Scene classification – is amplitude or phase more important?

Introduction

When we walk around, there is a great deal of information available to our brain to be processed to produce our perception and action. However, it is impossible to account for every bit of the available cues because the cognitive load would crash and burn up our brain. Hence, there must be a mechanism to filter out the evolutionarily unimportant details in order to quickly comprehend the events and situations at first sight. For instance, when looking at the scene in Fig. 1, we can quickly tell what it is about at a glance without paying much attention to every details of the images. So it appears that there exists the gist of every scene that we use for this kind of quick perception. The question is how we can quantify the vague concept of the gist of a scene.

Figure 1

There are some efforts along this line of research in the literature. In Oliva & Torralba (2001), the gist of a scene was posited to be contained in the statistical regularities of the amplitude spectra along some primary spatial envelope dimensions. This spatial envelope theory teases apart

several properties of a scene such as naturalness, openness, ruggedness, expansion, etc. based on the amplitude information. By reducing the dimensionality of the image using Gabor wavelets and performing simple training with discriminant analysis or linear regression, the authors proved that they can robustly classify the scenes in all proposed dimensions (naturalness, openness, etc.) with highly significant rate above chance (greater than 86%). In the same vein, Guyder et at. (2004) attempted to support the dominance of amplitude spectrum in scene gist recognition with psychophysical experiments. Specifically, they employed a priming paradigm in which either the amplitude or phase spectrum of the prime image is similar to the target scene while the other spectral feature is disrupted. The reaction time result shows significant differences between consistent and inconsistent priming for amplitude-modulated primes but not for phase-modulated primes.

Nevertheless, we have reasons to believe that the phase information plays a significant role in the rapid scene recognition due to the abundance of evidence on the importance of phase both in signal processing and human perception (Oppenheim and Lim,1981; Thomson, Foster & Summers, 2000). A demonstration in Fig. 2 and 3 can give us some intuition on this. Each image in Fig. 2 has the phase preserved and the amplitude progressively replaced by the amplitude of the other image. On the other hand, it is reverse in Fig. 3 with the preserved amplitude and contaminated phase (the technical details are discussed in Method section). In Loschky et al. (2007), the authors used the masking paradigm to probe the time course of scene gist perception. Based on the findings of previous studies that the conceptual and spectral similarity of the mask to the target is inversely correlated with the target detection, they systematically modulated the phase and amplitude of the mask with respect to the target and measured the human performance in a yes/no scene classification task. Their main findings is that the phase contributes more to the scene perception than the amplitude which is contradictory to the spatial envelope theory.

Therefore, it is interesting to investigate the role of phase and amplitude spectrum in human rapid scene recognition in another psychophysical paradigm. In this project, instead of indirect methods in previous studies, I directly manipulated the phase and amplitude spectrum of the target images

and observe the effects on human performance in natural/manmade scene classification task. I also tested the performance of spatial envelope model in the same task and on the same stimuli set for comparison with the experimental result.

Method

Stimuli

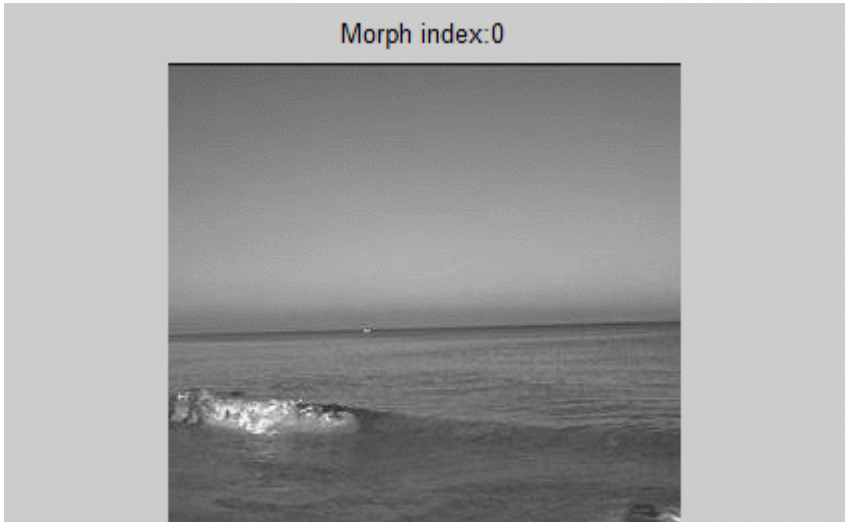
The scene images used in the experiment were downloaded from the SUN image database. There are 2688 color images (256 x 256 pixels) categorized into four manmade (building, street, inside city, highway) and four natural (mountain, coast, forest, open country) scene types. The number of manmade and natural scenes are roughly equal (1472 and 1216). All images were converted to gray scale with preserved luminance. In order to study the contribution of phase and amplitude spectrum to scene gist perception, the phase and amplitude of the images were manipulated independently. In the amplitude morphing condition, the phase of the seed image was fixed while the amplitude was modulated as follows:

$$Am = (1-morphIndex) * A1 + morphIndex * A2$$

where Am is the morphed amplitude, A1 is the amplitude of the seed image, A2 is the amplitude of an arbitrary image of different scene type (natural if seed is manmade and vice versa), morphIndex is the amount of morphing with 0 representing no morphing (the seed image's amplitude was used), 1 representing maximum morphing (another image's amplitude was used) and intermediate levels of morphing in between. It is similar for the phase morphing condition in which the amplitude of the seed image was preserved and the phase was modulated the same way as the amplitude above.

Fig.2 and 3 illustrate the morphing effect on a pair of natural and manmade scenes. In the left panel of Fig. 2, the amplitude of the seed manmande image on the left was morphed towards that of the natural image on the right. It is opposite for the animation on the right with the natural seed

image morphed toward the manmade image. On the other hand, Fig. 3 shows how the phase morphing works on the same pair of images. The morphing indices are from 0 to 1 in steps of 0.025 and each morphed image was presented for 100 ms.



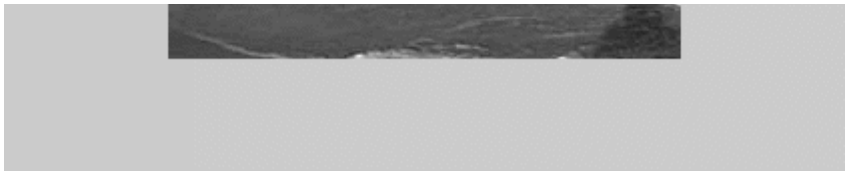
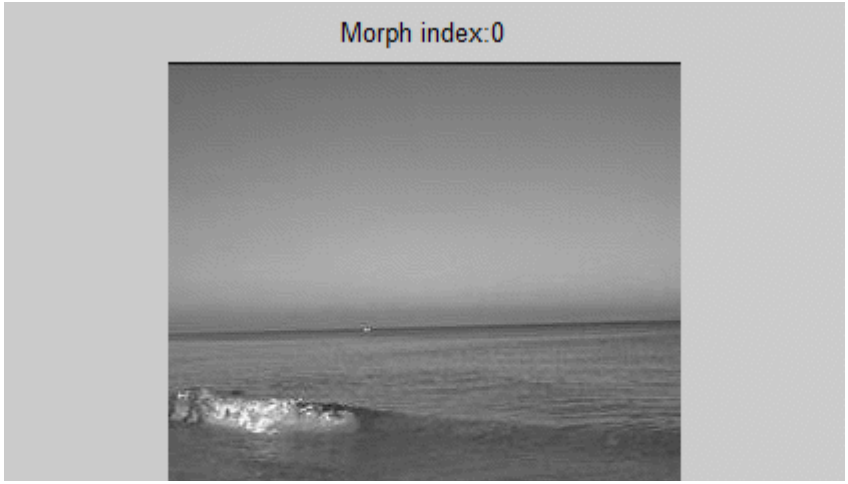


