ACCELERATED KNOWLEDGE-DRIVEN IMAGE MINING SYSTEM FOR DATA FUSION IN BIG DATA

Kevin Alonso; Mihai Datcu
DLR German Aerospace Center, Münchner Str. 20, D-82234 Weßling, Germany

REMOTE SENSING TECHNOLOGY INSTITUTE

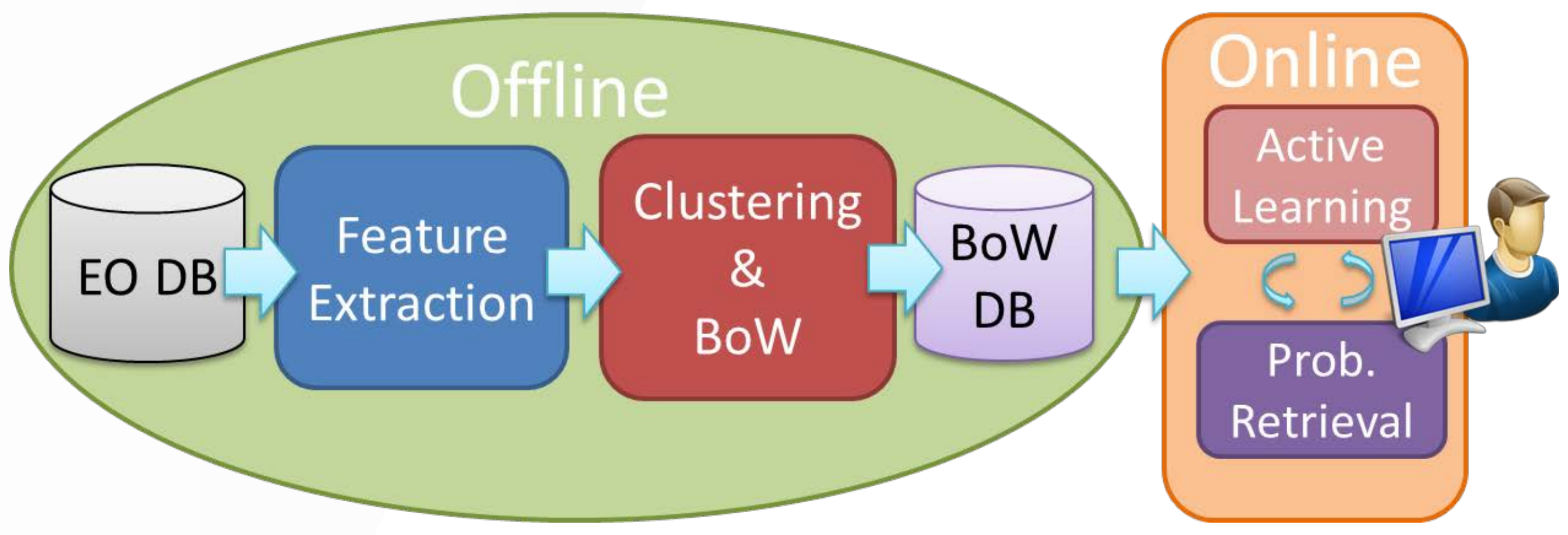
Abstract

In this poster, we present a knowledge-driven content-based information mining system for data fusion in Big Data. The tool combines, at pixel level, the unsupervised clustering results of different number of features, extracted from different image types, with a user given semantic concepts in order to calculate the posterior probability that allows the final search. The system is able to learn different semantic labels based on Bayesian networks and retrieve the related images with only a few user interactions, greatly optimizing the computational costs and over performing existing similar systems in various orders of magnitude.

System Overview

The system is composed by interconnected independent modules. The system Modules and their connections are shown. The initial step of the offline part can take hours but it can be view as real time considering the context of the input stream of the EO products and their generation. The system gets the EO product to be analyzed from the database. In the next step different types of features can be extracted at pixel level. These features are clustered automatically using k-means unsupervised clustering. The clustering results are used for the calculation of a probability density function (PDF) vector for every feature. The PDF is generated using the histogram of the clustering classes occurrence inside the image. This PDF is used as image identification signature following Bag of Words (BoW) [1] principles. At the end of the offline processes the calculated BoW signatures are stored in a database.

