Natural Versus Artificial Scene Classification by Ordering Discrete Fourier Power Spectra

G.M. Farinella[1], S. Battiato[1], G. Gallo[1], and R. Cipolla[2]

[1] University of Catania, IT
{gfarinella,battiato,gallo}@dmi.unict.it
[2] University of Cambridge, UK
cipolla@eng.cam.ac.uk

Abstract. Holistic representations of natural scenes is an effective and powerful source of information for semantic classification and analysis of arbitrary images. Recently, the frequency domain has been successfully exploited to holistically encode the content of natural scenes in order to obtain a robust representation for scene classification. In this paper, we present a new approach to *naturalness* classification of scenes using frequency domain. The proposed method is based on the ordering of the Discrete Fourier Power Spectra. Features extracted from this ordering are shown sufficient to build a robust holistic representation for *Natural* vs. *Artificial* scene classification. Experiments show that the proposed frequency domain method matches the accuracy of other state-of-the-art solutions.

1 Introduction and Motivations

Humans are able to recognize complex visual scenes at a single glance, despite the number of objects with different poses, colors, shadows and textures that may be contained in the scenes. Recent studies suggest that the human visual system achieves its discrimination power using global information about the overall structure of the scene [1].

Many computer vision researchers [2,3,4,5] have proved that holistic approaches can be efficiently used to solve the problem of rapid and automatic scene classification. In particular holistic approaches are able to recognize a scene bypassing the recognition of the objects and the details inside the scene. All of the proposed holistic approaches share the same basic structure that can be schematically summarized as follows:

1. A suitable features space is built (e.g. textons vocabulary). This space must emphasize specific image cues such as, for example,
   corner, oriented edges, etc.
2. Each image under consideration is projected into this space. A descriptor, as a whole entity, of the image projection in the feature space is built (e.g. textons histograms).

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 137–146, 2008.

3. Scene classification is obtained by using a probabilistic model on the new holistic representation of the images.

A wide class of classification algorithms based on the above scheme work extracting features on perceptually uniform color spaces (e.g. CIELab). Typically, filter banks [3] or local invariant descriptors [5] are employed to capture image cues and to build the visual vocabulary to be used in a bag of visual words model [2]. An image is considered as a distribution of visual words and this holistic representation is used to perform classification. Eventually local spatial constraints are added in order to capture the spatial layout of the visual words within images [2].

Alternatively, as Torralba and Oliva have shown in several papers, the frequency domain can be a useful and effective source of information to encode holistically an image for scene understanding. The statistics of natural images on frequency domain [6] reveal that there are different spectral signatures for different image categories. In particular by considering the shape of the spectrum of an image it is possible to address scene category [4], scene depth [7], and object priming [8] such as identity, scale and location.

This paper propose a new holistic representation obtained in the discrete Fourier frequency domain. This representation is used to perform classification at the superordinate level of description of scenes [4]. Specifically we are interested in discriminating *Natural* vs. *Artificial* scenes[1].

The new and main contribution of the present work is to demonstrate that the ordering of the frequencies is a useful feature for *naturalness* classification. Ordering the frequencies indeed enables capture the overall shape of the scene in frequency domain. Similar to [4] we infer the category of an image from the shape of the scene spectrum in the frequency domain, but differently from [4], we use the relative position of pre-selected ordered frequencies as a global holistic cue.

This paper is organized as follows: Section 2 describes the model we have used and how the *discriminative frequencies* are selected. Section 3 illustrates the dataset, the setup of our experiments and the results obtained using the proposed method. Finally in Section 4 we conclude and describe future works.

2 Ordering Discrete Fourier Power Spectra

The classifier proposed in this paper has been obtained taking the following statements as starting points:

Statement-I: The energy distribution over the spectrum space, as captured by the spectra *signature* or *shape*, is very distinctive for different scene categories. This statement is strongly supported by the results in [4,6].

