Feature Extraction Using Polynomial and Sigmoidal Kernels for Classification of Radar SAR Images

Cyrille Jean Enderli
Radar department
Thalès Airborne Systems
2, avenue Gay Lussac CN/648, 78851 Elancourt, France
cyrille-jean.enderli@fr.thalesgroup.com
tel. (+33)(0)134819973, fax. (+33)(0)134815281

Abstract: **This paper investigates the interest of nonlinear feature extraction for classification of Radar SAR images. It is shown that polynomial and sigmoidal filter models allow to significantly improve performances of standard classifiers.**

Key words: **Feature extraction, Classification, SAR images, Nonlinear Filtering**

I. INTRODUCTION

In many pattern recognition problems, a popular parametric method for feature extraction and classification is the Linear Discriminant Analysis (LDA) [1]. The purpose of LDA is to find an orthogonal transformation maximizing a criterion that measures class separability. Criteria used for LDA involve only the mean and covariance matrix of the data of each class. A common LDA solution is the projection upon the eigenvectors of the scatter matrix $\mathbf{S_w}^{-1}\mathbf{S_b}$. For C classes, $\mathbf{S_w}$ and $\mathbf{S_b}$ are defined as $\mathbf{S_w}=\Sigma_{i=1}^{C}\, P_i\mathbf{S_i}$, where $\mathbf{S_i}$ is the covariance matrix of class i and P_i is the class prior, and $\mathbf{S_b}=\Sigma_{i=1}^{C}\, P_i(\mathbf{M_i}-\mathbf{M})(\mathbf{M_i}-\mathbf{M})^T$ where $\mathbf{M_i}$ is the mean vector of class i and $\mathbf{M}=\Sigma_{i=1}^{C}\, P_i\mathbf{M_i}$. This LDA solution is known for yielding the optimal features with respect to the error probability when the data are Gaussian with equal covariance matrix [2]. In practice, the quantities $\mathbf{M_i}$ and $\mathbf{S_i}$ are generally estimated with the training data samples via the usual estimators: $\mathbf{M_i}=\Sigma_{j=1}^{ni}\, \mathbf{x_j^{(i)}}/n_i$ and $\mathbf{S_i}= \Sigma_{j=1}^{ni}\, (\mathbf{x_j^{(i)}})(\mathbf{x_j^{(i)}})^T/n_i - \mathbf{M_i}\mathbf{M_i}^T$ where $\mathbf{x_j^{(i)}}$ is the j-th sample vector of class i and n_i is the number of samples in class i. In this paper, the priors P_i are replaced by their estimates: $P_i \sim n_i/n$ where $n= \Sigma_{i=1}^{C}\, n_i$.

As estimated with a finite dataset, the matrix $\mathbf{S_w}$ is singular when the data size is greater than the number of data samples, which is the case in most applications [3]. Numerous solutions to overcome this problem have been proposed [3-10]. Among them, an efficient implementation of LDA can be obtained by first performing Principal Component Analysis (PCA) [1] of the data. LDA is then computed in the PCA transformed space. This method (PCA+LDA) has been shown to be equivalent to LDA for small datasets in Euclidian spaces, which may account for many successful applications [11].

Recently, nonlinear extensions of LDA (N-LDA) have received a great amount of interest for face and handwritten numerals recognition [4,12,13,14]. N-LDA methods can be described as LDA applied to the data that have been transformed in a nonlinear fashion. Very promising results have been obtained, but very few Radar applications seem to have been considered. In this paper, we describe an original application of N-LDA to Radar data identification, and show the interest of nonlinear filter models for SAR image identification.

This paper focuses on the use of Kernel Principal Component Analysis (K-PCA) [15] for extraction of features from Radar SAR images. Given a learning dataset of vectors in R^M, K-PCA can be described as PCA performed in a space where the data have been mapped to, according to some nonlinear projection model. Compared to the original data space, the mapped space is usually of very high – possibly infinite – dimension. Implementation of PCA in the mapped space is possible using the “kernel trick” [16] that allows computing scalar products of mapped vectors in terms of the original data, through the use of a kernel function. Extracted features are then passed through an LDA classifier and, for comparison, through a Nearest Neighbor (NN) classifier [1].

