PARAMETER-FREE CLUSTERING: APPLICATION TO FAWNS DETECTION

Daniele Cerra[1], *Martin Israel*[1] *and Mihai Datcu* [1,2]

[1] German Aerospace Center DLR, Remote Sensing Technology Institute IMF,
82234 Wessling, Germany, {Daniele.Cerra, Martin.Israel, Mihai.Datcu}@dlr.de
[2] Télécom Paris, 46 rue Barrault, 75634 Paris, France

ABSTRACT

Many fawns and other wild animals are killed by mowing machines every year. To prevent them from being killed or injured, a sensor system is being developed to detect the fawns hidden in meadows under mowing. Beside a microwave radar system, two cameras (thermal infrared and RGB) take a picture at the mower's current location. This paper focuses on the parameter-free algorithm that will be adopted to detect the locations containing a fawn hiding in the grass.

Index Terms— Parameter free, similarity measure

1. INTRODUCTION

About 500000 wild animals are killed by mowing machines every year in Germany. In particular, during the first cutting of grass in May or June, many young fawns are killed in their first days of life. Within the research project "Game Guard"[1], a sensor system is being developed for agricultural mowing machines to detect fawns hidden in meadows under mowing: when an alarm is raised appropriate rescue procedures will save the fawns from being injured or killed by the mower. Beside infrared detectors [2] a microwave radar system [3] and cameras (thermal and visible) are scanning the meadows. On the pictures of these cameras the detection algorithm of this paper will be applied.

This paper proposes a parameter-free, model independent methodology based on data compression as a valid alternative to classic image analysis methodologies, which are heavily dependant on the assumed data models. These methods are totally model-free and data-driven, and may be successfully employed for image classification and indexing regardless of sensor characteristics: this allows exploiting with the same approach the information contained in both the optical and infrared images for each location considered.

2. THE DATASET

Due to the fact that until now there exists no mowing machine mounted fawn detector, the pictures were taken manually by a handheld infrared camera (E45 by FLIR) mounted on a stand with a height of 1,20m and a water-level to verify that the viewing direction of the camera has constantly a nadir angle of 25 degree. The used E45 has an uncooled microbolometer focal plane array with 160 x120 pixel and a lens with 25 deg. field-of-view. The rawdata was extracted from the radiometric jpeg for this dataset.

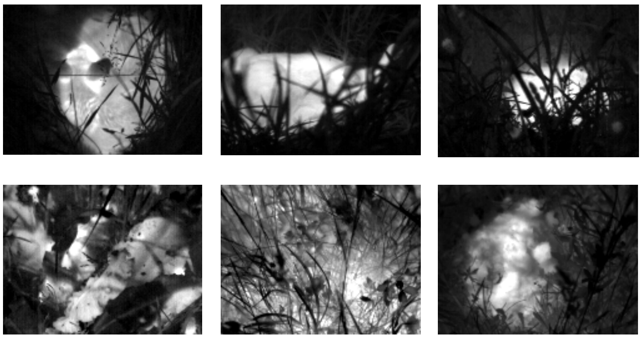

Fig. 1. Part of the dataset. On the upper pictures there are fawns; on the lower pictures there are no fawns.

The dataset consists of 103 pictures. 26 of them contain a fawn hidden in the grass, 74 of them contain meadow without fawns and 3 of them contain neither fawn nor meadow.

3. PARAMETER-FREE CLUSTERING

The Normalized Compression Distance (*NCD*) is a general similarity metric based on data compression, regarded as a way to estimate approximated complexities, proposed by Vitanyi et al.[4]. The *NCD* is based on the concept of algorithmic complexity: the Kolmogorov complexity of a string x, denoted with $K(x)$, is the length of the shortest program that outputs x and halts on an appropriate universal machine, such as a universal Turing Machine. On the basis of $K(x)$ the authors derive the similarity metric Normalized Information Distance (*NID*) as:

$$NID(x,y) = \frac{K(x, y) \text{ - } \min\ \{K(y), K(x)\}}{\max\ \{K(x), K(y)\}} \quad , (2)$$

with $K(x,y)$ being the complexity of *x* and *y* merged. Since $K(x)$ is not computable, a suitable approximation of the *NID* is obtained by estimating the complexity of each object *x* with the size *C*(*x*) of the object after being compressed by a standard compressor, as illustrated in Fig. 1.