[1] In this work the term *Artificial* refers to images in which are depicted man-made environments (cities, buildings, streets, etc) whereas *Natural* refers to images in which natural landscapes are represented (open country, mountain, forest, coast, etc).

Statement-II: It is possible to experimentally observe [4,6] that the spectrum of *Natural* scenes is quite isotropic with no preferred direction, whereas the spectrum of *Artificial* scenes have strong "vertical" and "horizontal" axis. Thus the "diagonal[2]", "horizontal[3]" and "vertical[4]" frequencies should be particularly powerful in scene *naturalness* discrimination. Specifically, *Artificial* scenes have well marked "horizontal" and "vertical" structure in spatial domain, so "vertical" and "horizontal" frequencies are stronger than other frequencies. On the other hand in *Natural* scenes the "diagonal" frequencies are stronger than "vertical" and "horizontal" frequencies. A classic exception to the isotropy of *Natural* scene spectrum is related to the scenes in which "landscapes" are depicted. In these scenes there is a well marked "horizontal" structure in spatial domain, so "vertical" frequencies are stronger than "horizontal" frequencies. Taking into account this last specific case the "diagonal" and the "horizontal" frequencies should be particularly powerful in classifying a scene as *Natural* or *Artificial*.

Statement-III: Shape descriptors that take into account the relative positions of shape contour points have proved very powerful for shape recognition [9]. Although this property has been demonstrated in the spatial domain it is straightforward to adopt this kind of strategy to recognize *shape* or *signatures* in the frequency domain.

A key point of the proposed classification scheme indeed is to look at each specific frequency $f_{x,y}$ within its *context*, i.e. according to the relationship that $f_{x,y}$ has with the other frequencies in the Fourier space. To capture the shape signature of images in the frequency domain we use the position of the frequencies after ordering them according to their magnitudes. The proposed holistic image representation is then inspired by the following rationale.

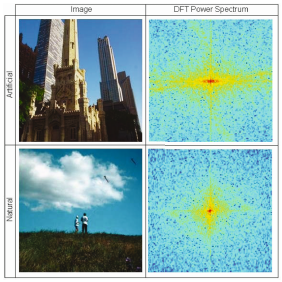

Fig. 1. Examples of two images from Natural (top-left) and Artificial (Bottom-Left) classes and their corresponding Discrete Fourier Power Spectrum (Right column)

[2] "Diagonal" frequencies correspond to any kind of structure without a preferred "horizontal" or "vertical" direction on spatial domain.
[3] "Horizontal" frequencies correspond to "vertical" structure on spatial domain.
[4] "Vertical" frequencies correspond to "horizontal" structure on spatial domain.

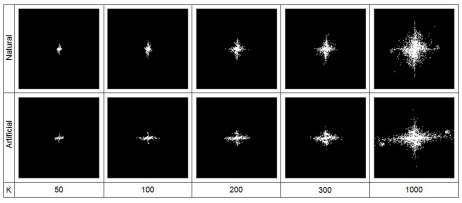

Fig. 2. The first K∈{50,100,200,300,1000} frequencies in the magnitude ordering of the power spectrum corresponding the two images reported in Figure 1

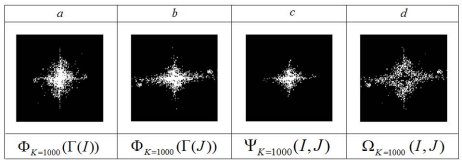

Fig. 3. The K-bounded descending ordering of Γ for the two images reported in Figure 1 allows to capture the shape of their related category ((*a*) *Natural*, (*b*) *Artificial*). The frequencies present in both ((*c*) not Discriminative) and just in one ((*d*) Discriminative) of the K-bounded ordering are shown.

Let I be an image and let $\Gamma(I)$ be the corresponding Discrete Fourier Power Spectrum (Figure 1). Ordering the frequencies in the Discrete Power Spectrum by their magnitude and selecting the first K frequencies in the descending order, it is possible to capture the specific shape signature of the image class (Figure 2).