The use of K-PCA has been suggested as an efficient preprocessing step before performing LDA for face recognition applications [13,14]. It has been shown that K-PCA+LDA is equivalent to N-LDA for small datasets and interesting results have been obtained with a 2nd order polynomial kernel. We propose here an original application of K-PCA+LDA to Radar, for feature extraction and classification of SAR images. Furthermore, we compare 2 types of kernels, namely the polynomial and sigmoidal kernels. Results obtained show that increasing the order of the polynomial kernel up to 5 leads to an improvement of

0-7803-9582-4/06/$20.00 ©2006 IEEE

classification performance and then leads to the overfitting problem (i.e. loss of generalization ability). Similar results are obtained with the sigmoidal kernel as its parameter varies.

II. PCA OF NONLINEARLY MAPPED DATA

A. *Filtering models*

An important nonlinear filter model for signal processing is the Volterra model. It is a polynomial model used in various fields like optics [17], accoustics [18,19], image processing [20], telecommunications [21], signal detection/estimation [22], classification [23], system identification [24], or frequency domain filtering [25]. For an input vector $\mathbf{x}\in R^M$, the output of a d-th order Volterra filter is the output of a *linear filter applied to the d-th order Volterra vector*. Denoted by $\phi^{(d)}(\mathbf{x}) \in R^{Nd}$, the Volterra vector is defined as the vector of all the nonredundant multiproducts of **x**'s components, up to the d-th order. Note that $\phi^{(1)}(x)=x$. The 1st order Volterra model is thus the linear model. The size of $\phi^{(d)}(\mathbf{x})$ is N_d>>M (one shows that $N_d=O(M^d)$ where M is the size of **x** [26]). Therefore, Volterra vectors cannot actually be computed in most applications when $d\geq 2$. However, a dot product of Volterra vectors is closely related to the dot product of the corresponding original data vectors through the following polynomial kernel function $k^{(d)}$ [27]:

$$\phi^{(d)}(\mathbf{x})^T\phi^{(d)}(\mathbf{y}) = k^{(d)}(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y})+\ldots+(\mathbf{x}^T\mathbf{y})^d \quad (1)$$

The kernel of (1) allows one to obtain the dot product of Volterra vectors without explicitly computing them. It may not be the optimal model with respect to the classification error in the Radar application considered in this paper. For comparison, we will also consider the sigmoidal kernel, corresponding to another filter model. This kernel has parameter κ and is defined as $k^{(\kappa)}(\mathbf{x},\mathbf{y})$:

$$k^{(\kappa)}(\mathbf{x},\mathbf{y})=\tanh(\kappa\mathbf{x}^T\mathbf{y}) \quad (2)$$

The sigmoidal kernel is often cited for Support Vector Machine (SVM) classification [16] and is sometimes used for recognition tasks [28]. Many other kernels may also be used, provided that they satisfy the Mercer condition [16].

B. *Implementation of K-PCA*

1) Eigenanalysis of training mapped samples: Let us consider the arrays $\mathbf{X_1},\ldots,\mathbf{X_c}$ of very high dimensional mapped data samples of C classes. $\mathbf{X_i}$ is the array of the original data samples of class i that have been mapped according to some model with associated kernel k(.,.). For example, $\mathbf{X_i}$ can be the array of the Volterra vectors of class i. By definition, the j-th column of $\mathbf{X_i}$ is the j-th mapped vector of class i, with i=1…C and j=1…n_i. Let us denote by N the size of the mapped vectors. $\mathbf{X_i}$ is thus of size (N,n_i), n_i being the number of samples of class i. The complete mapped dataset is denoted by $\mathbf{X}=[\mathbf{X_1},\ldots,\mathbf{X_c}]$ and is of size (N,n) where $n=\Sigma_{i=1}^{C}\, n_i$ <<N. The PCA of **X** consists in computing the eigenvectors corresponding to the largest eigenvalues of the following matrix **G**:

$$\mathbf{G}=\mathbf{XQQX}^T / n \quad \text{where} \quad \mathbf{Q}=\mathbf{I_n}\text{-}\mathbf{J_n}/n \quad (3)$$

