

Entropy of Primitive: From Sparse Representation to Visual Information Evaluation

Siwei Ma, *Member, IEEE*, Xiang Zhang, Shiqi Wang, Jian Zhang, Huifang Sun, *Fellow, IEEE* and Wen Gao, *Fellow, IEEE*

Abstract—In this paper, we propose a novel concept in evaluating the visual information when perceiving natural images—the entropy of primitive (EoP). Sparse representation has been successfully applied in a wide variety of signal processing and analysis applications due to its high efficiency in dealing with rich, varied and directional information contained in natural scenes. Inspired by this observation, in this paper the visual signal can be decomposed into structural and non-structural layers according to the visual importance of sparse primitives. Accordingly, the entropy of primitive is developed in measuring visual information. It has been found that the EoP changing tendency in image sparse representation is highly relevant with the hierarchical perceptual cognitive process of human eyes. Extensive mathematical explanations as well as experimental verifications have been presented in order to support the hypothesis. The robustness of EoP is evaluated in terms of varied block sizes. The dictionary universality is also studied by employing both universal and adaptive dictionaries. With the convergence characteristics of EoP, a novel top-down just-noticeable difference (JND) profile is proposed. Simulation results have shown that the EoP based JND outperforms the state-of-the-art JND models according to the subjective evaluation.

Index Terms—Entropy of primitive, sparse representation, orthogonal matching pursuit, visual information, just noticeable difference

I. INTRODUCTION

THE human visual system (HVS) allows human beings to perceive visual information from the outside world, and the psychological process of visual information is known as visual perception. As the ultimate receiver of images and videos is the HVS, accurately evaluating visual information plays an important role in the fields of image and video processing. Generally, both near-threshold and supra-threshold quality assessment models are highly relevant with the perceptual cognition. It is widely believed that the HVS is not able to perceive the variations of visual information when the distortion is below a threshold, and this threshold is referred

This work was supported in part by the National Basic Research Program of China (973 Program, 2015CB351800), National Natural Science Foundation of China(61322106, 61572047, 61571017), Postdoctoral Science Foundation of China (2015M580018) and Shenzhen Peacock Plan, which are gratefully acknowledged.

S. Ma, X. Zhang, S. Wang, J. Zhang and W. Gao are with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China. S. Ma and W. Gao are also with the Cooperative Medianet Innovation Center (e-mail: {swma,x_zhang,sqwang,jian.zhang,wgao}@pku.edu.cn).

H. Sun is with Mitsubishi Electric Research Laboratories, USA (e-mail: hsun@merl.com).

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

as the just-noticeable difference (JND). Image with distortions below the threshold conveys approximately equal visual information to the original image for subjects. Recently, many near-threshold models [1]–[4] have been proposed to generate the JND profile for images and videos. However, very few studies have shown that these near-threshold models can be successfully applied to characterize perceptual distortions beyond the threshold level. In view of this, many supra-threshold image quality assessment algorithms have been proposed to evaluate the perceivable visual information loss [5]. However, as the limited information known about HVS, neither state-of-the-art near-threshold nor supra-threshold models has built an effective system that performs as the HVS in estimating visual information.

The near-threshold or typically the JND models have been successfully applied in many applications, especially in image/video coding [2], [4], [6] and quality assessment [7]. The JND models can be classified into two categories, i.e. the spatial-domain JND and the transform-domain JND [7]. Both of them take advantages of visual characteristics of the HVS, including contrast sensitivity function (CSF) [8], luminance adaptation, and texture masking. Therefore, these kinds of methodologies can be referred to as the “bottom-up” methods [5], which are simulating the functionalities of the HVS.

In supra-threshold models, the popular mean square error (MSE) and peak signal-to-noise ratio (PSNR) metrics have been successfully employed in many realtime applications due to its calculational efficiency. However they cannot correlate well with the visual quality [9], [10]. It has motivated many researchers to involve in developing more accurate image quality assessment algorithm, such as structural similarity (SSIM) [11], visual information fidelity [12], feature similarity [13] and free-energy principle [14]. Recently, it has been found that the natural image is highly structured and the HVS can adapt to the structural information in natural scenes. Therefore the SSIM and its variants [15]–[17], have been well advocated in both academic and industry endeavors due to its best tradeoff among accuracy and efficiency.