Let $\Phi_K(\Gamma(I))$ be the set of the top K frequencies when $\Gamma(I)$ is ordered by decreasing magnitude. Figure 2 shows $\Phi_K(\Gamma(I))$ sets for different K values relative the images in Figure 1. If I is a *Natural* scene image and J is an *Artificial* scene image, the frequencies in

$$\Psi_K(I,J) = \Phi_K(\Gamma(I)) \bigcap \Phi_K(\Gamma(J)) \tag{1}$$

are likely to be not discriminative for our classification task. On the other hand the frequencies in

$$\Omega_K(I,J) = (\Phi_K(\Gamma(I)) \bigcup \Phi_K(\Gamma(J))) \setminus \Psi_K(I,J) \tag{2}$$

are likely to be useful for the recognition of the signature of each class because they belong only to one of the two sets $\Phi_K(\Gamma(I))$, $\Phi_K(\Gamma(J))$.

Figure 3 shows the $\Psi_K(I,J)$ and $\Omega_K(I,J)$ sets (K=1000) relative to the image pair in Figure 1. The value K=1000 has been choosen only as an example. Automatic selection of this parameter is thoroughly discussed in Section 2.1.

Frequencies in $\Omega_K(I,J)$ are called *disciminative frequencies*.

It is straightforward to generalize the $\Omega_K(I,J)$ to capture the shape of the whole classes of *Natural* and *Artificial* scenes by an averaging process. Let $\boldsymbol{N}$ and $\boldsymbol{A}$ be two ensembles of images respectively in the *Natural* and *Artificial*

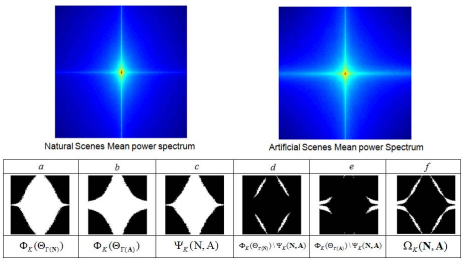

Fig. 4. Discriminative frequencies: (d) the frequencies selected from the *Natural* power spectrum span mostly "diagonally" whereas (e) the frequencies selected from the *Artificial* power spectrum span mostly "horizontally"

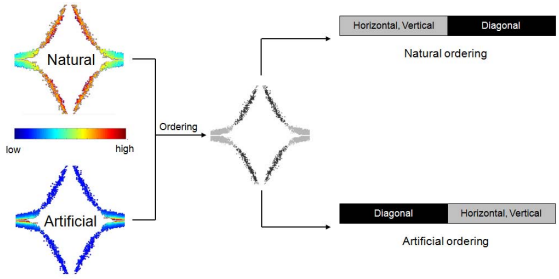

Fig. 5. Discriminative frequencies are reported on the left side. The magnitudes are color coded. In the middle, "diagonal" frequencies are represented in black while "horizontal" frequencies are represented in gray. On the right the *discriminative frequencies*, coded with the same colors than in the middle, are ordered. Separation properties characterizing the *Natural* vs. *Artificial* classes are evident in this schematic graphic diagram. The probability to find a "diagonal" frequency in the lower part of the ordering is low (high) for the *Natural* (*Artificial*) class. The probability to find a "horizontal" and "vertical" frequency in the lower part of the ordering is high (low) for *Natural* (*Artificial*) class.

classes. We indicate with $\Theta_{\Gamma(\boldsymbol{N})}$ and $\Theta_{\Gamma(\boldsymbol{A})}$ the mean power spectra obtained by respectively averaging over the power spectra of the two ensembles $\boldsymbol{N}$ and $\boldsymbol{A}$. $\Psi_K(\boldsymbol{N}, \boldsymbol{A})$ and $\Omega_K(\boldsymbol{N}, \boldsymbol{A})$ may hence be defined as:

$$\Psi_K(\boldsymbol{N}, \boldsymbol{A}) = \Phi_K(\Theta_{\Gamma(\boldsymbol{N})}) \bigcap \Phi_K(\Theta_{\Gamma(\boldsymbol{A})}) \tag{3}$$

$$\Omega_K(\boldsymbol{N}, \boldsymbol{A}) = (\Phi_K(\Theta_{\Gamma(\boldsymbol{N})}) \bigcup \Phi_K(\Theta_{\Gamma(\boldsymbol{A})})) \setminus \Psi_K(\boldsymbol{N}, \boldsymbol{A}) \tag{4}$$

