BIOLOGICALLY PLAUSIBLE CONTEXT RECOGNITION ALGORITHMS

Makiese Mibulumukini, Nicolas Riche, Matei Mancas, Bernard Gosselin, Thierry Dutoit

University of Mons (UMONS), Faculty of Engineering (FPMs)
20, Place du Parc, 7000 Mons, Belgium
Mibulumukini.Makiese@student.umons.ac.be
{Nicolas.Riche, Matei.Mancas, Bernard.Gosselin, Thierry.Dutoit}@umons.ac.be

ABSTRACT

In this paper, four new approaches of global context recognition algorithms (gist) are introduced. They are able to automatically distinguish context differences like buildings, coast, home (indoor), mountain or streets. All proposed models are biologically plausible and are able to deal with both color and gray-level images. They use Gabor or Log-Gabor filters to extract features that better mimic human visual perception. Those features are then classified using a Mahalanobis space (when a subset of features is extracted) or in a high-dimensional Gaussian space (when all features are taken into account) with Support Vector Machines (SVM). The proposed models are compared to a standard state of the art gist model to proof their efficiency.

Index Terms— Context recognition, Gabor and Log-Gabor filtering, Principal Component Analysis, Visual Cortex, Biologically plausible algorithms

1. INTRODUCTION

In the case of humans, scene or object recognition is generally fast, automatic and reliable. This simplicity contrasts with the difficulty of modeling computer vision recognition algorithms that is simple, effective and robust.

From the pioneering work of Hubel [1], a large majority of the recognition systems use bandpass oriented filters (Gabor filters, Gaussian functions ...) according to coding strategies based on those of the visual cortex [2,3].

Other experiments carried out in 1987 on the striate cortex of cats (very close to the one of humans) by Jones and Palmer [4] showed that receptive fields (RF) of simple cells in the cortex are similar to 2D Gabor filters. The RF of some cells was accurately measured by projecting a stimulus as a point on a screen. The resulting responses demonstrated the similarity of the simple cells with 2D Gabor filters.

Thus, several researchers focus on Gabor wavelets or Gaussian models both in facial or scene recognition [5,6,7]. Our study deals with scene recognition and biologically plausible approach for context recognition.

While, in a first stage, object recognition techniques focus on the object itself, the importance of the scene context is growing in several areas of computer vision like scene understanding and classification, object detection and even with saliency algorithms in natural scenes [8,18] .

The context of a scene provides a lot of cues about circumstances and conditions that surround objects. Context recognition is one of the basic bricks of artificial intelligence because it allows smart systems to have a more targeted deployment but also contextual reactions and a similar grouping in a large database.

Torralba *et al.* [7] was the first to suggest that we can define features correlated with scene properties without having to specify individual objects within a scene, just as we can build face templates without needing to specify facial features. This approach, also called "gist", consider the scene as a whole where global features are extracted in order to recognize a context. Torralba *et al.* uses wavelet image decomposition tuned to 6 orientations and 4 scales. The gist vector computed is then reduced to 80 dimensions using Principal Component Analysis (PCA). The classification is achieved by finding the minimum Euclidean distance between the gist vectors of the input images and those of the training set. The use of PCA, during scene analysis, is also biologically inspired because our visual system tends to reduce statistical redundancy of data by enhancing more contrasted information [2,3].

Another global approach that claims to be biologically plausible was proposed by Siagian and Itti [9].

This model makes use of center-surround features from orientation, color, and intensity channels. For each of the 34 sub-channels, it computes the average values from a predefined 4 by 4 grid (16 values) for a total of 544 raw gist values. These values are then reduced using PCA, followed by Independent Component Analysis (ICA), to 80 features which are used to classify scenes with neural networks. Both models [7,9] make use of uncorrelated coding (where relevant features are selected from raw gist values by dimension reduction made by PCA or ICA).

In this paper we propose a comparison between four novel context recognition models and the one of Torralba *et al*.

The proposed models are either based on the uncorrelated coding (where only the less correlated features are used) or distributed coding (where all features extracted are used for classification with no reduction). Also, some are based on luminance and chrominance features and Gabor or Log-Gabor filters.
In the next section we propose our gist approach which uses or not the chrominance information and we compare it with the one of Torralba. We thus show the contribution of luminance in our model. Section 3 provides a more biologically plausible algorithm, a comparison between uncorrelated coding and distributed coding is also made. The last section provides a discussion and conclusion.

