Image watermarking capacity analysis based on Hopfield neural network

Fan Zhang (张 帆) and Hongbin Zhang (张鸿宾)

The College of Computer Science, Beijing University of Technology, Beijing 100022

Received April 19, 2004

In watermarking schemes, watermarking can be viewed as a form of communication problems. Almost all of previous works on image watermarking capacity are based on information theory, using Shannon formula to calculate the capacity of watermarking. In this paper, we present a blind watermarking algorithm using Hopfield neural network, and analyze watermarking capacity based on neural network. In our watermarking algorithm, watermarking capacity is decided by attraction basin of associative memory.

OCIS codes: 100.2000, 100.2960.

Capacity is a very important property of digital watermarking. The purpose of watermarking capacity research is to analyze the limit of watermark information while satisfying watermarking invisibility and robustness.

Several works on watermarking capacity have been presented in recent years. Servetto considered each pixel as an independent channel and calculated the capacity based on the theory of parallel Gaussian channels (PGC)[1]. Barni's research focused on the image watermarking capacity in discrete cosine transform (DCT) and discrete Fourier transform (DFT) domain[2]. Moulin's work studied a kind of watermarking capacity problem under attacks[3,4]. Lin presented zero-error information hiding capacity analysis method in joint photographic experts group (JPEG) compressed domain using adjacency-reducing mapping technique[5,6].

Almost all of previous works on image watermarking capacity are based on information theory, using Shannon formula to calculate the capacity of watermarking. In this paper, we present a blind digital watermarking algorithm using Hopfield neural network, and analyze watermarking capacity based on neural network.

Hopfield neural network is a nonlinear dynamical system, using computational energy function to evaluate the stability property. The energy function always decreases toward a state of lowest energy. Starting from any point of state space, system always evolves to a stable state, an attractor. The set of initial conditions, which initiates the evolution terminating in the attractor, is the basin of attraction.

Neurons state of Hopfield neural network are usually binary $\{+1, -1\}$. For the sake of neural network to store a standard grayscale test image, we decompose image into eight bit planes. For each pixel, we decide whether the pixel should be modified randomly by using a random function. Then we can get a matrix that composes of $\{0,1\}$, 0 denotes not modifying this pixel and 1 denotes modifying this pixel. We name the matrix watermark bit plane (WBP). Actually, the WBP marks the points of watermark embedding. Watermark amplitude is decided as follow.

The noise visibility function (NVF)[7] is the function that characterizes local image properties, identifying textured and edge regions where the watermark should be more strongly embedded. Assuming that the host image subjects to generalized Gaussian distribution, the NVF at each pixel position can be written as

$$\mathrm{NVF}(i,j) = \frac{1}{1+\sigma_x^2(i,j)}, \tag{1}$$

where $\sigma_x^2(i,j)$ is the local variance of the image in a window centered on the pixel with coordinates (i,j). Once we computed the NVF, we can obtain the allowable distortions of each pixel by computing,

$$\Delta(i,j) = [1-\mathrm{NVF}(i,j)]\cdot S_0 + \mathrm{NVF}(i,j)\cdot S_1, \tag{2}$$

where S_0 and S_1 are the maximum allowable distortions in textured and flat region, respectively. Typically S_0 may be as high as 30 while S_1 is usually about 3. In flat regions the NVF tends to 1, so the first term of Eq. (2) tends to 0, and consequently the allowable pixel distortion is dependent on S_1. Intuitively this makes sense since we expect that the watermark distortions will be visible in flat regions and less visible in textured regions. According to above equation, the watermark embedded in the texture or the edge regions is stronger than that in the flat regions. If we embed maximum allowable watermark in each pixel, the robustness of watermarking will have a good performance. By this way, we can achieve the best trade-off between robustness and invisibility.

During the learning of neural network, both image bit planes and watermark bit plane are stored by Hopfield neural network. In watermark extracting, network recalls the host (original) image and WBP, and then we compare the retrieved host image with stego (watermarked) image, extract watermark according a threshold. Finally, comparing retrieved WBP with extracted watermark, we can judge whether watermark exists in the image using correlation test.

In watermarking schemes, watermarking can be considered as a form of communications, image can be considered as a communication channel to transmit messages, and the watermark is the message to be transmitted[8]. So, watermarking capacity problem can be solved using traditional information theory.

Considering image as an additive white Gaussian noise (AWGN) channel, and assuming P_{S} denotes watermark power constraint and P_{N} denotes noise power constraint,

then we apply Shanon channel capacity formula. Watermarking capacity in non-blind watermarking scenario is

$$C = \frac{1}{2}\log_2\left(1 + \frac{P_S}{P_N}\right). \quad (3)$$

Almost all of previous works on image watermarking capacity are based on this formula.

