density function of X and Y, and $p_1(x)$ and $p_2(x)$ represent their marginal probability density functions. For discrete random variables the integrals are substituted by summations.

The mutual information tells us about the information that X and Y share and represents a measure of dependence. When applied to feature vectors of images, one can derive compact sets of relevant compound descriptors. Thus, we avoid inefficient (redundant) feature combinations. If the selection is done at low computational cost and if the finally selected features allow accurate object classification for the images at hand [7], mutual information can become an attractive selection tool.

One of the advantages of mutual information is its capability to handle correlations as well as anti-correlations (i.e. negative correlation coefficients). The only pre-requisite is that the probability density functions contain sufficient information. Therefore, the probability density functions must be derived from a sufficient number of samples in order to suppress the effects of noisy data and of irregular outliers. Here again, cross-validation can be used as a validation technique.

C. *Fisher Information*

The Fisher information allows us to measure the amount of information that a random variable contains about an unknown parameter with a likelihood function. Under regular conditions, the Fisher information can be considered as a measure of the sensitivity of a likelihood function near its maximum.

Thus, we can use the Fisher information to compare various observation methods of a random process (i.e. we can use it for a comparison of the sensitivity of feature vectors).

The Fisher information can be written as follows [8]:

$$\mathcal{I}(\theta) = \mathrm{E}\left\{ \left[\frac{\partial}{\partial\theta} \ln f(X;\theta)\right]^2 \middle|\, \theta \right\},$$

where

E is the expected value
$\partial/\partial..$ is the partial derivative
ln $f(..)$ is the natural logarithm of the function f
$f(X;\theta)$ is the likelihood function
X is the random variable considered
θ is the unknown parameter

If we can obtain the derivative of the logarithm of the likelihood function with respect to θ in a numerically stable way as a smooth function (no jagged profiles), then the Fisher information becomes a useful inter-comparison tool. Otherwise, numerical problems may occur.

Thus, the usefulness of the Fisher information depends on the numerical accuracy of the logarithm of the derivative of the likelihood function. As we work with discrete functions, we need sufficient and reliable training data to derive the required likelihood functions.

The Fisher information can be used as a measure of system disorder (large I: low disorder; small I: high disorder); it can be compared with Shannon's form of entropy and is related to Kullback-Leibler entropy. For more details, see [9].

D. *Support Vector Machines*

During recent years, the concept of support vector machines (SVMs) has gained much attraction for general classification tasks. One of the reasons for its popularity is the availability of easily portable open source code coming with sufficient documentation [10].

A support vector machine is a software package for classification tasks that is trained with user-selected examples; the trained examples are used to define multi-dimensional planes that separate the classes.

In the field of remote sensing, a number of typical application studies have been published (e.g., [11]); however, systematic studies of machine learning and classification accuracy are to be found elsewhere. Thus, an IIM system containing images decomposed into feature vectors is an attractive field for information theoretic studies: the goal is to produce efficient models predicting classes based on feature combinations.

The SVM capability to run classification tasks based on small training datasets is of particular interest, too; a comparison with neural networks is a mandatory step that needs more detailed analysis than simple application studies.

In addition, the handling of high-dimensional data is another promising characteristic of SVMs. Many SVM packages offer selectable kernel functions. One can hope that some of them support a user-oriented classification of targets. If these kernels result in distance functions beyond simple least squares, machine learning could rival human interpretation.

VIII. Outlook

The optimum use of high resolution images should not be impaired by sub-optimal feature extraction tools. New remote sensing instruments and upcoming application fields for the acquired data will necessitate a number of additional feature extraction and selection methods. Therefore, our concept with upgradeable feature extraction and interactive analysis of class files should be a promising candidate for future image analysis systems. The actual performance of the available algorithms has to be verified by validation tests. A consolidated list of algorithms for the generation of efficient compound features is given in Section VII. In addition, the specific characteristics of an instrument can call for adaptive pre-processing of image data [12].

Acknowledgment

The TerraSAR-X images have been provided by DLR under an approved TerraSAR-X research proposal [13].

References

[1] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P.G. Marchetti, and S.D'Elia, "Information Mining in Remote Sensing Image Archives – Part A: System Concepts," *IEEE Trans. Geoscience and Remote Sensing*, vol. 41, pp. 2923-2936, Dec. 2003.
[2] M. Costache and M. Datcu, "Learning and Unlearning for Mining High Resolution EO Images", *Proc. IGARSS07*, Barcelona, 2007.
[3] F. P. De Vries, *"Speckle Reduction in SAR Imaging by Various Multi-Look Techniques"*, http://handle.dtic.mil/100.2/ADA349197.
[4] http://en.wikipedia.org/wiki/Cross_validation.
[5] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", http://citeseer.ist.psu.edu/ kohavi95 study.html.
[6] http://en.wikipedia.org/wiki/Mutual_information.
[7] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226-1238, Aug. 2005.
[8] http://en.wikipedia.org/wiki/Fisher_information.
[9] B. R. Frieden, *Physics from Fisher Information: A Unification*, Cambridge UK: Cambridge University Press, 2004
[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[11] M. Pal, P. M. Mather, “Support vector machines for classification in remote sensing”, *International Journal of Remote Sensing*, vol. 26, pp. 1007-1011, Mar. 2005

[12] G. Schwarz, D. Espinoza Molina, H. Breit, and M. Datcu, ”Adapting multilooking for joint radiometrical and geometrical SAR image enhancement”, *these proceedings*, 2008

[13] http://sss.terrasar-x.dlr.de/how_to_submit_a_tsx_proposal.pdf.