2. GABOR BASED CONTEXT RECOGNITION MODELS

Here, we introduce a new gist algorithm based on Gabor magnitude features. When chromatic information is take into account, the model uses decorrelated principal components channels to extract textures with a set of Gabor filters tuned to 8 orientations and 3 scales. The filtering result (for each orientation and scale) is divided into 4×4 pixels blocks where each block contains the average of its pixels.
The gist is a 384-dimensional (3×8×4×4) feature vector for gray-level images (case of monochromatic approach with intensity features only). When color images are used (or when intensity and chromatic information are used), a 1152-dimensional (384 per decorrelated channel) feature vector is extracted. This vector is then reduced using PCA in a Mahalanobis space. Classification is operated by using an SVM classifier with a Gaussian kernel [10].
The decorrelation of channel components precedes Gabor filtering. For gray-level images, no decorrelation is made. For color images, after the decorrelation made by Karhunen-Loeve Transform from RGB space to the PCA space with all the components stored, the image (I_d) of 200×200×3 size is filtered through a bank of Gabor filter with 3 scales and 8 orientations.
The main parameter of Gabor filters is the frequency. With high frequency one may find very fine textures while the main directions will be captured by using lower frequencies. Our Gabor filter uses an empirically optimized minimum wavelength of $\lambda = 5$ which implies a maximum frequency of $f = 1/\lambda$.The gist extracted from a channel of an image Id is such that:

$$Gist_{channel} = \sum_{i=1}^{N_f} abs(I_i) * \omega_i \qquad (1)$$

Where N_f is the total number of filter (N_f=3×8=24); ω_i is a 4×4 averaging window apply to the magnitude of each filtered image I_i. Then, the total number of features for a channel is 4×4×N_f.

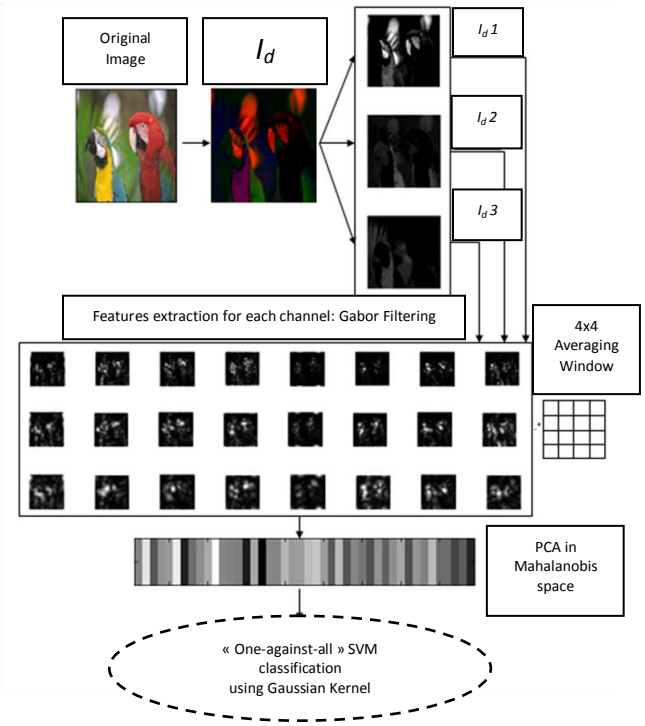

Fig. 1. *Gabor based context recognition architecture with chromatic information: original image is decorrelated (all components are kept after PCA) before to be filtered with Gabor function (with 3 scale and 8 orientations). A 4×4 averaging window selects 16 features for each filtered channel. A total of 384 features (3×8×4×4) per image are reduced in Mahalanobis space before to train a one-against-all SVM that will be used for classification. When chromatic information is not used, the Gabor filtering process is made with intensity channel extract from original image.*

The gist is first projected in a PCA space by extracting the mean Mx of training samples which are then multiplied by the eigenvectors $eVect$ from the covariance matrix of the data previously centered.
In a second step, the gist is projected on a Mahalanobis space by dividing each component by the square root of the corresponding eigenvalue $eVal$. Thus, we get the Mahalanobis projection matrix $wPCA = eVect / \sqrt{eVal}$. The Mahalanobis space is such as the variance along a dimension is 1. The main advantage of the Mahalanobis space with respect to a conventional PCA is that the transformation provides a better similarity measure between the vectors which increases the performance of a recognition system [5,12].
The gist obtained in the Mahalanobis space is reduced by selecting the k most significant components of each channel. A component is significant if its eigenvalue is greater than the hundredth of the maximum eigenvalue.
Most significant components are then used for training a one-against-all SVM classifier and select the class with the highest prediction.