In Figure 4 the mean power spectra of the two classes, *Natural* and *Artificial* are reported. The mean power spectrum of the *Natural* class has been obtained averaging the spectra of images within the basic classes *Forest*, *Coast*, *Mountain* and *Open Country*. The mean power spectrum of the *Artificial* class has been obtained averaging the spectra of images within the basic classes *Building*, *Street*, *Highway* and *City*. Figure 4 confirms the *Statement-II* of this section.

The classification scheme can benefit from the corresponding relative order between the selected *discriminative frequencies* (Figure 5). In particular the following two cases can be considered:

- *The image is Natural*: "horizontal" and "vertical" *discriminative frequencies* typically rank before "diagonal" *discriminative frequencies* in the magnitude ascendent ordering;
- *The image is Artificial*: "diagonal" *discriminative frequencies* typically rank before "horizontal" and "vertical" *discriminative frequencies* in the magnitude ascendent ordering.

Our classification model should capture the fact that in the ascending ordering of the power spectrum of a *Natural* (*Artificial*) image there is high (low) probability that the ascendent ordering position of a "diagonal" frequency is higher (lower) than the ascendent ordering position of "horizontal" and "vertical" frequencies.

The above rationale suggest that the ordered *discriminative frequencies* provide an effective feature set for *Natural* vs. *Artificial* classification, indeed it is able to capture the discriminative shape signature in frequency domain. In some sense the ranking order of a *discriminative frequency* $f_{x,y}$ can be thought as the context in which $f_{x,y}$ lie with respect to the other *discriminative frequencies*. This capture the essence of the *Statement-III* of this section.

Let $\boldsymbol{r}$ be the vector of the relative position order of the *discriminative frequencies* after their selection and let $\boldsymbol{s}$ be the corresponding vector of their magnitudes. We use these two feature vectors alone or in combination to holistically represent the scene.

2.1 Discriminative Frequencies Selection

A key point in the construction of our holistic representation is the selection of the discriminative frequencies. This selection, as pointed above, is parameterized by the number K of the highest ranking frequencies in the mean power spectra. The selection process is hence reduced in choosing a suitable value for K. The idea is to select a maximal set of *discriminative frequencies*. We summarize the selection process as follows:

1. Compute the mean power spectrums $\Theta_{\Gamma(\boldsymbol{N})}$ and $\Theta_{\Gamma(\boldsymbol{A})}$ of the two classes using respectively the *Natural* and *Artificial* training images;
2. Sort $\Theta_{\Gamma(\boldsymbol{N})}$ and $\Theta_{\Gamma(\boldsymbol{A})}$ and to each frequency $f_{x,y}$ in the mean spectrums assign the ranking positions $Pos(f_{x,y}$,*Natural*) and $Pos(f_{x,y}$,*Artificial*);
3. Let $\Phi_n(\Theta_{\Gamma(\boldsymbol{N})})$ and $\Phi_n(\Theta_{\Gamma(\boldsymbol{A})})$ be the set of the first n frequencies in the descending ordering. Compute the set of the best *discriminative frequencies* $\Omega_K(\boldsymbol{N}, \boldsymbol{A})$ such that $K = \arg\max_n(\mid \Omega_n(\boldsymbol{N}, \boldsymbol{A}) \mid)$.