Sparse representation is an emerging and powerful method to describe signals based on the sparsity and redundancy of their representations [18] and is efficient in dealing with rich, varied and directional information contained in natural scenes [19]. In the literature [20], it is also interesting to find that the primitive, or the basis in sparse representation has the properties of spatially localized, oriented and bandpass, which closely correspond to the characteristics of receptive fields of simple cells. To perform sparse representation, an over-

complete dictionary should be obtained first, where the typical K-SVD [21] algorithm is employed for dictionary training. The well-known matching pursuit family (MPF) algorithms have been proposed for solving the sparse decomposition problem [22]. The orthogonal matching pursuit (OMP) [23] is one of the MPFs that works in a greedy fashion. Extensive new algorithms have been developed by taking advantage of the sparse characteristics to achieve the state-of-the-art performance in a wide variety of applications, such as image quality assessment [24]–[30], image denoising [31], image restoration [32]–[35] and image/video coding [36]–[38].

Inspired by these works, we employ the technique of sparse representation to interpret the perceptual cognition process by decomposing the natural scene into structural and non-structural layers. Along this vein, we further study the inverse problem: how much information is contained or preserved in the visual scene in terms of sparse decomposition? To address this issue, instead of using numerical measures such as SSIM, we analyze the visual scene based on the novel study of visual perception, the *internal generative mechanism (IGM)*. In *IGM*, the brain optimizes the visual scene by predicting the primary information [39], which is also called structural information in this paper. The structural and non-structural visual signals are distinguished by using different types of primitives to best match the original visual signal. Motivated by the relationship between the responses of neurons or sparse primitives, we develop a novel concept in evaluating the visual information: entropy of primitive (EoP). The basic idea of the EoP is to employ the information conveyed by primitives to characterize how much surprise there is when encountering a visual scene. Interestingly, it has been found that the EoP can efficiently estimate the visual information in perceptual cognitive process. When the structural information achieves saturation, the EoP curve tends to be stable and the reconstructed image reaches the state without noticeable visual distortions. It motivates us to propose a “top-down” JND model with the principle of EoP based hierarchical visual signal representation.

Above all, the major contributions in this work include:

- The sparse primitives are classified into three groups to characterize the inherent property, and the visual signal is accordingly separated into several layers with different visual importance. The structural and non-structural visual information can be well represented with these layers.
- The entropy of primitive (EoP) is proposed to measure the visual information in the natural scenes, supported by complete mathematical analysis and experimental verifications.
- With the concept of the EoP, a novel “top-down” JND profile is proposed and outperforms state-of-the-art JND models according to the subjective results.

In this paper, 50 images drawn from public databases are utilized as shown in Fig. 1.

The rest of the paper is organized as follows. In Section II, the visual perception is interpreted as a hierarchical signal decomposition based on the sparse representation. In Section III, the EoP is proposed to evaluate the visual information. Mathematical interpretations and experimental verifications



Fig. 1. 50 test images used in this work.

are presented to demonstrate the reliability and robustness of the EoP. We also propose an EoP based approach to generate the JND profile in Section IV. Section V concludes this paper and emphasizes the future works.

II. HIERARCHICAL VISUAL SIGNAL SPARSE REPRESENTATION

A. Sparse Representation Dictionary Learning

Image primitive coding is based on the Sparseland model, which assumes that natural images $x(x \in \mathbb{R}^n)$ can be approximated by a linear combination over an over-complete dictionary. Put more formally, this can be written as $\forall x, x \approx \Psi\alpha$ and $\|\alpha\|_0 \ll n$, where $\Psi(\Psi \in \mathbb{R}^{n \times k})$ is the over-complete dictionary, and $\alpha(\alpha \in \mathbb{R}^k)$ is the representation vector. The notation $\|\bullet\|_0$ represents the l_0 norm. Typically, we assume that $k > n$, implying the dictionary Ψ is redundant to x .