In (3), $\mathbf{I_n}$ is the n-dimensional identity matrix and $\mathbf{J_n}$ is full of ones. The purpose of **Q** is to center **X** (i.e. removing its mean vector) by right side multiplication. The size of **G** is (N,N) and as N is very large, direct computation of **G**'s eigenelements is intractable. However, **G** and the following matrix **A** share the same nonzero eigenvalues [1]:

$$\mathbf{A}=\mathbf{Q}^T\mathbf{X}^T\mathbf{XQ} / n \quad (4)$$

Furthermore, if **Λ** and **Θ** are the diagonal eigenvalue matrix and the corresponding eigenvector matrix of **A**, then the eigenvector matrix of **G** corresponding to its nonzero eigenvalues is given by [1]:

$$\boxed{\mathbf{\Psi}=\mathbf{XB}} \quad \text{where} \quad \boxed{\mathbf{B}=\mathbf{\Theta\Lambda}^{-1/2}/\sqrt{n}} \quad (5)$$

As a conclusion, the relevant eigenelements of **G** (**Ψ** and **Λ**) can be deduced from the eigenelements of **A** (**Θ** and **Λ**). It should be noted that in the expression of **A** (4), each entry of $\mathbf{X}^T\mathbf{X}$ is a a dot product that can be computed by using the kernel k(.,.) associated with the data projection model considered. If the Volterra model is considered, each entry of $\mathbf{X}^T\mathbf{X}$ is a dot product of Volterra vectors, which can be computed by using the kernel $k^{(d)}$ of (1) without explicitly computing the Volterra vectors. The sigmoidal kernel (2) can be used instead of $k^{(d)}$. Also, since **A** is of size (n,n), where n is the total number of training samples, computation of **A**'s eigenelements is fast when a small number of sample vectors are available in the database.

2) Feature extraction of a new mapped sample: Feature extraction of a new mapped data sample $\mathbf{z}\in R^N$ is achieved by the projection of **z** in the columns of **Ψ** (5) corresponding to the m largest eigenvalues of **G** (3), with m<n:

$$\mathbf{z'}=\mathbf{\Psi_m}^T\mathbf{z} \quad (6)$$

where $\mathbf{\Psi_m}$ is the m columns of **Ψ** corresponding to the m largest eigenvalues of **G**. Let us assume that $\mathbf{\Psi_m}$ is the first m columns of **Ψ**. Reporting (5) in (6) yields $\mathbf{z'}=\mathbf{B_m}^T\mathbf{X}^T\mathbf{z}$ where $\mathbf{B_m}$ is the first m columns of **B** (5). In this last expression of **z'**, the entries of the vector $\mathbf{X}^T\mathbf{z}$ are dot products of mapped data samples. Therefore, the kernel associated with the considered model allows to compute the features **z'** of (6) without actually mapping the original data samples.

3) Number of extracted features: The number m of retained eigenvalues of **G** may have great influence on classification performance. In the following, m will be set to a fraction of the eigenvalues' energy E. E is defined as the sum of **G**'s eigenvalues, and m is set to a value such that

$$\Sigma_{i=1}^{m} \lambda_i = pE \qquad (7)$$

where λ_i are **G**'s eigenvalues and p varies from 0.1 to 0.99. The value for p (and m) is selected such that classification performance is maximum.

III. Classification of SAR images

In this section, features are extracted from Radar SAR images through K-PCA. An LDA classifier and a Nearest Neighbor (NN) classifier are then applied to the extracted features. The LDA classifier used here works as follows

- (Training Stage) Estimate the mean vectors $\mathbf{M_i}$ of the features extracted from the training samples of each class. Estimate the (assumed) common covariance matrix $\mathbf{\Gamma}$ of the pooled features of all the classes.
- (Test Stage) Given a test vector **z'** of features, assign the class with highest posterior. Bayes' rule and Gaussian assumption yield the class such that $(\mathbf{z'}-\mathbf{M_i})^T\mathbf{\Gamma}^{-1}(\mathbf{z'}-\mathbf{M_i}) - \ln(P_i)$ is minimum [1]. Recall that P_i is the prior of class i. It has been defined as $P_i=n_i/n$ where n_i is the number of samples of class i and n is the total number of samples. Since the quadratic part of the decision rule $(\mathbf{z'}^T\mathbf{\Gamma}^{-1}\mathbf{z'})$ is independent of the class label i, the decision rule simplifies in a linear expression: assign to **z'** the class such that $\mathbf{m_i}^T\mathbf{\Gamma}^{-1}(2\mathbf{z'}-\mathbf{m_i}) + \ln(P_i)$ is maximum.

