Intrinsic prior for Bayesian classification of texture images

Aurélien Schutz, Lionel Bombrun and Yannick Berthoumieu

Université de Bordeaux, Institut Polytechnique de Bordeaux, Laboratoire IMS, Groupe Signal et Image
{*aurelien.schutz, lionel.bombrun, yannick.berthoumieu* }*@ims-bordeaux.fr*

ABSTRACT

This paper introduces an intrinsic prior distribution for supervised classification of texture images. First, we introduce the intrinsic prior distribution as the normal law on a Riemannian manifold. Next, based on this definition, we derive the estimation and classification schemes. Finally, we propose an application for the classification of texture images. Experiments on the VisTex texture database are conducted and demonstrate the interest of the proposed intrinsic classification algorithm.

Index Terms— Information geometry, intrinsic prior, texture classification.

1. INTRODUCTION

In the framework of natural texture image processing, parametric stochastic models have been well studied. When parametric stochastic modeling is associated with scale-space representation, very effective algorithms can be designed, obtaining successful results for a large class of natural textures. Providing an unified view of model estimation, classification and synthesis, multiscale scheme enables us to prototype stochastic models which have a non-Gaussian distribution.

However, based on scale-space decomposition and stochastic modeling, none of the existing works have addressed the problem of the intra-class diversity characterization to increase the classification performance. In this paper, we employ the Bayesian framework to take into account the intra-class diversity. From the concept of intrinsic prior, we develop a new supervised parametric classification algorithm.

Parametric classification is closely linked with the theory of statistical manifolds, which aims at providing a Riemannian structure to the parameters space of probability density functions (pdf). Many works have proposed extrinsic tools by embedding the manifold in an Euclidean space [1]. Nevertheless, to compare two observations in a parameter space, the geometry of the Riemannian manifold should be considered which is the main objective of information geometry theory [2]. In this context, the notion of intrinsic tools is the key point. For example, in an estimation problem, the estimate should be invariant under any reparametrization of the parameter space. This notion of intrinsic has hence been considered in a wide range of disciplines such as in the definition of intrinsic loss function [3], intrinsic discrepancy in the context of Bayesian estimation [4] and detection [5], and also in intrinsic version of Cramér-Rao bound [6]. In this paper, our contributions are threefold. First, in the Bayesian framework, we introduce the concept of intrinsic prior distribution as the normal law on a Riemannian manifold when the Jeffrey divergence is considered and derive an intrinsic estimation and classification scheme. Second, based on the intrinsic prior, we show that the optimal decision characterizing the classification procedure leads to a decision directly on the parameter space (Riemannian manifold). Third, we propose to validate the proposed methodology in a texture based image retrieval experiments.

The paper is structured as follows. Section 2 gives a state-of-the-art about intrinsic distribution and introduces the proposed intrinsic prior. Based on this definition, Section 3 derives the estimation and classification scheme. Section 4 introduces an application for the classification of texture images, and some experiments results are displayed on the VisTex database. Conclusions and future works are finally reported in Section 5.

2. INTRINSIC PRIOR DISTRIBUTION

Let $\chi = (\chi_1, \ldots, \chi_K)$ be K independent sets of N_i independent and identically distributed random variables (vectors) $\mathbf{x}$ according to a parametric model $p(\mathbf{x}|\theta)$. Let $(\hat{\theta}_1, \ldots, \hat{\theta}_K) \in \Theta$ be the K maximum likelihood estimates computed on these sets $(\chi_1, \ldots, \chi_K)$. This collection of K parametric vectors can be described by its first and second order characteristics: the most central element (*i.e.* the centroid $\bar{\theta}$) and the standard deviation σ around this central point. Based on the definition of the entropic prior and the normal law on a Riemannian manifold, we introduce the notion of intrinsic prior distribution $p(\theta|\bar{\theta}, \sigma)$.