Our model with only luminance information					
%	*B	*C	*H	*M	*S
*B	76	3	13	1	7
*C	0	95	1	4	0
*H	13	6	71	4	6
*M	0	2	1	95	2
*S	3	0	13	4	80
Average on the diagonal of confusion matrix : **83.4%** Standard deviation : **11.1%**					

Our model with intensity and chromatic information					
%	*B	*C	*H	*M	*S
*B	67	4	17	4	8
*C	1	93	0	5	1
*H	13	5	71	4	7
*M	2	5	3	90	0
*S	3	1	19	3	74
Average on the diagonal of confusion matrix : **79.0 %** Standard deviation: **11.7%**					

Model of Torralba [7]					
%	*B	*C	*H	*M	*S
*B	70	2	17	6	5
*C	0	91	2	6	1
*H	17	4	58	10	11
*M	3	16	6	72	3
*S	3	2	4	5	86
Average on the diagonal of confusion matrix : **75.4%** Standard deviation : **13.2%**					

Tab. 1. *Performances of our Gabor based models compared to the gist of Torralba. The real classes are located in row and the predicted classes in column.*B: building and inside city, *C: coast, *H: home, *M: mountain, *S: street.*

During classification, gist features are first projected in PCA-Mahalanobis space then classified in a high dimensional space using SVM parameters (*Figure 1*). The SVM classifier will have to recognize five classes of natural images, namely city and buildings, coast, home, mountain, street. We trained our SVM with 100 color images per class (500 images in total). We compared our model with the gist of Torralba *et al.* [7] (The current state of the art shows that all gist models perform roughly the same [13] in terms of performance and accuracy) using 100 other color images per class for the test (see *Table 1*). All images were taken from Torralba database [14,15].

In *Table 1*, we see that our models perform better than Torralba *et al*. algorithm. This is mainly due to Mahalanobis space projection which provides better results for grouping features when a Gaussian kernel is used instead of a simple Euclidean distance like in [7].

We also see that the contribution of chromatic information is useless comparing to the luminance approach: a lot of redundancy data are eliminated when intensity is used, thus, they already lead us to the most significant features.

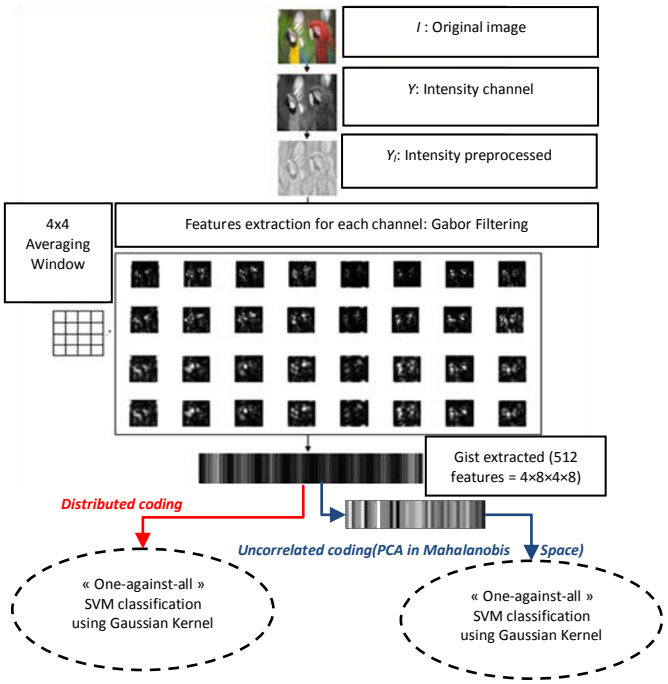

Fig. 2. *Intensity channel (Y) is extracted from original image (I) and preprocessed (Y_l) before being filtered by Log-Gabor function (4 scale and 8 orientations). A 4×4 averaging window selects 16 features for each filtered channel. A total of 512 features (4×8×4×4) per image are reduced in Mahalanobis space (in case of uncorrelated coding) before to train a one-against-all SVM. For distributed approach all gist features are used for classification.*

