Argument for Scene Categorization with Image Amplitude Spectra

Purpose

Although many researches on scene recognition have shown that semantic information is mainly conveyed by the phase spectrum [1, 2], most of psychophysics works [3-5] paradoxically use the amplitude spectrum as the principal component of a natural scene. The aim of our work was to test whether amplitude without phase spectrum data contain sufficient information to classify natural scenes. Ninety-two noisy natural scenes belonging to four categories (Beach, City, Forest and Indoor Scene) were presented both to human subjects and to a classification model which reproduces early human visual processes (retina, V1 cortical cells). The shapes of the noise's amplitude spectrum used were those of cities (N_n), of beaches (N_n), of generic natural scenes [6], or white noise. Behavioral and simulation data show that scene *categorization was easier for pictures composed of a congruent noise than for incongruent ones*. This leads us to think that amplitude without phase spectrum could be sufficient to make natural scene classification.

Amplitude spectrum

The categorical noise contains only information relative to the scene category in the amplitude spectrum. Then, all (other) things being equal, if performance on scene categorization varies as a function of the "*category*" of the noise, then the amplitude spectrum carries some useful category information processed by the visual system.

Stimuli

92 stimuli were built from 28 natural scene grey-level pictures and noises from four categories. 7 pictures (I_p) from four categories (Beach I_b, City I_c, Forest I_f and Indoor Scene I_i) were chosen according to their amplitude spectrum's shape: vertical for beach, horizontal for city, both for indoor and isotropic for forest [8, 9].
Four classes of noise (N_m) were built: horizontal N_h, vertical N_v, white N_w and natural N_n. The shapes of noises were obtained (cf. equation 1) by multiplying white noise *w(x)* with the mean amplitude spectra *mm(x)* of city and beach (for N_h and N_v) or with *1/f* spectra (for N_n) [7].

$$N_m(x) = \frac{DFT^{-1}\left[DFT\left[w(x)\right] * DFT\left[m_m(x)\right]\right]}{\varepsilon(N_m)} * \varepsilon(I_p) \quad (1)$$

$$\varepsilon\left[f(x)\right] = \sum \left|DFT\left[f(x)\right]\right|^2 \quad (2)$$

Each noise is normalized (cf. equation 2), and a coefficient *k* is applied in order to ensure a sufficient masking. On the basis of a pre-test experiment, we choose 1.75 as coefficient value, which resulted in a categorization performance superior to 80% accuracy.
When adding noise to natural scenes (cf. equation 3), their respective amplitude spectra could be neutral, congruent (Beach with vertical noise or City with horizontal noise) or incongruent according to the scene category (cf. figure 1).

$$I_{pm}(x) = I_p(x) + k * N_m(x) \quad (3)$$

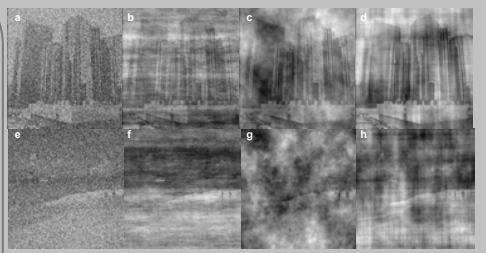

Figure 1: Examples of beach (1st row) and city (2nd row) with the four categorical noises N_w (*a* & *e*), N_h (*b* & *f*), N_n (*c* & *g*) and N_v (*d* & *h*). Congruent noisy scenes are pictures *b* and *h*, *d* and *f* are incongruent.

As *w(x)* is a white noise, its module is constant and its phase random,

$$N_m(x) = \left|M_m(x)\right| e^{j \cdot rand(x)} \quad (4)$$

If ones swaps the phase spectrum of a scene, it contains no more image specific information and therefore the picture could not be recognisable [2]. We could assess that if the phase spectrum of a signal is multiplied with a random signal, the resulting signal is random:

$$N_m(x) = \left|M_m(x)\right| e^{j\varphi(M(x)) * rand(x)} \quad (5)$$

Loci of Information

Gabor Model

Former studies have shown that the global distribution of the local dominant orientations appears to be a powerful feature for discriminating between four semantic categories of real world scenes [10].

Retinal pre-processing

The retinal photoreceptors make a space-time high-pass filtering after an adaptive compression process. This results in a contrast equalization of the image and spectral whitening which compensates for the 1/f image amplitude spectrum [11].

Cortical filtering

In area V1 of the visual cortex, the retinal image is decomposed by the filtering of cortical neurons, which are sensitive to various spatial frequency bands and various orientations of the stimuli. Here, we aim at categorizing and not describing scenes, so we simulate complex cells which are invariant to object position in the scene. These cells, described by the means of Gabor wavelets, provide the local energy of images. Images are filtered by Gabor wavelets into 7 frequency bands and 7 different orientations. According to the physiological data about the visual cells [12], the relative radial bandwidth of the Gabor filters is fixed at 1 octave. So, each image is analyzed with a bank of 49 Gabor filters, the output energies of which provide a point in a 49-dimensions space.

Behavioral Measures

Participants

Eleven students from the National Polytechnic Institute of Grenoble (ages between 18 and 25, with normal or corrected vision) were volunteers to participate in the experiments.