2.1. Entropic prior

Given a parametric model $p(\cdot|\theta)$, the entropic prior on θ is given by [7, 8]:

$$p\left(\theta|\theta_0, \alpha\right) \propto \frac{1}{|G(\theta)|^{-\frac{1}{2}}} \exp\left\{-\alpha I(p(\cdot|\theta_0), p(\cdot|\theta))\right\} \quad (1)$$

where α is a positive scalar parameter, $|G(\theta)|$ is the determinant of the Fisher information matrix computed at point θ

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

and $I(p(\cdot|\theta_0), p(\cdot|\theta))$ is a divergence between the probability measures $p(\cdot|\theta_0)$ and $p(\cdot|\theta)$.

As observed, this entropic prior is invariant under any change of coordinate in the parameter space. This prior is hence intrinsic and assigns to θ a probability which decreases exponentially with the divergence. In (1), the parameter α controls the sensitivity to changes in distance. Note that when α tends toward 0, the entropic prior reduces to the non-informative Jeffrey prior.

2.2. Normal law on a manifold

In [9], Pennec introduces the concept of normal law on a Riemannian manifold $\mathcal{M}$ knowing the mean value $\bar{\theta}$ and covariance matrix Σ as:

$$p(\theta|\bar{\theta}, \Gamma) = k \exp\left(-\frac{\overrightarrow{\bar{\theta}\theta}^T \Gamma \overrightarrow{\bar{\theta}\theta}}{2}\right) \tag{2}$$

where k is the normalizing constant and Γ the concentration matrix linked to the covariance matrix by:

$$\Sigma = k \int_{\mathcal{M}} \overrightarrow{\bar{\theta}\theta}\, \overrightarrow{\bar{\theta}\theta}^T \exp\left(-\frac{\overrightarrow{\bar{\theta}\theta}^T \Gamma \overrightarrow{\bar{\theta}\theta}}{2}\right) \mathrm{d}\mathcal{M}(\theta) \tag{3}$$

The model defined in (2) approximates the normal model by the usual Gaussian distribution in the tangent space $\mathrm{T}_{\bar{\theta}}\mathcal{M}$ at the mean value $\bar{\theta}$. In this definition, the projection from the Riemannian manifold to the tangent space is given by the exponential map. Here, the geodesic distance (GD) induced by the Riemannian metric, derived from the Fisher information matrix is considered to compute the proximity between two observations, *i.e.* $\mathrm{GD}(p(\cdot|\bar{\theta}), p(\cdot|\theta)) = \|\overrightarrow{\bar{\theta}\theta}\|$.

This model has notably been considered in [10] for the segmentation of magnetic resonance images.

2.3. Intrinsic prior

Inspired from the definitions of the entropic prior (1) and the normal law on a Riemannian manifold (2), we introduce the definition of the intrinsic prior as the normal law on a manifold when the Jeffrey divergence is considered instead of the Riemannian metric as:

$$p\left(\theta|\bar{\theta}, \sigma\right) = \frac{1}{(2\pi)^{d/2}\sigma^d |G(\theta)|^{-\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2} J(p(\cdot|\theta), p(\cdot|\bar{\theta}))\right\} \tag{4}$$

where d is the dimension of the parameter space Θ and $J(p(\cdot|\theta), p(\cdot|\bar{\theta}))$ is the Jeffrey divergence computed between the parametric models $p(\cdot|\theta)$ and $p(\cdot|\bar{\theta})$.

3. INTRINSIC ESTIMATION AND CLASSIFICATION

3.1. Intrinsic estimation

By considering the notation introduced in Section 2, the maximum likelihood estimators of the centroid $\bar{\theta}$ and standard deviation σ are obtained as solution of:

$$\hat{\lambda} = \arg\max_{\lambda} p(\chi|\lambda), \tag{5}$$

where $\lambda = \{\bar{\theta}, \sigma\}$ is the set of hyperparameters, and

$$p(\chi|\lambda) = \prod_{i=1}^{K} p(\chi_i|\lambda) = \prod_{i=1}^{K} \int_{\Theta} p(\chi_i|\theta_i)\, p(\theta_i|\lambda)\, \mathrm{d}\mathcal{M}(\theta_i), \tag{6}$$