In order to train the over-complete dictionary, the K-SVD algorithm [21] is employed in this work, which consists of two process, the sparse coding and dictionary updating. The training samples are generated by partitioning the input image X into non-overlapped patches $x_1, x_2, \dots, x_i, \dots$. Assuming a local Sparseland model on image patches, given the set of training samples, the dictionary that leads to the best representation of the training samples under the sparsity constraint is generated. This process can be formulated as follows,

$$(\Psi, \{\alpha_i\}) = \arg \min_{\Psi, \{\alpha_i\}} \sum_k \|x_i - \Psi\alpha_i\|_2^2, \text{s.t. } \|\alpha_i\|_0 < \mathbb{L}, \quad (1)$$

where \mathbb{L} controls the sparse level.

Fig. 2 illustrates an example of the dictionary trained by the 8×8 patches partitioned from *Lena* image. In this paper, the atom of dictionary Ψ is called the primitive, denoted by ψ_i hereinafter.

B. Sparse Primitives Classification

The receptive fields of simple cells in human primary visual system can be characterized as being spatially localized, oriented and bandpass [20]. Sparse coding has been shown to be an effective representation approach in accounting for these properties. In this subsection, we first employ a strategy to divide the sparse primitives into multiple categories in terms of their inherent properties regarding to the perceptual information.

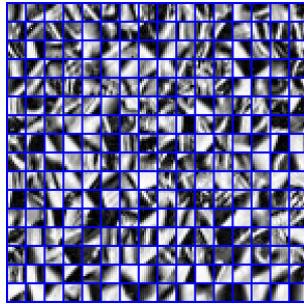


Fig. 2. Illustration of 256 primitives trained by non-overlapped 8×8 patches from *Lena* image.

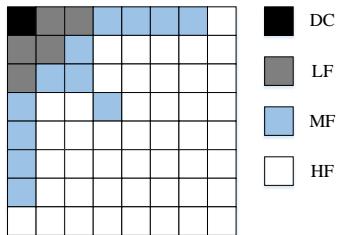


Fig. 3. DCT coefficients classification [40].

In [40], an image patch can be classified into distinct types depending on its energy distribution in frequency domain, where several thresholds are predefined for clustering. It has been widely recognized that the Discrete Cosine Transform (DCT) is highly effective in decomposing image signal to different frequency sub-bands, as illustrated in Fig. 3. In this work, we apply k-means cluster algorithm to adaptively divide primitives into different categories, namely the *primary*, *sketch* and *texture*, respectively. For dimensional reduction, we extract two features from the DCT domain and one feature from spatial domain, and employ these features in primitive classification.

The two DCT domain features are defined as follows,

$$f_1 = \overline{LF}, \quad (2)$$

$$f_2 = \overline{LF} / (\overline{MF} + \overline{HF}), \quad (3)$$

where \overline{LF} , \overline{MF} and \overline{HF} refer to the mean values of LFs, MFs and HFs in Fig. 3, respectively.

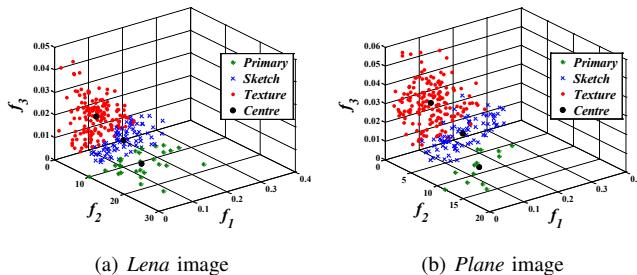


Fig. 4. 3D plots of primitive classification using k-means algorithm. (The red dots, blue crosses and green stars represent texture, sketch and primary primitives, respectively. The three black solid dots denote the central points of each classification.)