In Figure 6 the selection process relative to the mean power spectra of Figure 4 is illustrated. The discriminative set of frequencies, corresponding to the global maximum of the function $\mid \Omega_n(\boldsymbol{N}, \boldsymbol{A}) \mid$, is selected. Clearly, $\mid \Omega_K(\boldsymbol{N}, \boldsymbol{A}) \mid$, which is the number of *discriminative frequencies*, is smaller than K (which is the

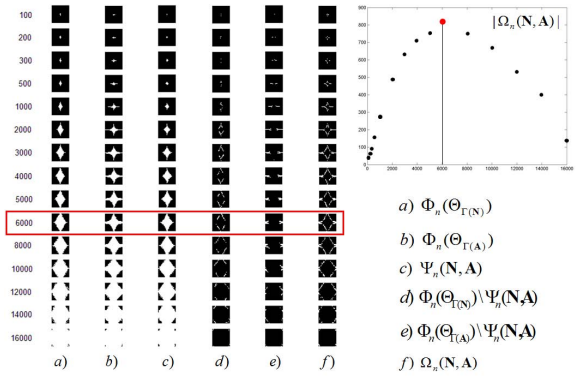

Fig. 6. Discriminative frequencies selection

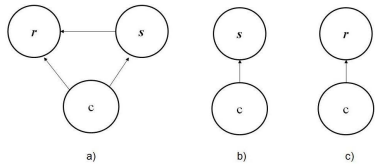

Fig. 7. The Generative Models used for the classification task. In the cnfiguration a) the power spectrum of the *discriminative frequencies* (node $\boldsymbol{s}$) and their ordering (node $\boldsymbol{r}$) are depending from the class of an image (node $\boldsymbol{c}$). Moreover, the ordering (node $\boldsymbol{r}$) depends from the magnitude of the power spectra (node $\boldsymbol{s}$). The configuration b) takes into account only the magnitude of the power spectrum, whereas the configuration c) takes into account only the ordering position.

maximum order considered). In Figure 6, we can see that $\mid \Omega_n(\boldsymbol{N}, \boldsymbol{A}) \mid$ is less than 850 for $K = 6000$.

2.2 Classification Model

Given the holistic representation of a new unclassified image I, a simple generative model together with a MAP classification rule can be used as a framework to establish which category c$\in\{$*Natural*, *Artificial*$\}$ the image I belongs to. Specifically, we use three generative models as shown in Figure 7. Let be $f_{x,y}^{(1)} \ldots f_{x,y}^{(H)}$ the H pre-selected frequencies in $\Omega_K(\boldsymbol{N}, \boldsymbol{A})$. For each $f_{x,y}^{(i)}$ the values $\boldsymbol{s}_i$ and $\boldsymbol{r}_i$ refer respectively to the power spectrum value and its ordering position.
When the spectrum magnitudes and the ordering positions of *discriminative frequencies* are used in combination as holistic representation (Figure 7.a), the classification problem requires the evaluation of the class likelihood function P(c$|\boldsymbol{s},\boldsymbol{r}$). This posterior probability can be factorized as follows:

$$P(c|\boldsymbol{s},\boldsymbol{r}) = \frac{P(\boldsymbol{r}|\boldsymbol{s},c)P(\boldsymbol{s}|c)}{\sum_c P(\boldsymbol{r}|\boldsymbol{s},c)P(\boldsymbol{s}|c)} \tag{5}$$

where equiprobability is assumed for the prior P(c).

The meanings of the factors involved in the model are:

- *Power Spectrum Probability*: $P(\boldsymbol{s}|\mathrm{c})$ gives the most likely spectrum values given a scene category.
- *Relative Position Probability* : $P(\boldsymbol{r}|\boldsymbol{s},\mathrm{c})$ gives the most likely ordering for a scene category given power spectrum information;

When the spectrum magnitudes and the ordering positions of *discriminative frequencies* are used separately as holistic representation (Figure 7.b and 7.c) the posterior probability can be obtained by using Bayes Theorem. In such case we use respectively the *Power Spectrum Probability* or the *Relative Position Probability* when spectrum magnitudes features or the ordering positions are involved.