where $\mathrm{d}\mathcal{M}(\theta_i) = |G(\theta_i)|^{\frac{1}{2}} \mathrm{d}\theta_i$ is the volume element on a Riemannian manifold. In (6), the main difficulty relies on the computation of the integral. Many works have been dedicated to this problem including Monte Carlo integration, variational based approaches or Laplace approximation [11, 12]. This latter has been successfully validated by many authors in applied mathematics (see [13] for instance). Next, after developping $p(\chi_i|\theta_i)$ as $\prod_{j=1}^{N_i} p(x_j|\theta_i)$ and introducing the Laplace approximation and the expression of the proposed intrinsic prior (4) in (6), one can derive the maximum likelihood estimates of the centroid $\hat{\bar{\theta}}$ and standard deviation $\hat{\sigma}$. After some cumbersome computations and removing the terms independent of $\bar{\theta}$ and σ, it yields:

$$\hat{\bar{\theta}} = \arg\min_{\bar{\theta}} \frac{1}{K} \sum_{i=1}^{K} J(p(\cdot|\hat{\theta}_i), p(\cdot|\bar{\theta})) \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{dK} \sum_{i=1}^{K} J(p(\cdot|\hat{\theta}_i), p(\cdot|\hat{\bar{\theta}})) \tag{8}$$

First, it can be noticed that the maximum likelihood estimators of $\bar{\theta}$ and σ do not depend on the set χ. They are directly expressed as a function of the maximum likelihood estimators $\hat{\theta}_i$. Note also that, equations (7) and (8) coincide with the maximum likelihood estimators of $p(\theta|\lambda)$ where $\theta = \{\hat{\theta}_1, \ldots, \hat{\theta}_K\}$.

To estimate $\hat{\bar{\theta}}$ from (7), a stochastic gradient descent algorithm can be considered [2, 14, 15]. $\hat{\bar{\theta}}$ is obtained as the fixed point solution of

$$\theta_{t+1} = \theta_t - \eta_t C(\theta_t) \nabla l(\theta_t) \tag{9}$$

where η_t is the step size which may depend on iteration t, $C(\theta_t)$ is a positive definite matrix and $\nabla l(\theta_t)$ is the gradient of the cost function defined in (7). When $C(\theta_t)$ is equal to $G^{-1}(\theta_t)$, the inverse of the Fisher information matrix, (9) corresponds to the natural gradient descent. This latter is intrinsic since it does not depend on the chosen parametrization θ of the pdf [2, 14, 15].

3.2. Classification

Let χ_t be a set of N_t independent and identically distributed random vectors $\mathbf{x}$. Let $\hat{\theta}_t$ be the maximum likelihood estimate computed on χ_t. Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_C$ be a collection of C hyperparameters, corresponding to C classes, estimated according to (7) and (8). The sample χ_t is classified to the class

c maximizing the likelihood $p(\chi_t|\lambda_c)$, i.e.

$$\hat{c} = \arg\max_c p(\chi_t|\hat{\lambda}_c) \tag{10}$$

By following the same procedure as described in Section. 3.1 for the estimation process, one can rewrite the expression of the decision rule (10) as:

$$\hat{c} = \arg\min_c d\ln\hat{\sigma}_c + \frac{1}{2\hat{\sigma}_c^2} J(p(\cdot|\hat{\theta}_t), p(\cdot|\hat{\bar{\theta}}_c)) \tag{11}$$

Under the homoscedasticity assumption, $\forall\{i \in 1,\ldots,c\}, \hat{\sigma}_i = \hat{\sigma}$, (11) reduces to the decision rule

$$\hat{c} = \arg\min_c J(p(\cdot|\hat{\theta}_t), p(\cdot|\hat{\bar{\theta}})). \tag{12}$$

Equations (12) and (11) can respectively be interpreted as linear (resp. quadratic) discriminant analysis on a Riemannian manifold. Note that those decision rules are decisions on the parameter space Θ (not on the original space χ), hence reducing the computational complexity.

In the next section, we propose an application of this intrinsic classification scheme for the recognition of texture images.