In the online part of the system the user interaction enters in scene. This part is real time from the point of view of the user. She/He expects a relatively fast response of the System to her/his actions. The user can introduce positive and negative examples about the specific semantic he/she is looking for and do a search by similarity matrix or probability of the label in the archive.



System Architecture

Active Learning

The Active Learning is performed using Bayesian networks. The learning uses the posterior probabilities expressed as,

$$p(L|D) = \sum_{j} p(L|\omega_i) \cdot p(\omega_i|D)$$

And applying Bayes' formula as,

$$p(L|D) = p(L) \cdot \sum_{i} \frac{p(\omega_i|L) \cdot p(\omega_i|D)}{p(\omega_i)}$$

Being $p(L)$ the prior probability of the semantic label L, $p(\omega_i|D)$ the probability of the classes in the data, $p(\omega_i|L$ the probability of the classes in the label, which can be expressed as the PDF of the classes updated with the user positive and negative examples. Finally $p(\omega_i)$ the prior of signal classes ω_i as,

$$p(\omega_i) = \sum\nolimits_{L} p(\omega_i|L) \cdot p(L)$$

Typically the use of only one feature is not enough to correctly characterize the user semantic L. For this reason the fusion or merging of different features is needed. For linking different features in statistics the assumption of full statistical independence must be taken into account. In [2], the full statistical assumption was made over the features resulting in,

$$p(\omega_{jk\ldots}|L) = p(\omega_j|L) \cdot p(\omega_k|L) \cdot \ldots$$

In [3], the presented algorithm introduces a new statistical assumption, where the statistical independence is enhanced to include the posterior probability. This extension simplifies the calculation complexity of the posterior probabilities from polynomial to linear, deriving in a calculation speed improvement up to four orders of magnitude for the 4 model fusion case.

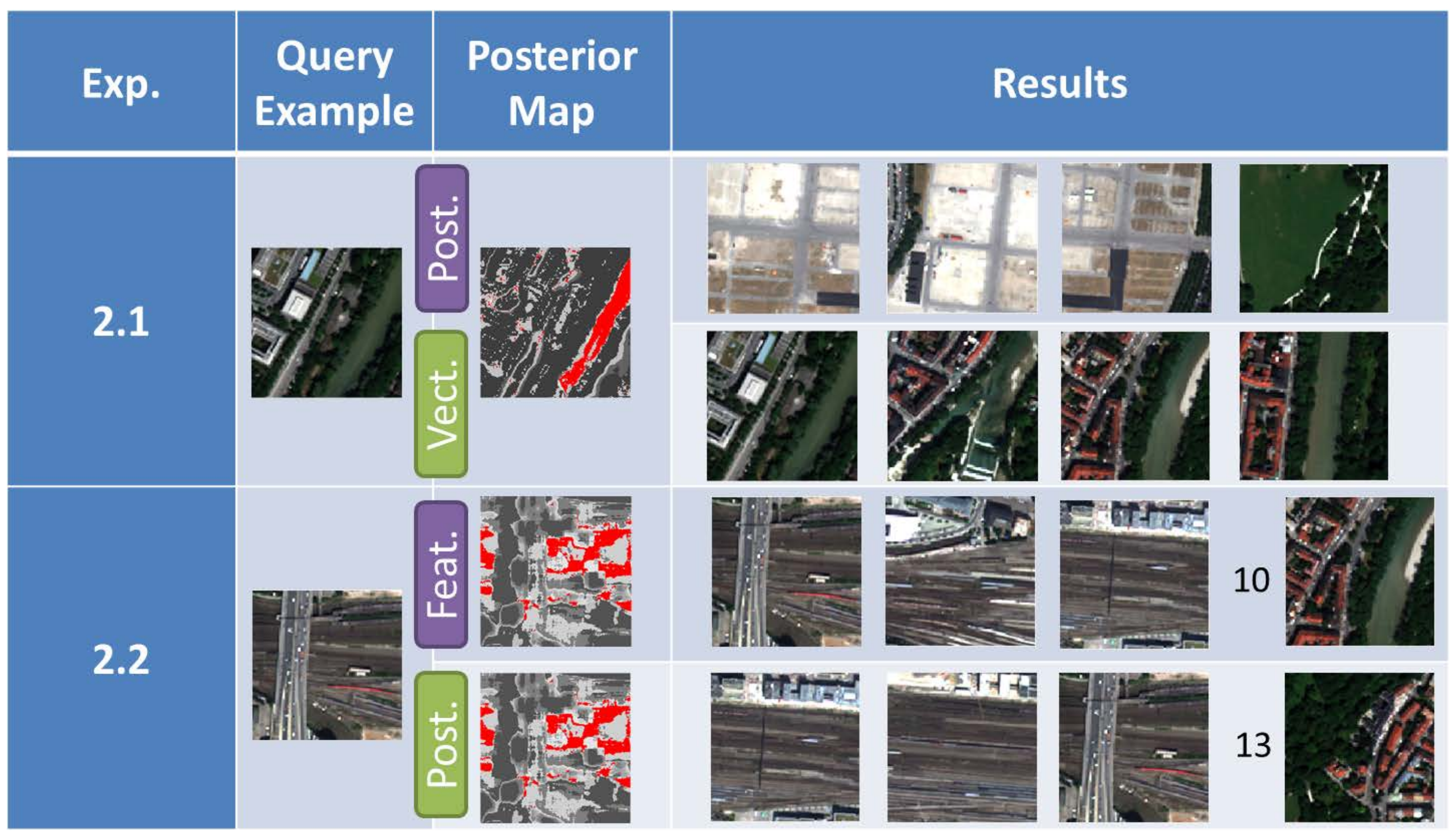
$$p(L|D) = \sum_{j} p(L|\omega_j) \cdot p(\omega_j|D) \cdot \sum_{k} p(L|\omega_k) \cdot p(\omega_k|D) \cdot \ldots$$