Hopfield neural network functions as an associative memory, the patterns are stored as dynamical attractors, and the network has error-correcting capability. If initialized with a corrupted or incomplete pattern, the network will be guessed and relaxed to the nearest stored pattern. The radius of attraction basin is defined as the largest Hamming distance within which almost all states flow to the pattern. For a trained network, the average attraction radiuses of stored patterns give a measure of the network completion capability.

The Hamming distance of two vectors S^1 and S^2 is the number of components and different from each other. We denote it as $d_h(S^1, S^2)$. The total number of vectors which Hamming distance to vector S^p is less than r constitutes a r-Hamming sphere

$$B_r(S^p) = \{S^q | d_h(S^p, S^q) = r\}. \quad (4)$$

Basin of attraction can be denoted by Hamming distance or Hamming sphere, which represents the error-correcting capability of neural network.

Assume P denotes the number of stored patterns and Ndenotes the number of neurons. For Hopfield neural network, when stored patterns are orthogonal, the basin of attraction of each stored pattern is

$$d_h \le \frac{N-P}{2P}. \quad (5)$$

Assume the attraction basin of each stored pattern is decided by d_h. When the probe pattern's Hamming distance away from the stored pattern is less than Eq. (5), probe pattern that is one of stored patterns will be attracted to a stored pattern, otherwise, associative recall fails.

In watermarking schemes, watermark is viewed as noise. Assuming watermark can be embedded in every pixel, eight bit planes of host images are stored patterns in our watermarking algorithm. If we modify amplitude of some pixels, some changes occur in corresponding place on bit planes. This means that those bit planes are polluted. The more watermark is embeded, the bigger Hamming distance between stego image bit planes with host image bit planes is. When the Hamming distance is out of the bound of attraction basin, neural network cannot recall host image bit planes correctly. So, the bound of attraction basin confines the number of points in the image that can be modified, furthermore, confines the capacity of watermarking.

Equation (5) is derived when stored patterns are orthogonal. Obviously, if stored patterns are nonorthogonal, basin of attraction is less than Eq. (5). For analyzing watermarking capacity and the limit of information, we assume stored patterns are orthogonal, which is reasonable. It is just like we assume information source and noise to be subjected to Gaussian distribution during channel capacity analysis in information theory.

According to our watermarking algorithm, there are nine probe patterns, $P = 9$. Then, according to Eq. (5), the attraction basin of each stored pattern $d_h \le 3640$. If number of modified points of a bit planes is less than d_h, neural network can associatively recall the bit planes successfully.

Modified a pixel may results in several bit planes's change in corresponding point. If a bit plane is modified, corresponding pixel is modified. Contrarily, if the point of a bit plane is not modified, we cannot affirm that this pixel was not embedded watermark, maybe other bit planes of this pixel are modified.

Because the maximum watermark amplitude is no more than 30, at least one of five low bit planes is modified in corresponding point when embedding watermark in a pixel. As the analysis above, the maximum number of modifiable point is 3640. In an extreme case, modified place in five low bit planes may be different to each other, then the maximum number of watermarking pixels is $n = 5\times3640 = 18200$. In our watermarking algorithm, watermark is a binary sequence, and we assume the length is n. State of each component in the sequence is random. An n-length binary sequence has 2^n combination in all, each combination appears in probability $1/2^n$. According to information theory, information of an n-length binary sequence is

$$C = -\log_2\left(\frac{1}{2^n}\right). \quad (6)$$

So, in our watermarking algorithm, watermarking capacity is 18200 bits.

In experiments, three 256×256 standard test images Baboon, Peppers, and Lena are used. Discrete Hopfield neural network is used in our experiments. Assume noise is white Gaussian noise which variance equals to 4.

If the number of neurons equals to the number of pixels, the computation will be very complex. So we should reduce the dimension of neural network to reduce the commutative complexity. The NVF is calculated in local image region, a window of which is centered on a pixel. We divide image into many non-overlap regions according to the window size of the NVF. Each region corresponds to a neuron. If the NVF calculation region is a window of size 5×5, then a 256×256 image can be divided into 51×51 regions, and the dimension of neural network is 51×51.

According to Eq. (2), we can calculate the maximum allowable watermark amplitude value of each pixel while keeping watermark's invisibility. All watermark amplitude value can build an image, named maximum watermark image (MWI). Figure 1 shows Lena's MWI.