In the proposed classification frameworks, the *Power Spectrum Probability* and the *Relative Position Probability* are modeled as H-dimensional multivariate gaussian distribution:

$$P(\boldsymbol{s}|c) = \prod_{i=1}^{H} \frac{1}{\sqrt{2\pi\sigma_{i,c}}} \exp\left[-\frac{1}{2}\left(\frac{s_i - \mu_{i,c}}{\sigma_{i,c}}\right)^2\right] \tag{6}$$

$$P(\boldsymbol{r}|\boldsymbol{s},c) = \prod_{i=1}^{H} \frac{1}{\sqrt{2\pi H}} \exp\left[-\frac{1}{2}\left(\frac{r_i - Pos(f_{x,y}^{(i)},c)}{H}\right)^2\right] \tag{7}$$

For the distribution of *Power Spectrum Probability* $\sigma_{i,c}$ are obtained considering the frequency $f_{x,y}^{(i)}$ of all the training images belonging to the class c. For the distribution of the *Relative Position Probability* the i-th gaussian is centered at the relative position $Pos(f_{x,y}^{(i)},\mathrm{c})$ that assumes the frequency $f_{x,y}^{(i)}$ in the ordering of the mean power spectrum $\Theta_{\Gamma(c)}$. In $P(\boldsymbol{r}|\boldsymbol{s},c)$ all the standard deviations are set to be equal to H, which is the maximum absolute difference that can be observed between r_i and $Pos(f_{x,y}^{(i)},\mathrm{c})$.

3 Dataset, Experimental Setup and Results

The experiments described in this section are aimed to demonstrate that the relative order of the discrete Fourier power spectra can be a useful feature for *Natural* vs. *Artificial* scene classification. The database used in our tests is composed of eight basic scene categories collected by Oliva and Torralba [4]: *Coast*, *Forest*, *Open Country*, *Mountain*, *Highway*, *Building*, *City*, *Street*. The first four basic classes have been considered as *Natural* scenes whereas the other classes have been considered *Artificial*. The overall database contains 2360 labeled images of 256×256 pixel in size. covering a large variety of *Natural* and *Artificial* outdoor places. As in [4], strongly ambiguous scenes in term of *naturalness* were discarded. Images have been converted in gray scale after performing histogram equalization on color domain. A logarithmic transformation of the discrete Fourier power spectrum of each image have been applied normalizing

the spectrum in [0,1]. All experiments have been repeated ten times with different randomly selected training (75%) and test images (25%). The parameters involved in the classification models have been learned from the training set at each run (see Section 2) and the measures of Accuracy, Recall and Specificity were recorded at each run.

In Table 1, 2, 3 the measures of classification performances obtained at each run are reported. The results are shown with respect to the three proposed holistic representation: magnitude $\boldsymbol{s}$, relative position order $\boldsymbol{r}$, magnitude and relative position order $(\boldsymbol{r},\boldsymbol{s})$.

Table 1. Percentage of test images that were classified correctly

ACCURACY	1	2	3	4	5	6	7	8	9	10	Average
$\boldsymbol{s}$	81.44	83.26	84.46	86.16	83.46	81.10	84.65	80.80	84.55	84.72	**83.46**
$\boldsymbol{r}$	93.50	92.97	90.77	93.09	93.92	92.38	94.14	91.08	92.65	92.28	**92.68**
$(\boldsymbol{r},\boldsymbol{s})$	92.57	91.27	92.53	91.44	92.81	90.72	93.00	90.61	93.90	93.05	**92.19**

Table 2. Percentage of *Natural* test images that were classified correctly

RECALL	1	2	3	4	5	6	7	8	9	10	Average
$\boldsymbol{s}$	97.98	96.87	98.54	98.20	98.67	98.98	97.78	97.53	97.70	98.96	**98.12**
$\boldsymbol{r}$	93.46	92.31	88.46	92.76	93.73	92.38	93.03	88.84	90.80	90.10	**91.59**
$(\boldsymbol{r},\boldsymbol{s})$	96.47	94.88	95.66	93.58	98.53	96.23	94.61	95.42	98.09	96.68	**96.02**