4. INTRINSIC TEXTURE CLASSIFICATION

4.1. Context

Many works in texture image recognition have shown that the wavelet representation is a well-adapted domain to characterize the texture, yielding to a multiscale analysis scheme which consists in modeling each wavelet subband. Let I be a texture image. Let N_o and N_s be respectively the number of orientation and scale of a multi-scale decomposition. I is hence decomposed into $N_o \times N_s$ sub-bands. Let us consider the parametric vector $\theta_{s,o}$ of the pdf associated to each sub-band. The collection T_I of those parametric vectors will represent the texture image I.

$$T_I = \{\theta_{s,o}|s = 1,\ldots,N_s, o = 1,\ldots,N_o\}. \tag{13}$$

4.2. Intrinsic texture estimation

Let $(T_{c,1}, \ldots, T_{c,N_{Tr}})$ be N_{Tr} training samples from the same class c. From this collection of samples, the class c is represented by a collection of centroids $\bar{\theta}_{c,s,o}$ and standard deviation $\sigma_{c,s,o}$ computed on each subband, since the subbands of the wavelet decomposition are assumed to be independent. Here, the subscripts c, s, o refer respectively to the texture class c, the scale s and the orientation o of the wavelet subband. Hence, for each wavelet subband, one centroid $\bar{\theta}_{c,s,o}$ and standard deviation $\sigma_{c,s,o}$ are estimated according to the intrinsic estimation scheme developed in Section 3.1, see equations (7) and (8). As observed in Fig. 1, the natural intra-class diversity of texture images is captured by the proposed normal law on a Riemannian manifold (intrinsic prior).

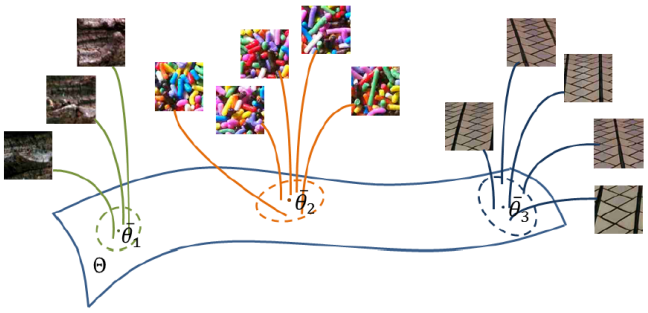

Fig. 1. Representation of the intra-class diversity modeled by the proposed intrinsic prior distribution.

In this paper, we propose an application to texture image recognition based on a multivariate modeling context. The wavelet coefficients located around the neighborhood $p \times q$ of the current spatial position are clustered in the random vector $\mathbf{x}$. The realizations of vector $\mathbf{x}$ characterize the spatial dependency of wavelet coefficients. Spherically Invariant Random Vectors are a class of stochastic models which have shown promising results for modeling the spatial dependency of wavelet coefficients [16]. Hence, we propose to model those observations based on the SIRV representation. Let $\mathbf{x}$ be a $pq-$dimensional vector following a SIRV distribution, it yields that $\mathbf{x}$ admits the stochastic representation $\mathbf{x} = \sqrt{\tau}\mathbf{g}$ where τ is a scalar random variable called multiplier ($\tau \in \mathbb{R}^+$) and $\mathbf{g}$ a real Gaussian vector with zero mean and covariance matrix $\Sigma = \mathbb{E}\{\mathbf{g}\mathbf{g}^T\}$. By exploiting the independence of the processes τ and $\mathbf{g}$ and by working on the joint vector $\mathbf{y} = (\tau, \mathbf{g})$, the Jeffrey divergence of the joint model can be expressed as the sum of the Jeffrey divergence for the multivariate Gaussian process and the Jeffrey divergence for the multiplier part. Note that both terms admit a closed-form expression recalled in [17]. It yields that the centroid for a SIRV model $\mathbf{y}$ is composed by two centroids: one for the Gaussian part and one for the multiplier part. For more information dealing with the implementation of those centroids estimators, the interested reader is referred to [17].