In Fig. 1, the bright parts are the regions that allow bigger watermark amplitude value; the dark parts are the regions that allow smaller watermark amplitude value. In Fig. 1, an approximate outline of Lena can be identified. In the complex texture regions or in the edge regions, for example, the regions of Lena's hair, the bigger watermark amplitude value is allowed.

In watermark extracting, we assume that the probe

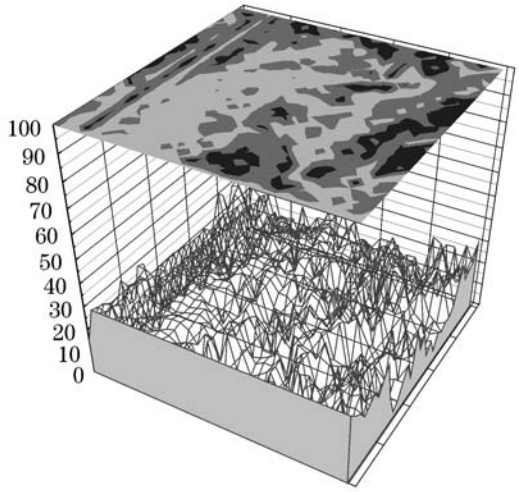

Fig. 1. Lena image's maximum watermark image.

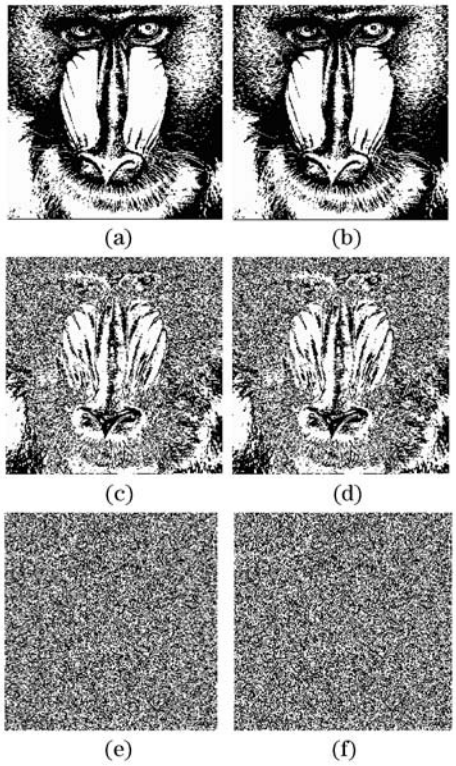

Fig. 2. Original Baboon image bit planes (a), (c), (e) and retrieved image bit planes (b), (d), (f).

Table 1. Time of Positive Detection (Presence of Watermark)

	Lena	Baboon	Peppers
Stego Image	100	100	100
Noised Stego Image	100	100	100
Host Image	0	0	0
Noised Host Image	1	0	0

patterns are stego image, noised stego image, host image, and noised host image, respectively. We experimented 1200 times, one hundred times in above condition for each test image. Table 1 shows the results of our experiments. Figure 2 shows original Baboon image bit planes and retrieved image bit planes.

In conclusion, we present a blind watermarking algorithm based on Hopfield neural network, using the NVF for adaptive watermark embedding. And we analyze watermarking capacity based on neural network. In our watermarking algorithm, watermarking capacity is decided by attraction basin of associative memory.

This work was supported by the National Natural Science Foundation of China (No. 60075002) and the National High Technology Research and Development 863 Program of China (No. 2001AA144080). F. Zhang's e-mail address is justzf@sina.com.

References

1. S. D. Servetto, C. I. Podilchuk, and K. Ramchandran, in *Proceedings of IEEE International Conference on Image Processing* **1,** 445 (1998).
2. M. Barni, F. Bartolini, and A. De Rosa, Proc. SPIE **3657,** 437 (1999).
3. P. Moulin and M. K. Mihcak, IEEE Trans. Image Processing **11,** 1029 (2002).
4. P. Moulin, Signal Processing **81,** 1121 (2001).
5. C. Y. Lin and S. F. Chang, in *Proceedings of IEEE International Conference on Image Processing* **3,** 1007 (2001).
6. C. Y. Lin, Ph.D. Thesis (Columbia University, 2000).
7. S. Voloshynovskiy, S. Pereira, and V. Iquise, Signal Processing **81,** 1177 (2001).
8. J. Cox, M. L. Milller, and A. McKellips, Proceedings of the IEEE (Special Issue on Identification and Protection of Multimedia Information) **87,** 1127 (1999).