Figure 2



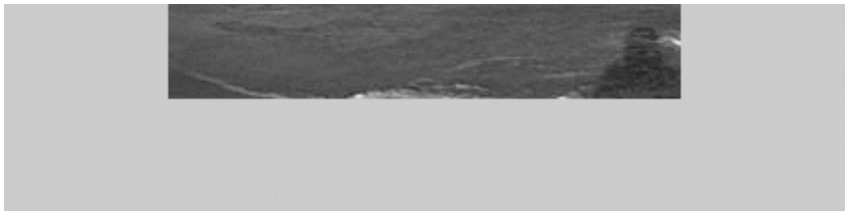


Figure 3

Experimental design

In the experiment, we used a Mac computer to control the presentation of stimuli on a LCD screen (1920 x 1200 pixels and 20.2 x 12.5 inches) with refresh rate of 60 Hz. There are 2 main experimental conditions: the amplitude morphing and phase morphing, each with 20 blocks. In each block, there are 22 trials corresponding to 11 morphing levels (0:0.1:1) for two scene types (natural and manmade). In one trial, a pair of natural and manmade scene was randomly drawn from the database. Then, the seed image was chosen based on the trial type and morphed towards the other image. The order of trials in each block was randomized. All images used in a morphing condition (amplitude or phase) were different. At the beginning of the experiment, there were 20 practice trials with unmorphed natural and manmade scenes and the feedback was given. There was no feedback in the following test trials.

The presentation of stimuli was as follows. An oval fixation first appeared at the center of the screen for 500 ms. Then it was followed by a morphed image which was presented for 30 ms. Right after that, a random noise mask was displayed and stayed on screen until the subjects responded. The subjects were instructed to choose the type of the presented scene (natural or manmade) by pressing the buttons on the joystick. The viewing distance was roughly 24 inches. For visual clarity, the images were resampled and displayed at 350 x 350 pixels. So, an image subtends 8.8 degree of visual angle.

Simulation

It is interesting to compare the performance of the spatial envelope model with that of human observers on the same set of stimuli. Therefore, besides the psychophysical experiment, I replicated the computer vision model in Oliva & Torralba (2001) and used it to do the same task on the same image set. Essentially, the gist of a scene was computed by projecting the high-dimensional image onto the low-dimensional Gabor wavelet space. For simplicity, Gabor filters of four frequency scales and four orientations were used. Therefore, the image of 65536 dimensions in pixel space was reduced to 16 dimensions of Gabor wavelet basis. Due to the binary nature of the classification task (natural v.s. manmade scene), the quadratic discrimination analysis was employed for the model. In the training stage, the gists of 400 images (200 natural and 200 manmade) were computed and used for the training. To validate the model in normal classification task, all images in the database not used in the training were classified by the algorithm. The average percent correct of the validating classification is 86% which is equal to the accuracy reported in Oliva & Torralba (2001).

In the testing stage, 1216 natural scenes and 1216 manmade scenes were amplitude- and phase-morphed with the morphing index ranging from 0 to 1 in step of 0.05. The above classification model was applied to those images to determine whether it is natural or manmade.

Data analysis

The responses of all five subjects were aggregated and averaged across morphing types (amplitude/phase), morphing indices(0:0.1:1) and two seed image types (natural/manmade). The standard error was computed by bootstrapping method. Due to the disparate distinct trends of data between morphing conditions and scene types, the data were not fit by a functional form. Moreover, the data were clean enough that we can infer the general trend by simply connecting the data points.

Result

Simulation

Fig. 4 shows the result of spatial envelope model in the scene classification task. The left plot demonstrates the amplitude morphing condition and the right plot shows the phase morphing condition. In accordance with the convention in plotting psychometric function, instead of using the morphing index on the x-axis, I converted it to the percent preserved calculated as (1-morphIndex)*100% which denotes how much of the seed image is preserved in terms of amplitude or phase. The y-axis represents the percent the morphed image was classified as the seed.