3. LOG-GABOR BASED CONTEXT RECOGNITION MODELS AND CODING COMPARISON

Luminance is the feature which contains most of the information because the finest area of our eye, called fovea, is mainly photosensitive [11]. As shown in the previous section the chrominance does not bring any advantage, we thus focus here on the intensity only in order to build a model which focuses on the essence ("gist") of a scene.

In this section we focus on tuning our luminance model with Log-Gabor filters. The resulting model is used for comparison between uncorrelated coding and distributed coding (see *Figure 2*).

Log-Gabor filters proposed by Field [16], offer better performance than Gabor filters due to the fact that they are perfectly centered on their logarithmic axis, and thus, they have no continuous component between two successive filters. Then, the transfer function of Log-Gabor filters is extended to higher frequencies. Researches on the statistics of natural image indicate that natural images have spectral amplitude that decreases by approximately $1/\omega$ (where ω is the pulse) [17]. To encode images with these spectral

characteristics, we should use filters with the same spectrum shape. Another good point for the Log-Gabor filters is that they are consistent with measurements made in the visual systems of mammals indicating that we have cell responses that are symmetrical, seen on a logarithmic frequency scale [17]. The boosted gist is such that:

$$Gist = \sum_{i=1}^{s \times o} abs\left(IFFT2\left(H_i .* FFT2\left(Y_l \right) \right) \right) .* \omega_i \quad (2)$$

where ω_i is a 4×4 averaging window applied to each filtered image. The number of scales (*s*) is set to 4 and the number of orientations (*o*) to 8. Y_l (of 256×256 size) is the preprocessed luminance: whitening and local contrast normalization is applied (see *Figure 2*) in order to get more uncorrelated input data. In equation 2 and 3, H_i is the Log-Gabor function (of radial component H_r and angular component H_θ) for a frequency f and an orientation angle θ such as:

$$H_i(f,\theta) = H_r(f).H_\theta(\theta) \quad (3)$$

$$with\ H_r(f) = \exp\left\{ -\frac{\left[\ln(f/f_0)\right]^2}{2\left[\ln(\sigma_r/f_0)\right]^2} \right\}$$

$$and\ H_\theta(\theta) = \exp\left\{ -\frac{(\theta-\theta_0)^2}{2{\sigma_\theta}^2} \right\}$$

Our implementation of log-Gabor filter is based on Peter Kovesi's work [17], and all parameters (central frequency f_0, initial orientation θ_0, radial bandpass σ_r and angular bandpass σ_θ) are set to have the finest bandwidth (that means a very precise RF modeling).

Figure 2 summaries both Log-Gabor approaches proposed (with or without final PCA reduction). The uncorrelated coding provides better recognition result (see *Table 2 and Table 1*) with Log-Gabor filters and confirms that Log-Gabor features give better wavelet than classical Gabor functions: We can note that our uncorrelated coding approach achieves 86.4 % of average recognition rate when a Log-Gabor filter bank of 3 scales and 8 orientations are used (same parameters as those of section 2: without intensity preprocessing, 200×200 image size).

A little difference is observed between uncorrelated and distributed coding, which is mainly due to the fact that PCA reduction keeps relevant components that carry the entire scene contrast while distributed coding hold all features that contribute to the spatial envelope of the scene. On the other hand, the uncorrelated coding drastically decreases the number of features needed for the classifier training (68 against 512 for distributed coding).

4. DISCUSSIONS AND CONCLUSION

We introduced several biologically plausible algorithms that overcome the current state of the art of global approach (or gist approach). We also provide a comparison between uncorrelated and distributed coding that lead us to invest the relevance of PCA during scene of object analysis.

Results in *Table 2* show that PCA gives a very good approximation (comparing to the distributed coding) for efficient classification (88% of average recognition). During training and tests, only 68 features per image are used for uncorrelated coding Log-Gabor approach. Then, we can see the strength of selection made by PCA comparing to the distributed approach (where 512 features per image are used) or Torralba's algorithm (where 80 features are selected from the gist of each image). These results (see *Table 2*) prove that we can achieve a good classification by using the most decorrelated information instead of all features.