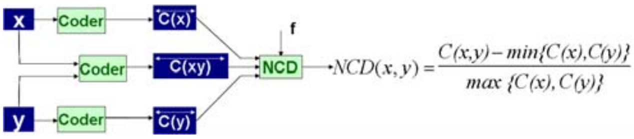
x
y
Coder
Coder
Coder
C(x)
C(xy)
C(y)
f
NCD
$NCD(x,y) = \frac{C(x,y) - min\{C(x),C(y)\}}{max\ \{C(x),C(y)\}}$

Fig. 2. ***NCD*** **schema. In this sketch, the coder represents a general lossless compressor (such as Gzip). The lengths of the compressed files *C*(*x*), *C*(*y*) and *C*(*xy*) are used with the function f to compute the information distance between two objects *x* and *y*, where *xy* represents the two objects merged in one. The basic idea is that if the two objects share common information, *C*(*xy*) will be smaller than *C*(*x*) + *C*(*y*).**

The indices are stored in a distance matrix that can be used for classification [4] by applying on it hierarchical clustering.

It has to be remarked that the approximation of *K*(*x*) with *C*(*x*) is data dependant: since current compressors are built based on different hypothesis, with some being more efficient than others on certain data types. Therefore, the dependence on the choice of the compressor is not a free parameter in itself, and for each dataset a compression algorithm able to fully exploit the redundancies in that kind of data should be adopted [5]: better compression, in fact, means better approximation of the Kolmogorov complexity.

Performance comparisons for general compression algorithms have shown that this dependence is generally loose [6], but increases when compressors for specific data types are adopted. Consider the case of images: an image consists of a number of independent observations, with each of those represented by a pixel value in the image grid. These observations constitute a stochastic process characterized by spatial relation inter pixels, since the value of each pixel is dependant not only on the previous one but also on the values of its neighbours, including the vertical and diagonal ones. Good results have been then achieved when using compression algorithms suited for images such as Jpeg2000 [7] in applications to satellite images [8]: compression with Jpeg2000 allows keeping the vertical spatial information contained within the images, exploiting it intrinsically within the computation of the information distance, whereas a general compressor, such as one belonging to the LZ family [9], is limited since it linearly scans the data and thus may fail at capturing the full information about the spatial distribution of the pixels. A good solution for the detection of fawns would then be to inject Jpeg2000 as a compressor in the NCD similarity measure, considering as *C*(*xy*) the size of the image obtained appending *x* and *y*, after being compressed in a lossless way by the Jpeg2000 algorithm.

4. RESULTS AND DISCUSSION

A totally unsupervised clustering experiment has been carried out to assess the discrimination power of the proposed method. The dataset used consists of 103 infrared images of size 160x120, of which roughly one third contains a fawn. In a first step, a distance matrix containing the similarity indices obtained with the NCD+JPG2000 similarity measure has been computed. The open-source utility Complearn [10] is then used to perform an unsupervised clustering, generating a dendrogram which fits (suboptimally) the distance matrix. Results in fig. 1 show that the two classes are well separated with only two “false alarms”. The patches containing fawns belong to a separate cluster, even when the animals are almost totally covered by vegetation: this confirms that the compression-based similarity measures introduced in the previous section are a powerful tool to discover similarities within images with a totally data-driven, model-free approach.

The final application should possibly employ a procedure similar to the one illustrated by Cilibrasi in[11]: here *n* objects are chosen as anchors to represent a set of classes in order to avoid the computation of a full distance matrix and to compute only *n* distances for a test object; afterwards the distance values are used to build a feature vector of *n* dimensions that can be used as an input for a Support Vector Machine[12] to perform classification.

5. ACKNOWLEDGEMENTS

Part of this research was carried out in the frame of the joint DLR-CNES-ENST Centre of Competence (CoC) for Information Extraction and Image Understanding for Earth Observation.

6. REFERENCES

[1] [Online] http://www.mstonline.de/foerderung/projektliste/detail_html?vb_nr=V3019

[2] P. Haschberger, M. Bundschuh, and V. Tank, “Infrared Sensor for the Detection and Protection of Wildlife”, *Optical Engineering*, 35: 882 – 889, 1996.

[3] *A. Patrovsky and E. Biebl, “Microwave sensors for detection of wild animals during pasture mowing”, Advances in Radio Science, 3, 211–217, 2005.*

[4] M. Li, X. Chen, X. Li, B. Ma and P.M.B. Vitányi, “The Similarity Metric”, *IEEE Trans Inf Theory*, 50|12:3250-3264, 2004.

[5] E. J. Keogh, S. Lonardi, C. Ratanamahatana, “Towards Parameter-Free Data Mining”, *SIGKDD 2004*.

[6] M. Cebrian, M. Alfonseca, and A. Ortega, “Common Pitfalls Using the Normalized Compression Distance: What to Watch Out for in a Compressor”, *Communications in Information and Systems*, vol.5, no. 4, pp. 367-384, 2005.

[7] Jpg2k for Java [Online] http://jj2000.epfl.ch/

[8] D. Cerra and M. Datcu, “Algorithmic Information Theory Based Analysis of Earth Observation Images: an Assessment”, *IEEE Geoscience and Remote Sensing Letters*, in press, 2009.