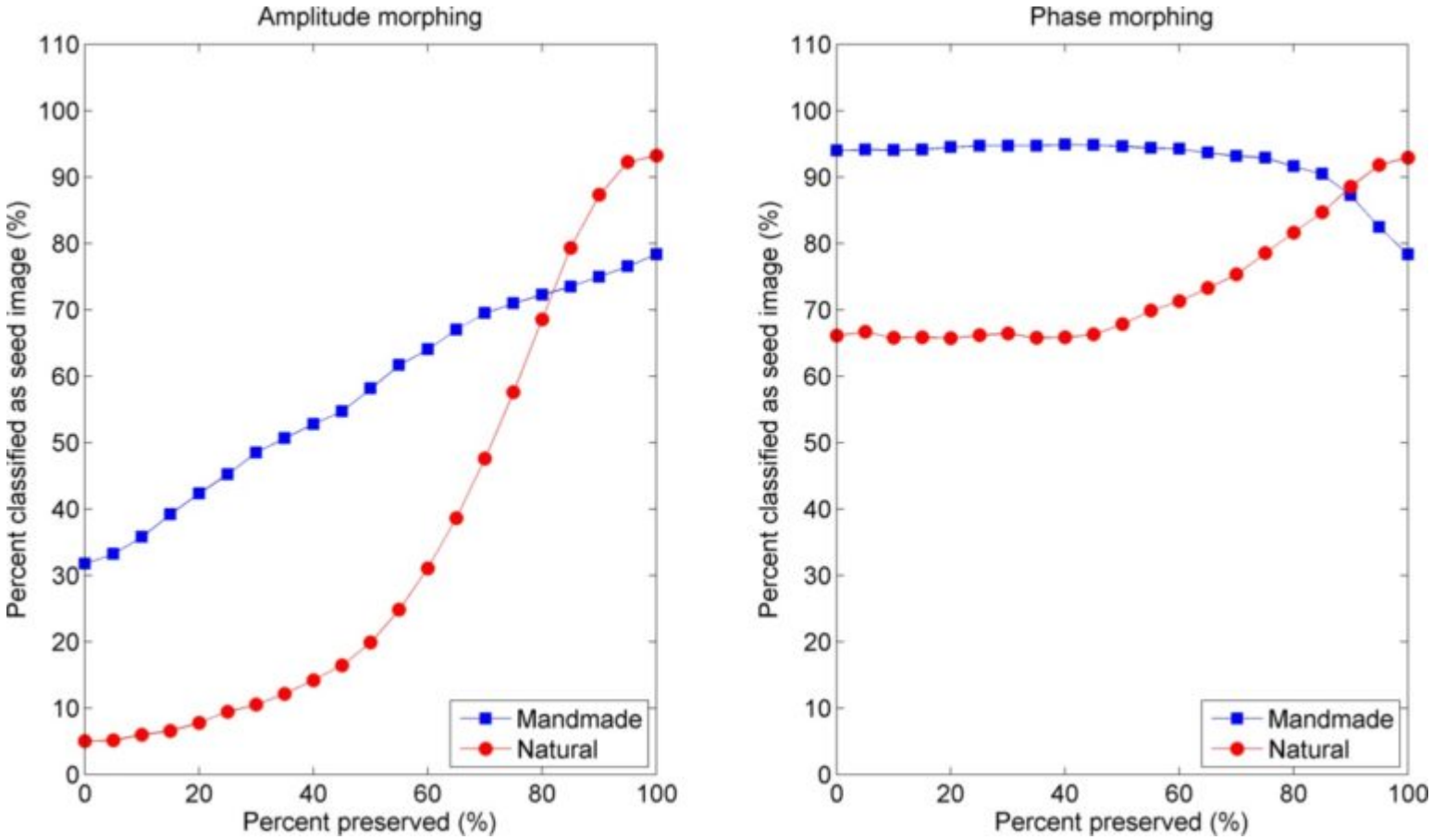


Figure 4

In general, the result conforms to the tenet of spatial envelope theory that the gist of a scene is

predominantly conveyed in its amplitude spectrum. Specifically, the amplitude information appears to uniquely determine the identity of the scene as shown in the left plot in which the more the amplitude of seed image was morphed, the less likely the model classified it as the seed. On the other hand, the right plot implies that the phase information does not contribute much in the classification task. Furthermore, it is strange that when the manmade scene's phase was morphed, it increases the chance of being classified as the seed.

As a cross-validation, we observe that the 0 percent preserved for manmade condition on the right plot will correspond to the 0 percent preserved for natural condition on the left plot because they are both manmade amplitude mixed with natural phase. So the percents classified as seed image of these 2 conditions should be roughly one minus the other due to different seeds. We can check it for the other case as well and it assures us that we are doing the right things. So maybe the strange shape of manmade condition on the right is simply because the phase is not a good criterion for the spatial envelope model, at least for the manmade seed image.

Another observation is that the algorithm appears to be more sensitive to the spectral modulation of natural scenes with the steep slope of the red curve on the left plot whereas the manmade scenes are more resistant to the morphing effect. In other words, the model is somewhat biased to manmade scenes.