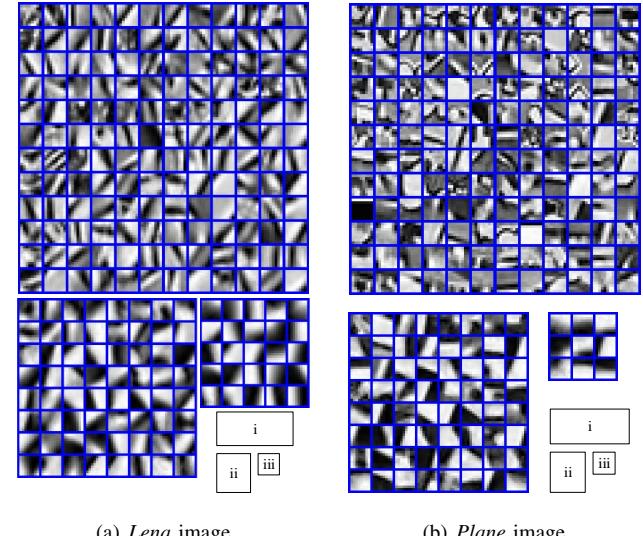


Fig. 5. Primitives classification of *Lena* and *Plane* images, where the primitives are classified as: (i) texture, (ii) sketch, (iii) primary.

Moreover, the *Laplacian* operator, which is efficient in edge detection, is performed to obtain the effective feature in spatial domain. Let the third feature f_3 denote the mean value of the *Laplacian* map generated by a primitive.

Finally, these features can be combined as a feature vector notated by $F = \{f_1, f_2, f_3\}$. For each sparse primitive ψ_i , we extract the feature vector F_i . The k-means algorithm is then applied to divide the feature space $\{F\}$ into several parts, as shown in Fig. 4 & 5. Primary primitives with smooth changes have relatively higher f_1 and lower f_2 and f_3 , while the texture ones perform on the contrary and the sketch primitives are located between them. Note that most of the primitives are classified into the texture type, only a small part belongs to the primary type according to the results.

C. Hierarchical Visual Signal Sparse Representation

Based on the *IGM* theory, visual signal is formed by the primary visual information and uncertain information. The primary visual information is predicted by correlations among stimuli as well as priori knowledge in brain. In [39], the visual perception is partitioned into two portions, the predicted portion and the disorderly portion, corresponding to the primary and uncertain information, respectively. Note that the predicted portion contains mostly the content-dependent structural information, while the disorderly portion corresponds to the non-structural information. Generally, the non-structural information is content-independent. In this subsection, we propose the hierarchical visual signal representation based on the sparse coding, where the images can be decomposed into structural and non-structural layers.

Sparse coding aims at obtaining the optimal representation coefficients α_i in terms of the trained dictionary Ψ , under the accuracy and sparsity constraints. It can be formulated as follows,

$$\alpha_i = \arg \min_{\alpha_i} \sum_k \|x_i - \Psi \alpha_i\|_2^2, \text{s.t. } \|\alpha_i\|_0 < \mathbb{L}, \quad (4)$$

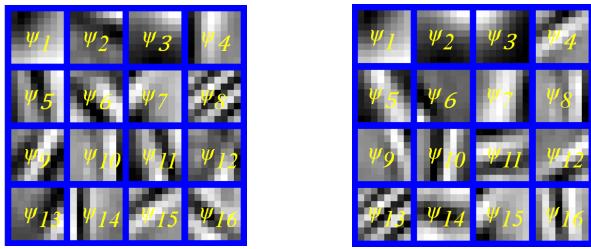


Fig. 6. Sixteen primitives selected by *OMP* algorithm (raster order).

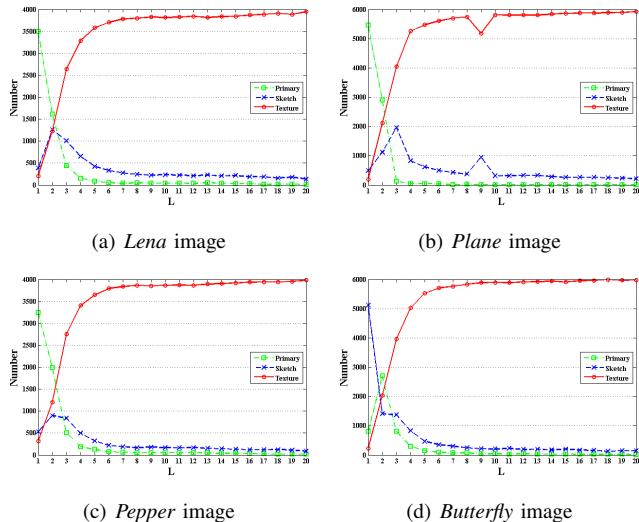


Fig. 7. Plots of the primitive numbers in terms of the iteration L . The green, blue and red lines indicate primary, sketch and texture primitives, respectively.

where α_i is the sparse representation vector. This can be solved by a number of approximation algorithms. In this paper, the typical *orthogonal matching pursuit (OMP)* algorithm [23] is employed to solve this problem.