The NN classifier computes the Euclidian distances between **z'** and all the feature vectors extracted from the training data samples. The class assigned to **z'** is the class corresponding to the feature vector that has minimum distance to **z'**. Classification performances of both LDA and NN methods is evaluated by the mean classification error ε:

$$\varepsilon = \Sigma_{i=1}^{C} P_i\, P[\neg C_i|C_i] \qquad (8)$$

where C is the number of classes and $P[\neg C_i|C_i]$ stands for the probability of error for class i. This latter probability is estimated by the number of misclassified test samples of class i, divided by the number of test samples of class i.

A. *Description of the training and testing datasets*

In this section, data samples are SAR images of 10 classes from the MSTAR database [29]. Resolution is 30cm along both range and azimuth axes. The Radar grazing angle and the target aspect angle characterize each target image. Training images are obtained at 15° grazing angle and test images are obtained at 17° grazing angle.

The classifier is trained over a limited range of aspect angles, referred to as the "training sector". If the training sector were too small, there would not be enough data samples to describe the classes accurately. Preliminary studies have shown that training sectors of 60° are good solutions to have enough training samples with limited variations due to aspect angle. The training sector used in our tests is [0,60], grouped with [180,240] because of the front/back ambiguity of a target image.

The "testing sectors" are defined as the ranges of aspect angle of the test images. The testing sector used for classification performance evaluation here is [0,45] grouped with [180,225]. Table 1 shows the classes' names and types, and the number of training and test samples.

TABLE I. Description of the datasets

Class	Name/type	# train / test samples
C_1	2s1 / gun	100 / 82
C_2	Bmp2 / char	171 / 166
C_3	Brdm2 / truck	104 / 82
C_4	Btr60 / transport	67 / 59
C_5	Btr70 / transport	72 / 60
C_6	D7 / bulldozer	75 / 56
C_7	T62 / char	95 / 79
C_8	T72 / char	171 / 162
C_9	Zil131 / truck	99 / 81
C_{10}	Zsu234 / gun	102 / 82

B. *Preprocessing of the data*

In each train or test image, the Radar reverberation of the ground (i.e. the "clutter") is removed using a CFAR processor (Fig.1b) [30].

The image is then rotated as illustrated in Fig. 1c. The aim of image rotation is to ensure that for different aspect angles, the same target pixel of the rotated image corresponds to the same physical area of the target. Given the target aspect angle, α, the image is rotated by an angle of 180-α (if α<180) or -α (if α>180). The rotation centre is the centre of the image.

After rotation , the pixels corresponding to the target are the data used for classification (Fig.1d). All the target images are added rows and columns of zeros to the right and bottom sides, in order to fit in a common bounding box. The common bounding box is here of size 36 x 24 pixels. An image is reshaped to a vector by stacking its columns. The data vectors are of length M = 36 x 24 = 864.

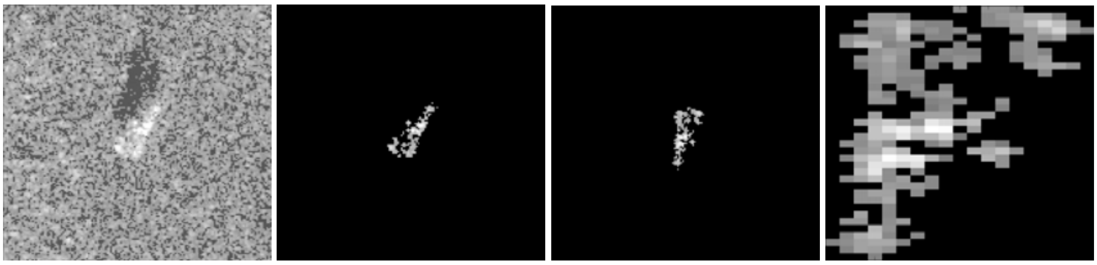

Figure 1. From left to right: (a) Original SAR image; (b) Ground clutter suppression; (c) Image rotation; (d) Data used for target recognition

C. *Classification results*

The feature extraction method through K-PCA (section II) has been applied to the preprocessed SAR images with the polynomial (1) and sigmoidal (2) kernels. Influence of the parameters d and κ of the kernels have been investigated for d=1,2,…15 and κ=0.1,0.2,…,7. For a fixed value of a kernel parameter, recall that the number of retained features with K-PCA corresponds to the proportion of the eigenvalues' energy leading to the least classification error (8). The minimum classification error is reported on fig.2 for both LDA and NN classification methods as the parameters of the kernels vary.