Experiment

In a similar format to Fig. 4 of simulation result, Fig. 5 shows the psychophysical results averaged across all 5 subjects. The error bar is +/- 1 SEM. Strikingly, a first glance at the result gives us the impression that it is almost opposite to the prediction of spatial envelope model. In particular, it appears that the human observers based their judgement more on the phase information than the amplitude with steeper slopes of performance curves in phase morphing condition. However, the human perception of scene gist appears to be modulated to some degree by the spectral amplitude change although it is much less noticeable.

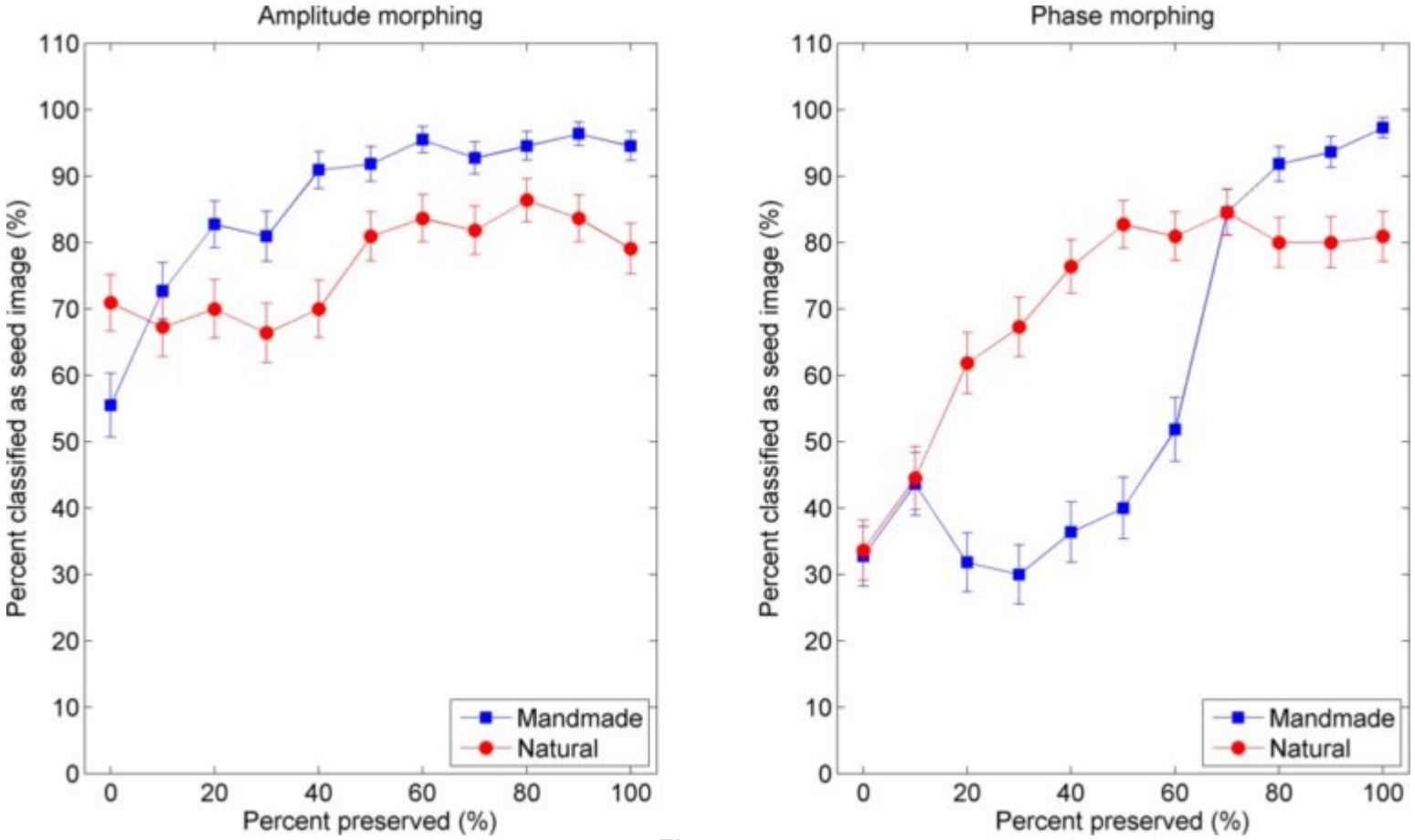


Figure 5

A second discrepancy between the model and human observer is that human perception seems to be more biased to the natural scene in the sense that the recognizability of natural images is less corrupted by both kinds of spectral modulation. For both plots, it is clear that the slope of performance curve of natural seed is always less shallow than that of manmade seed.