Experiments and results

For validating the system we have chosen Munich city EO optical images from WorldView-2 and SAR images from TerraSAR-X both with 1.25 meter resolution. The size of the image is 200x200 pixel, and the total number of images is 500. The parameters used are on the one hand spectral parameters, like multispectral features or Intensity values; and on the other hand texture features, like Weber's Law Descriptor (WLD) [4].

	Search Proc.	1 Model	2 Model	4 Model
Feature Independence	Posterior Probability	0.62s	6.39s	2354s
	Vector Similarity	0.045s	0.15s	43.5s
Posterior Independence	Posterior Probability	0.155s	0.196s	0.31s
	Vector Similarity	0.042s	0.126s	34.61s

System query performance time for different statistical assumptions, query ranking types and model number. First row is used as threshold and corresponds to an emulation of the original KIM implementation. The difference among the performance in the case of four models using the new statistical assumption is four orders of magnitude better.



System query results. First experiment shows a better performance of vector similarity based search approach in a initial learning stages. The second experiment shows how the simplification of the calculation processes in order to speed up the search process is not being affected with the use of posterior probability assumption. Moreover, for some cases the query results are even better.

Conclusions

- We have presented a new implementation of a KIM system.
- We have introduced new search methods and theoretical probabilistic assumptions which outperform in speed the original ones.
- The proposed probabilistic search based on the vector distance between posterior probability vector and images BoW in the databases performs better in weakly defined labels.
- Also we demonstrated a great speed up of the original KIM system with the introduction of posterior probability statistical independence assumption, which after initial tests does not seem to introduce biases in the learning processes.
- As future work, a more intensive evaluation of the obtained results are planned introducing more features for the user-specific semantic definition.
- In long term the introduction and fusion of data features from other type of images (Optical, Multispectral, Multitemporal, GIS) is planned.

References

[1] Y. Huang, Z. Wu, L. Wang, and T. Tan, “Feature Coding in Image Classification: A Comprehensive Study,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, pp. 1, 2013.
[2] 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: system concepts,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2923–2936, Dec. 2003.
[3] K. Alonso and M. Datcu, “Image information mining: an accelerated bayesian algorithm for data fusion of SAR big data,” in *10th European Conference on Synthetic Aperture Radar (EUSAR 2014)*, Berlin, Germany, June 2014.
[4] J. Chen, S. Shan, G. Zhao, and X. Chen, “A robust descriptor based on weber’s law,” in *Computer Vision and Pattern Recognition*, 2008.

Deutsches Zentrum
für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

DLR – German Aerospace Center
Remote Sensing Technology Institute
D-82234 Weßling
http://www.dlr.de/