1) Results with the polynomial kernel: Fig.2a shows the performances obtained with the polynomial kernel as a function of the kernel parameter d. For LDA classification, the overall minimum error is reached for d=5 and equals 5.96%. In this case, the number of extracted features is such that 95% of the eigenvalues' energy (7) is retained. K-PCA improves performances up to d=5. In particular, the linear model (d=1) yields an error of 14.5% which shows the interest of nonlinear feature extraction. When d>5, the LDA classification performances decrease. Recall that the test data are images obtained at a different grazing angle than the training data. If one uses the same data for both training and testing, classification error remains 0% for d>5. Therefore, setting d>5 leads to the overfitting phenomenon (i.e. loss of generalization ability).

For NN classification, the minimum error is reached for the linear model d=1 and equals 6.05%. The number of extracted feature through PCA is such that 70% of the eigenvalues' energy (7) is retained. In this case, polynomial feature extraction is not relevant. However, NN classification of the original data (without PCA feature extraction) leads to classification error of 12.5%, which shows that feature extraction can improve NN classification performances.

2) Results with the sigmoidal kernel: Fig.2b shows the classification error as a function of the sigmoidal kernel parameter κ. For LDA classification, the overall minimum classification error is reached for κ=2.0 and equals 6.2%. The number of extracted features is such that 90% of the eigenvalues' energy (7) is retained. For NN classification, minimum error is 5.55%, reached for κ=2.2 and for a number of extracted features such that 75% of the eigenvalues' energy (7) is retained. Both curves of classification error versus κ have a minimum around κ=2. Influence of κ seems to be less important with the NN classifier because performances are quite stable for κ∈]0,3]. On the other hand, κ may have great influence on LDA clasification performances. When the value of κ increases above 3, the performances of both classifier are degraded and tend to be the same.

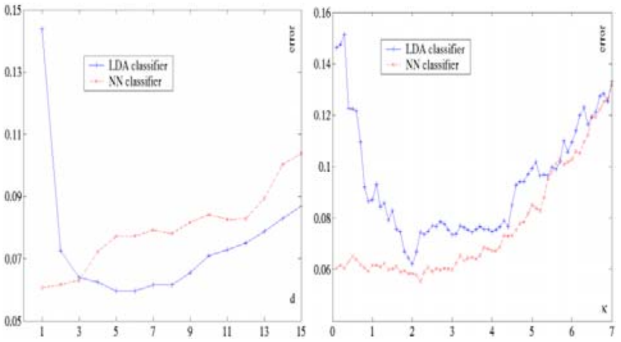

Figure 2. LDA and NN classification error as a function of the (a) left: polynomial and (b) right: sigmoidal kernel parameter.

IV. Conclusion

We have described a nonlinear feature extraction method and successfully applied it to classification of Radar SAR images. An advantage of the method is that it can be efficiently implemented with few training samples in the database through the use of kernel functions. Two nonlinear models (the Volterra polynomial and sigmoidal models) have been used in conjunction with two standard classification techniques (LDA and NN). Results obtained have shown that nonlinear feature extraction allows to significantly improve the classification performances. The least LDA classification error was 5.96% and was reached with the 5th order polynomial model, whereas in the NN case, the least error was 5.55% for the sigmoidal model with parameter κ=2.2. These results show relevance of both nonlinear models depending on which classification method is used.

Acknowledgment

The author would like to thank Mr Philippe Réfrégier from Institut Fresnel in Marseille, for his various advice on filtering and classification techniques, and Mr Laurent Savy from Thales Airborne Systems, for enlightening discussions on SAR image classification.

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed., New ork: Academic Press, 1990.

[2] K. Demuynck, J. Duchateau, D. Van Compernolle, "Optimal feature sub-space selection based on discriminant analysis", technical report D023, Katholik Universiteit Leuven, Belgium, 1999.

[3] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, "Face recognition using LDA based algorithms", IEEE Trans. Neural networks, vol. 14(1), 2003.

[4] W. Liu, Y. Wang, S.Z. Li, T. Tan, "Null space-based kernel Fisher discriminant analysis for face recognition", in proc. 6th IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, May 2004.

[5] C.E. Thomaz, D.F. Gillies, "A maximum uncertainty LDA-based approach for limited sample size problem – with application to face recognition", Technical Report, Imperial College London – Dpt. of computing, 2004.