Discussion

The psychophysical results tend to favor the dominant role of phase information in human perception of scene gist. It is quite congruous with our intuition when looking at the animations in Fig. 2 and 3. Furthermore, the demo seems to explain some other aspects in the shape of human performance curves. For instance, Fig. 2 gives us the impression that when the amplitude of manmade seed image is severely contaminated by that of natural scene, it looks more natural whereas the amplitude-morphed natural scene on the right always looks natural. That might explain the slopes of manmade and natural curve in low percent preserved range in the left plot of Fig. 5. Similarly, from the animation in Fig. 3, it seems that the phase morphing effect is faster and stronger for manmade scene than for natural scene. It is consistent with the steep slope of manmade curve and gradual slope of natural curve in the right plot of Fig. 5.

The agreement of experimental results and our impression in the demo is remarkable because it implies that we are able to extract enough information in a flash for a high-level representation of the world such as the naturalness. In addition, this information is not necessarily based on the amplitude spectrum as proposed by the spatial envelope theory but it may rely more on the phase spectrum. If this is true, it again proves how the human evolution is adaptive to the environment because the phase information is more important than the amplitude given the statistical regularities of the amplitude spectrum in natural images (Field, 1987).

One challenge to a theory on the importance of phase in rapid scene recognition is that the early visual system, esp. the primary visual area, mainly contains the cells tuning to the features of spectral amplitude of the stimuli such as contrast and orientation. The phase information appears to be represented later through the spatial configuration of the stimuli in higher areas such as MT. Therefore, if there is evidence on some shortcut pathway to the higher areas of the visual system, the phase theory will be neurally substantiated.

In terms of computational modeling, there is also some difficulties in advancing some phase-based model because the amplitude spectrum has been an established subject in signal analysis due to its straightforward interpretation and convenient manipulation. However, there is a growing interest in employing this long ignored information in several areas including speech processing

(Aarabi et al., 2005), image processing (Skarbnik et al., 2010) and medical imaging (Kothapalli et al., 2005). Therefore, it would be interesting to see how the progression in the field of signal processing could promote our understanding of human perception.

A future direction could be using the current paradigm to test other spatial envelope dimensions proposed in Oliva & Torralba (2001) such as the openness, ruggedness, etc. Also, because the subordinate category such as “beach” v.s. “city” was used in other psychophysical paradigms (Guyder et at., 2004; Loschky et al., 2007) instead of the superordinate category (natural v.s. manmade), it would be interesting to test the scene gist recognition on those levels as well.

Reference

Aarabi, P., Shi, G., Shanechi, M. M and Rabi, S.A (2005). Phase-Based Speech Processing, World Scientific.

Field, D. J. (1987). Relations between the statistics of natural images and response properties of cortical cells. Journal of the Optical Society of America A, 4, 2379–2394.

Guyader, N., Chauvin, A., Peyrin, C., He´rault, J., & Marendaz, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude based process. Comptes Rendus Biologies, 327, 313–318.

Loschky L. C., Sethi A., Simons D. J., Pydimari T., Ochs D., Corbeille J. (2007). The importance of information localization in scene gist recognition. Journal of Experimental Psychology: Human Perception and Performance, 33, 1431–1450.

Nikolay Skarbnik, Yehoshua Y. Zeevi, Chen Sagiv (2010). The Importance of Phase in Image Processing, CCIT report.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42, 145–175.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. Proceedings of the IEEE, 69, 529–541.

Sri-Rajasekhar Kothapalli, Chandra S. Yelleswarapu, and Sriram G. Naraharisetty (2005). Spectral phase based medical image processing. Acad Radiol.

Thomson, M. G. A., & Foster, D. H. (1997). Role of second- and third-order statistics in the discriminability of natural images. Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 14, 2081–2090.