4.3. Texture classification

Let T_t be a test image. According to the classification rule presented in Section 3.2, this image is labeled to the class $\hat{c}$, corresponding to the class maximizing the likelihood $p(T_t|\lambda_{\hat{c}})$. Since the subbands of the wavelet decomposition are independent, one can consider the chain rule principle to obtain the pdf of $p(T_t)$ as the product of the pdf computed for each subbands. After some computations, the decision rule is simply the sum of the decision rules (11) computed on each subband. It yields

$$\hat{c} = \arg\min_c \sum_{s,o} d\ln\hat{\sigma}_{c,s,o} + \frac{1}{2\hat{\sigma}_{c,s,o}^2} J(p(\cdot|\hat{\theta}_t), p(\cdot|\hat{\bar{\theta}}_{c,s,o})). \tag{14}$$

When considering the multivariate SIRV model with Weibull distributed multiplier, the dimension d of the parameter space Θ is equal to $\frac{(pq+1)pq}{2}+1$ since one covariance matrix of dimension $pq \times pq$ and one shape parameter for the multiplier τ should be estimated. When a univariate model is consider to represent the wavelet coefficients such as the 2-parameters generalized Gaussian distribution (GGD), d is equal to 2.

Similarly, when the homoscedasticity assumption holds, the decision rule reduces to a nearest neighbor classifier according to the Jeffrey divergence.

$$\hat{c} = \arg\min_{c} \sum_{s,o} J(p(\cdot|\hat{\theta}_t), p(\cdot|\hat{\bar{\theta}}_{c,s,o})). \qquad (15)$$

Note that to the best of our knowledge, even if this last decision rule has been previously proposed in [18, 19, 20], no previous works had been fully formalized in the Bayesian framework.

4.4. Results and discussion

To evaluate the performance of the proposed supervised classification algorithm, the database is split into a training database and a disjoint testing database. From a practical point of view, N_{Tr} training samples are randomly selected for each texture class, the remaining samples are taken as testing samples. In the following, 100 Monte Carlo runs are used to evaluate the performance of the proposed classifiers. Performances are evaluated in terms of kappa index. The kappa index refers to the proportion of consistent classifications observed beyond that expected by chance alone [21, 22].

This experiment is carried out on the MIT Vision texture (VisTex) [23]. This database is composed of 40 classes and 64 images per class of size 64×64 pixels. All texture images are normalized in intensity to have zero mean and unit standard deviation. This normalization gives invariance to affine transformations in the illumination intensity. Here, the stationary wavelet decomposition with 2 scales and Daubechies' filter db4 have been used and a 3×3 neighborhood has been considered to model the spatial dependency of the wavelet coefficients.

Fig. 2 draws the evolution of the average kappa index as a function of the number of training samples on the VisTex database. Results for both univariate (GGD) and multivariate (SIRV) models are respectively displayed in red and blue. Moreover, experiments are carried out to evaluate the influence of the standard deviation σ in the decision rule. The solid and dashed lines correspond to the classification results when the standard deviation is considered (14) and when the homoscedasticity assumption holds (15). As observed, a gain of about 3 points is observed when a multivariate model (such as the SIRV) is used to take into account the spatial dependency compared to an univariate model (such as the univariate GGD). Note also that the proposed classifier (dash lines) has a significant gain of 5 points compared to the decision rule when the homoscedasticity assumption is considered (solid lines).

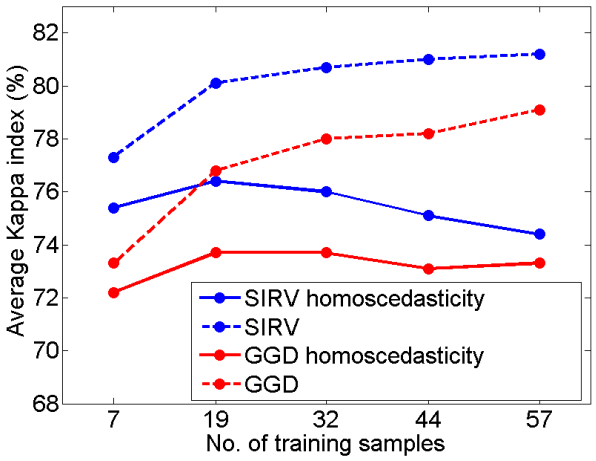

Fig. 2. Evolution of the average kappa index as a function of the number of training samples on the VisTex database for the univariate GGD and the SIRV models.