[6] L. Chen et al., "A new LDA-based face recognition system that can solve the small sample size problem", Pattern Recognition, vol. 33(10), pp. 1713-1726, 2000.

[7] W. Zhao, R. Chellappa, P.J. Philllips, "Subspace linear discriminant analysis for face recognition", Technical Report CS-TR4009, University of Maryland, 1999.

[8] K. Etemad, R. Chellappa, "Discriminant analysis for recognition of human face images", J. Opt. Soc. Am. A, vol 14(8), 1997.

[9] P.N. Belhumeur et al., "Eigenfaces vs. Fisherfaces: Recognition using class specific projection", IEEE Trans. Patt. Anal. Machine Intell., vol. 19(7), pp. 711-720, 1997.

[10] S. Fidler, A. Leonardis, "Robust LDA classification", in IEEE Workshop on Statistical Analysis in Computer Vision, 2003.

[11] J. Yang, J.-Y. Yang, "Why can LDA be performed in PCA transformed space?", Pattern Recognition, vol. 34, pp. 2067-2070, 2001.

[12] C. Park, H. Park, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition", SIAM Journal on Matrix Analysis and Applications, 2005.

[13] J. Yang, A.F. Frangi, J.Y. Yang, "A new kernel fisher discriminant algorithm with application to face recognition", Neurocomputing Letters, vol. 56, pp. 415-421, 2004.

[14] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, "KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition," IEEE Trans. Patt. Anal. Machine Intell., vol. 27, pp. 230-244, February 2005.

[15] B. Schölkopf, A. Smola and K.R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol. 10, pp. 1299-1319, 1998.

[16] C.J.C. Burges, "A tutorial on support vector machines," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.

[17] B.E.A. Saleh, "Bilinear transformations in optical signal processing", in proc. SPIE, vol. 373, 1981.

[18] R. Niemistö, T. Mäkelä, "Robust adaptive polynomial filters for accoustic echo cancellation", in 5th Nordic Signal Processing Symposium, Hurtigruten (Norway), 2002.

[19] J.F. Michaelides, P. Kabal, "Nonlinear adaptive filtering for echo cancellation", in Proc. IEEE International Conference on Communications, pp. 30.3.1-30.3.6, Philadelphia, June 1988.

[20] R. Bernstein, M. Moore, S. Mitra, "Adjustable quadratic filters for image enhancement", in proc. IEEE international conference on image processing, 1997.

[21] A. Zhu, T.J. Brazil, "An adaptive Volterra predistorter for the linearization of RF high power amplifiers", in Proc. IEEE IMS'02, pp.461-464, 2002.

[22] B. Picinbono, P. Duvaut, "Optimal linear-quadratic systems for detection and estimation", IEEE Trans. Inform. Theory, vol. 34(2), March 1988.

[23] C. Enderli, L. Savy, P. Réfrégier, "Application of the deflection criterion to classification of radar SAR images", IEEE Trans. Patt. Anal. Machine Intell., in press.

[24] T. Koh, E.J. Powers, "Second-order Volterra filtering and its application to nonlinear system identification", IEEE Trans. Accoustics, Speech and Signal Processing, vol. ASSP-33(6), 1985.

[25] S. Taylor, N. Haritos, K. Thiagarajan, "Optimal sample length for calculating transfer functions from discrete experimental data", ANZIAM, vol. 46, pp. C.572-C.587, 2005.

[26] R.D. Nowak, B.D. Van Veen, "Tensor product basis approximation for Volterra filters", IEEE Trans. Signal Processing, 1996.

[27] C.J. Enderli, "Classification par filtrage de Volterra optimal pour la deflexion. Application a l'identification de donnees Radar.", Ph.D. thesis, university of Marseille – France, n° 2006AIX30016, 2006.

[28] C.J. Bellman, M.R. Shortis, "A machine learning approach to building recognition in aerial photographs", International Archives of the Photogrammetry, Remote Sensing, Spatial and Information Sciences, Graz (Austria), 2002.

[29] Moving Stationary Targets Acquisition and Recognition, https://www.mbvlab.wpafb.af.mil/public/datasets/mstar/

[30] P.P. Ghandi, S.A. Kassam, "Analysis of CFAR processors in nonhomogeneous background", IEEE Trans. Aerosp. Electron. Systems, vol. 24(4), pp. 427-445, 1988.