Procedure

The participants sat in front of the screen at a distance of 1.2 meters. Stimuli subtended a visual angle of 9.7°. Each picture was displayed for 20 ms; it was preceded by a black central fixation dot for 1.5 s, and followed by a white screen for 1.5 s. Participants were to press a button as quickly and as accurately as possible if the picture was a beach or city, and not to do so for the other picture categories ('Go – no Go' procedure). Reaction times were recorded with a vocal key and responses were keyed to disk by the experimenter. No feedback was given. Each experiment started with 2 practice sets. The order of trials was random.

Results

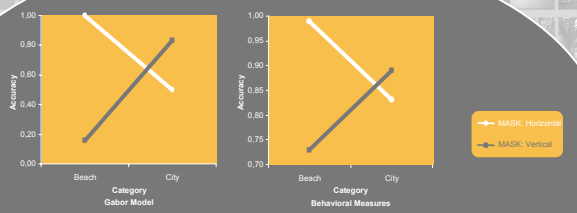

Figure 2 : Percentage of accuracy as a function of the scene Category (Beach *vs* City) and the amplitude spectrum characteristics of Masks (Horizontal *vs* Vertical). (The scales of plots differ.)

Results

Gabor Model

In order to measure our model accuracy, we compute the Euclidean distance between the amplitude spectrum of each noisy image (i.e Beach and City with N_h and N_v noises) and the mean amplitude spectra of beach and city categories. Images are associated to the closest category. Accuracy is the ratio of good associations to the cardinal of each category. (Figure 2a)

Behavioral Measures

A 2 x 2 ANOVA (Mask x Category), carried out on accuracy, showed no effect of Category ($F(1,10) < 1$), an effect of Mask ($F(1,10)=6.33$, $p<0.03$): Participants responded more accurately for horizontal mask (N_h) than for vertical mask (N_v), and a Mask x Category interaction ($F(1,10)=10.14$, $p<0.01$): Participants responded more accurately when the amplitude spectra of the mask and the category were congruent (I_{hb} and I_{cv} conditions) than incongruent (I_{hc} and I_{cb} conditions) ($F(1,10)=12.025$, $p<0.01$).
A 2 x 2 ANOVA (Mask x Category), carried out on correct response times, leads to similar results. (Figure 2b)

Discussion

With this work, we have shown that image amplitude spectrum is sufficient to make some natural scene clusters. So, shaping noise is an efficient paradigm in order to investigate low-level vision processes.
Moreover, these results provide us a validation of our model of scene categorization based on the global distribution of image energy.

There are still two main problems in the experiment:

- Noise shapes could be changed to render them more orthogonal to the shape of semantical category amplitude spectra [7, 10, 11]
- The signal to noise ratio (SNR) of the stimuli computed in the spatial domain can also explains our results (even if it is equal to 1 in the frequency domain). Now, the SNR cannot account for the structural organisation of a natural scene. Equalizing the histogram produces a SNR ≠1 without modifying the structure of the picture. However if one swaps each pixel randomly the SNR will be equal to 1 and the picture become noise

Alan Chauvin[∞♦], Nathalie Guyader[♣], Christian Marendaz[∞], Jeanny Hérault [♦];
∞:Laboratoire de Psychologie et NeuroCognition - UMR 5105, UPMF BP 47, 38040 Grenoble, France
♦: Laboratoire des Images et des Signaux - UMR 5083, INPG, 46 Av. Félix Viallet 38031 Grenoble Cedex,

Sponsors

1. Oppenheim, A.V. and Lim, J.S., *The importance of phase in signals*. Proc. IEEE, 1981. 69: p. 529-541.
2. Morgan, M.J., Ross, J., and Hayes, A., *The relative importance of local phase and local amplitude in patchwise image reconstruction*. Biological Cybernetics, 1991. 65: p. 113-119.
3. Field, D.J. and Brady, N., *Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes*. Vision Res, 1997. 37(23): p. 3367-83.
4. Ruderman, D.L., *Origins of scaling in natural images*. Vision Res, 1997. 37(23): p. 3385-98.
5. Tolhurst, D.J. and Tadmor, Y., *Band-limited contrast in natural images explains the detectability of changes in the amplitude spectra*. Vision Res, 1997. 37(23): p. 3203-15.
6. Van der Schaaf, A. and Van Hateren, J.H., *Modelling the power spectra of natural images: statistics and information*. Vision Res, 1996. 36(17): p. 2759-70.
7. Guérin-Dugué, A. and Oliva, A., *Classifications of scene photographs from local orientations features*. Pattern Recognition Letters, 2000. 21: p. 1135-1140.
8. Hérault, J., *De la rétine biologique aux circuit neuromorphiques, in Les Systèmes de Vision*, J.M. Jolion, Editor. 2001, Hermes.
9. De Valois, R.L. and De Valois, K.K., *Spatial Vision*. 1988, Oxford: Oxford University Press.
10. Oliva, A. and Torralba, A.B., *Modeling the Shape of the Scene: A holistic representation of the Spatial Envelope*. International Journal of Computer Vision, 2001. 42(3): p. 145-175.
11. Guyader, N. and Hérault, J., *Représentation échasse-fréquence pour la catégorisation d'images*. in *GRETSI*. 2001. Toulouse.