[9] J. Ziv and A. Lempel, “Compression of Individual Sequences Via Variable-Rate Coding”, *IEEE Trans. Inf. Theory*, 1978.

[10] Complearn tool by R. Cilibrasi, A. Cruz, S. de Rooij, and M. Keijzer, [Online], available at: http://www.complearn.org/index.html.

[11] R. Cilibrasi, "Statistical Inference Through Data Compression", *Lulu.com Press* , 2006.

[12] T. Joachims, B. Schölkopf, C. Burges, and A. Smola “Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning”, *MIT-Press*, 1999.

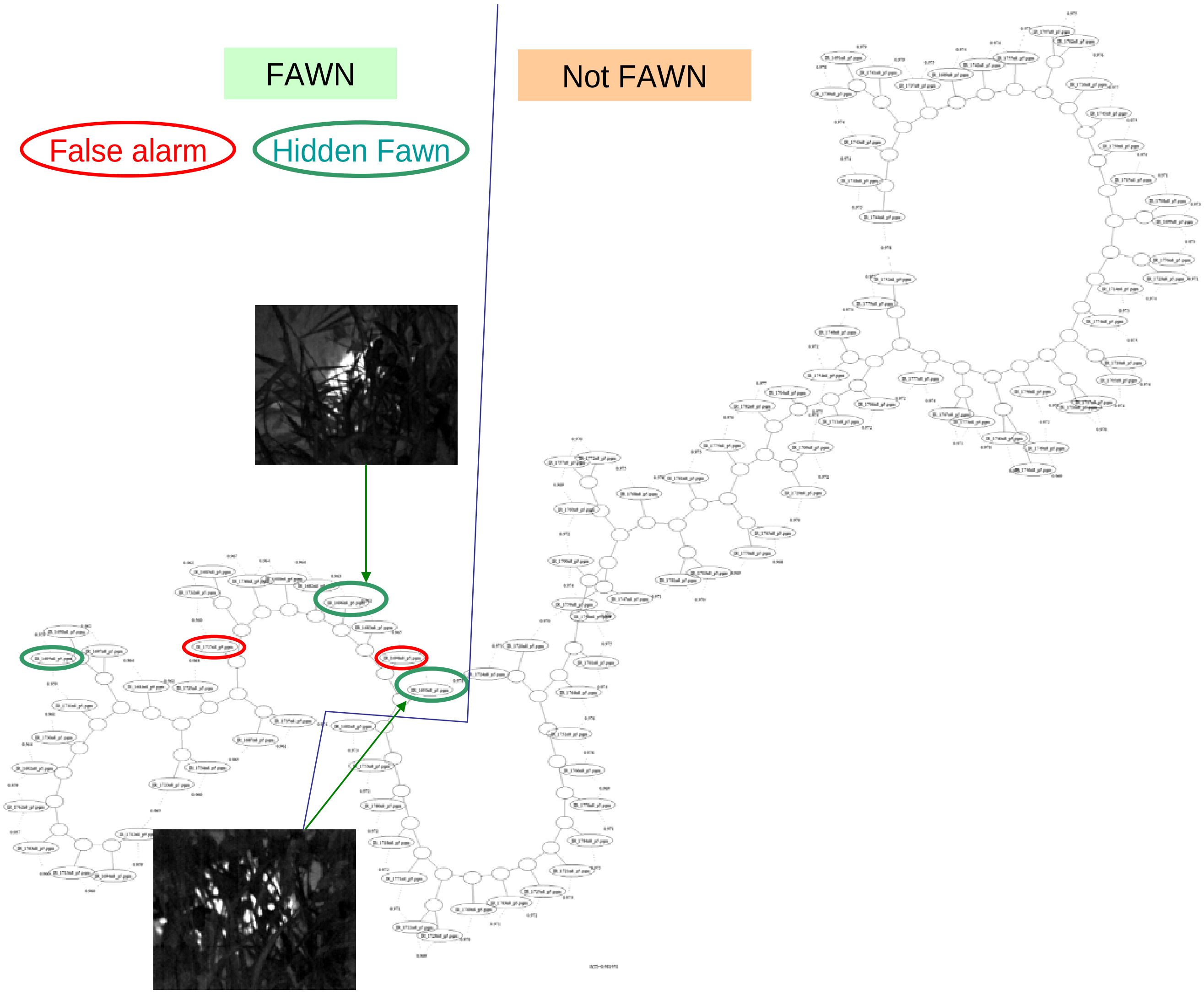

Fig. 3. Hhierarchical clustering of the similarity indices obtained with the NCD+JPG2000 similarity measure on a dataset of 103 infrared images of size 128x128. A line is drawn to separate the cluster of images containing a fawn. Two false alarms are circled in red. Fawns hidden behind the grass (circled in green), of which two samples are included, all lie inside the fawns cluster.