5. CONCLUSION

This paper has addressed the problem of classification based on an intrinsic prior. After introducing the proposed intrinsic prior distribution as the normal law on a Riemannian manifold when the Jeffrey divergence is considered, we have derived an intrinsic estimation and classification scheme. Next an application to supervised classification texture images has been proposed. Classification results on the VisTex database have shown a gain compared to other conventional approaches. Further works will deal with the extension of the proposed work to an intrinsic multi-barycentric classification algorithm in order to handle the intra-class diversity of natural texture images.

6. ACKNOWLEDGEMENT

This study has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the Investments for the future Programme IdEx Bordeaux (ANR-10-IDEX-03-02).

7. REFERENCES

[1] Q. Han, *Isometric Embedding of Riemannian Manifolds in Euclidean Spaces*, vol. 13, American Mathematical Society, 2006.

[2] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical monographs*, Oxford University Press, 2000.

[3] C. P. Robert, "Intrinsic loss functions," *Theory and Decision*, vol. 40, pp. 192–214, 1996.

[4] J.M. Bernardo and M.A. Juárez, *Intrinsic Estimation*, Bayesian statistics 7, Oxford University Press, 2003.

[5] J.M. Bernardo, *Integrated Objective Bayesian Estimation and Hypothesis Testing*, Bayesian statistics 9, Oxford University Press, 2010.

[6] S. T. Smith, "Covariance, subspace, and intrinsic Cramér-Rao bounds," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1610–1630, May 2005.

[7] C. C. Rodríguez, *Entropic Priors*, W. T. Grandy Jr. and L. H. Schick (Kluwer, Dordrecht), 1991.

[8] H. Snoussi and A. Mohammad-Djafari, *Information Geometry and Prior Selection*, Williams, C. (Ed.), Bayesian Inference and Maximum Entropy Methods, MaxEnt Workshops, Amer. Inst. Physics, New York, 2002.

[9] X. Pennec, "Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 1, pp. 127–154, 2006.

[10] C. Lenglet, M. Rousson, R. Deriche, and O. Faugeras, "Statistics on the manifold of multivariate Normal distributions: Theory and application to diffusion tensor MRI processing," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 3, pp. 423–444, 2006.

[11] R. E. Kass and D. Steffey, "Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models)," *Journal of the American Statistical Association*, vol. 84, no. 407, pp. 717–726, 1989.

[12] Y. Miyata, "Fully exponential laplace approximations using asymptotic modes," *Journal of the American Statistical Association*, vol. 99, no. 468, pp. 1037–1049, 2004.

[13] R. Johnson, "An asymptotic expansion for posterior distributions," *The Annals of Mathematical Statistics*, vol. 38, pp. 1899–1906, 1967.

[14] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, Feb. 1998.

[15] L. Arnold, A. Auger, N. Hansen, and Y. Ollivier, "Information-geometric optimization algorithms: A unifying picture via invariance principles," Tech. Rep., 2011.

[16] N.-E. Lasmar and Y. Berthoumieu, "Multivariate statistical modeling for texture analysis using wavelet transforms," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 790–793, 2010.

[17] A. Schutz, L. Bombrun, and Y. Berthoumieu, "Centroid-based texture classification using the SIRV representation," *IEEE International Conference on Image Processing*, pp. 3810–3814, 2013.

[18] S.-K. Choy and C.-S. Tong, "Supervised texture classification using characteristic generalized Gaussian density," *Journal of Mathematical Imaging and Vision*, vol. 29, pp. 35–47, Aug 2007.

[19] A. Schutz, L. Bombrun, and Y. Berthoumieu, "K-centroids based supervised classification of texture images: handling the intra-class diversity," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1498–1502, 2013.

[20] A. Shabbir, G. Verdoolaege, and G. Van Oost, "Multivariate texture discrimination based on geodesics to class centroids on a generalized Gaussian manifold," *Geometric Science of information*, pp. 853–860, 2013.

[21] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[22] D. Gomez and J. Montero, "Determining the accuracy in image supervised classification problems," *EUSFLAT*, vol. 1 - 1, pp. 342–349, Jul. 2011.

[23] "MIT Vision and Modeling Group. Vision Texture.," Available: http://vismod.media.mit.edu/pub/VisTex.