Table 3. Percentage of *Artificial* test images that were classified correctly

SPECIFICITY	1	2	3	4	5	6	7	8	9	10	Average
$\boldsymbol{s}$	55.01	63.02	62.54	67.55	58.25	53.88	63.21	55.57	60.53	62.47	**60.20**
$\boldsymbol{r}$	93.57	93.95	94.38	93.60	94.19	92.38	95.97	94.46	95.41	95.68	**94.36**
$(\boldsymbol{r},\boldsymbol{s})$	86.34	85.90	87.65	88.13	84.42	82.33	90.36	83.35	87.66	87.38	**86.35**

First, let us examine the performance when the magnitude of the *discriminative frequencies* is considered as holistic representation. The average accuracy rate is 83.46%, which is no higher than the best results obtained by the state of the art methods working on frequency domain. As shown in Table 2 and Table 3, by using the magnitude of the *discriminative frequencies* there is an average of 40% of false positive whereas the percentage of false negative is very low. This means that the classifier is not robust in recognizing *Artificial* scenes.

Next, let us examine the performances obtained by considering the order position of the *discriminative frequencies*. As shown in Table 1, the accuracy is greater than 90% in all tests. The average accuracy is 92.68%. In Figure 8 some images correctly classified by using the order position as holistic representation are reported according their increasing posterior probability. As shown by recall and specificity measures (Table 2 and Table 3) in this case the classifier show low percentage of misclassified test images with respect to a specific class. The average classification accuracy of the proposed method (92.68%) close match the results of state of art approaches working in the frequency domain (93.5%) [4].

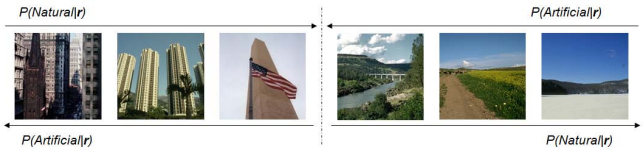

Fig. 8. Posterior Probability: On the left side images correctly classified as *Artificial* scenes. On the right side images correctly classified as *Natural* scenes. The images are ordered respect to the probability of belonging to the *Artificial* or *Natural* Class.

Finally, it is interesting to observe that a combination of the magnitude and the order position of the *discriminative frequencies* do not improve the classification performances. Indeed, recognition of the *Artificial* scenes decreases respect to considering the case when the order position is used alone (Table 1,2,3).

4 Conclusion

This paper has presented a new approach for *Naturalness* Classification of scenes based on ordering Discrete Fourier Power Spectra. The proposed method works by selecting a set of *discriminative frequencies*, building an holistic representation mainly based on the frequencies ordering and using a simple probabilistic model for classification. The experiments prove that the achieved performance match the state-of-art methods working on frequency domain. Future work on this technique requires the addressing of the problem related to the classification where image resolution is not fixed and other classes are taken into account.

References

1. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Visual Perception, Progress in Brain Research, 251–256 (2006)
2. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. II, pp. 2169–2178 (2006)
3. Renninger, L.W., Malik, J.: When is scene recognition just texture recognition? Vision Research 44, 2301–2311 (2004)
4. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175 (2001)
5. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proceedings of the European Conference on Computer Vision (2006)
6. Torralba, A., Oliva, A.: Statistics of natural image categories. Network: Computing in Neural Systems 14, 391–412 (2003)
7. Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Trans. Pattern Anal. Mach. Intell. 24(9), 1226–1238 (2002)
8. Torralba, A., Pawan, S.: Statistical context priming for object detection. In: Internation Conference on Computer Vision (2001)
9. Belongie, S., Malik, J., Puzicha, J.: Shape context: A new descriptor for shape matching and object recognition. In: Neural Information Processing Systems (2000)