All PCA-based methods proposed differ to the one of Torralba by the use of Mahalanobis space that offers a better similarity measure for Gaussian data than euclidean distance [5,12]. Our first approach is also chrominance-based and the use of Log-Gabor filters is a novel contribution to gist-based approaches.

Further work will deal with the role of context as a top-down bias in human visual attention modeling.

Our model with all features kept (distributed coding)					
%	*B	*C	*H	*M	*S
*B	89	1	8	0	2
*C	0	97	0	3	0
*H	6	1	87	3	3
*M	0	6	4	90	0
*S	6	0	6	0	88
Average on the diagonal of confusion matrix : **90.2%** Standard deviation : **4.0%**					

Our model with relevant features kept (uncorrelated coding)					
%	*B	*C	*H	*M	*S
*B	86	1	8	2	3
*C	0	95	0	5	0
*H	7	2	83	3	5
*M	0	7	3	89	1
*S	3	0	9	1	87
Average on the diagonal of confusion matrix : **88.0%** Standard deviation : **4.5%**					

Tab. 2. *Performances of our Log-Gabor based models.*B: building and inside city, *C: coast, *H: home, *M: mountain, *S: street.*

5. ACKNOWLEDGEMENTS

Part of this research is used by the NumediArt Institute (www.numediart.org) and supported by the FNRS/FRIA.

6. REFERENCES

[1] D. H. Hubel, Eye, Brain, and Vision, New York (USA): W. H. Freeman, 1989.

[2] M. Mibulumukini, "De la perception des images à l'algorithme Log-Gabor PCA," in *Workshop sur les Technologies de l'Information et de la Communication*, Casablanca (Morocco), 2011.

[3] A. Guérin-Dugué and H. L. Borgne, "Analyse de scènes par Composantes Indépendantes," in *De la séparation de sources à l'Analyse en Composantes Indéndantes*, Villard-de-Lans (Isère), Ch. Jutten & A. Guérin-Dugué, 2001.

[4] J. P. Jones et L. A. Palmer, «An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex,» *Journal of Neurophysiology*, vol. 58, n° %16, pp. 1233-1258, 1987.

[5] V. Perlibakas, «Face recognition using Principal Component Analysis and Log-Gabor Filters,» 2005.

[6] L. W. Renninger and J. Malik, "When is scene identification just texture recognition?," *Vision Research*, pp. 2301-2311, March 2004.

[7] A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition," *International Conference on Computer Vision*, Octobe 2003.

[8] M. Mancas, "Relative Influence of Bottom-up & Top-down Attention," *Attention in Cognitive Systems, Lecture Notes in Computer Science*, February 2009.

[9] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300-312, February 2007.

[10] S. Canu, Y. Grandvalet, V. Guigue and A. Rakotomamonjy, "SVM and Kernel Methods Matlab Toolbox," 2005. [Online]. Available: http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html.

[11] O. L. Meur, "Attention sélective en visualisation d'images fixes et animées affichées sur écran : modèles et évaluation de performances - applications," Nantes, 2005.

[12] V. Perlibakas, "Computerized face detection and recognition," Lithuanian, 2004.

[13] M. Viswanathan, C. Siagian and L. Itti, "Comparisons of Gist Models in Rapid Scene Categorization Tasks," *Proc. Vision Science Society Annual Meeting (VSS08)*, May 2008.

[14] A. Quattoni and A.Torralba, "Indoor scene Recogition," 2009. [Online]. Available: http://web.mit.edu/torralba/www/indoor.html.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42(3), pp. 145-175, 2001.

[16] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America*, vol. 4, no. 12, pp. 2379-2394, December 1987.

[17] P. Kovesi, "What Are Log-Gabor Filters and Why Are They Good?," [Online]. Available: http://www.csse.uwa.edu.au/~pk/research/matlabfns/.

[18] A. Torralba, A. Oliva, M. Castelhano and J. M. Henderson, "Contextual Guidance of Attention in Natural scenes: The role of Global features on object search", Psychological Review. vol. 113(4) 766-786, October 2006.