The *OMP* method works in a greedy fashion that chooses the most similar primitive with the residual at each iteration. Note that the “residual” at the first iteration is the original patch itself. Then the original signal is subtracted by the chosen primitive to update the residual. Note that the notation L will be used to indicate a specific iteration during the *OMP* algorithm.

Intuitively, the primitive which is most similar with the image patch is picked first, followed by some detailed primitives for shaping the visual contents. Let ψ_i denote the i^{th} selected primitive by the *OMP* algorithm. As shown in Fig. 6, sixteen picked primitives ($\psi_1 \sim \psi_{16}$) are arranged in raster order. Note that the first few primitives are visually smoother than others. It can be inferred that the primary primitives are captured before sketch and texture primitives in a statistical perspective. To verify this hypothesis, another simulation is conducted, where the number of each type of primitive in every iteration is recorded during the sparse reconstruction process. The results are depicted in Fig. 7, from which we can conclude that:

- The number of primary primitives dominates in the first iteration (i.e. when $L = 1$), though Fig. 5 has shown that

the number of the primary primitives is smallest. Then the plot of primary primitive drops dramatically with the increasing L .

- The number of sketch primitives is relatively small and decreases slowly to a low level. Note that for the case of *Butterfly* image in Fig. 7d, the number of sketch primitives is maximum at the first iteration. This can be accounted by the large areas of edge and texture contents in the *Butterfly*.
- The number of texture primitives is always smallest when $L = 1$ and achieves maximum with the increasing L .

It can be concluded that *OMP* scheme decomposes image signal into multiple layers, and these layers are naturally ordered by perceptual importance, as demonstrated in Fig. 8.

- The most significant structural information can be reconstructed by the first layer ($L = 1$), which is also referred to as the *primary layer*.
- The second layer ($L \in [2 \sim \tilde{L}]$), or *sketch layer*, recovers the detailed information. Note that \tilde{L} corresponds to the boundary between the second and the third layers, which will be specifically defined in Section III.
- With the first two layers, almost all the perceptual information that can be captured by the HVS, has been well represented. The combination of the two layers is named as the *structural layer*. Different from that, the perceptual information contained in the third layer ($L > \tilde{L}$), i.e. the *non-structural layer*, is negligible due to its low correlation with the visual experience.

Fig. 9 shows the sparse reconstruction process of the *Lena* image, where we can clearly observe that the image reconstructed by primary layer ($L = 1$) recovers the basic information but with remarkable blocking artifacts. With the help of the sketch primitives in the second layer, these artifacts are strongly alleviated and the blurry edges are rapidly sharpen. Most of the structural information has been recovered until $L = 6$. This observation accords well with the visual system mechanism, wherein the primary component (what object is it) is perceived before the details (what does the object look alike). However human can hardly distinguish the differences between the reconstructed images (from $L = 9$ to $L = 13$), as these visual elements are highly non-structured and insensitive to the HVS.

III. ENTROPY OF PRIMITIVE

In Section II the concept of hierarchical visual signal representation is introduced, where we incorporate the typical sparse reconstruction method (*OMP* algorithm) with human visual perception. In fact, the hierarchical structure in visual representation is closely related to the early work in [20]. It is stated that the major properties of sparse primitives (or receptive fields of simple cells) include spatially localized, oriented and bandpass, which has been utilized in Section II. In addition sparse primitives have been shown to be effective in dealing with rich, varied and directional information contained in natural scene. In this section, the novel concept of entropy of primitive (EoP) is proposed to bridge the sparse representation and visual information evaluation. In this way, the

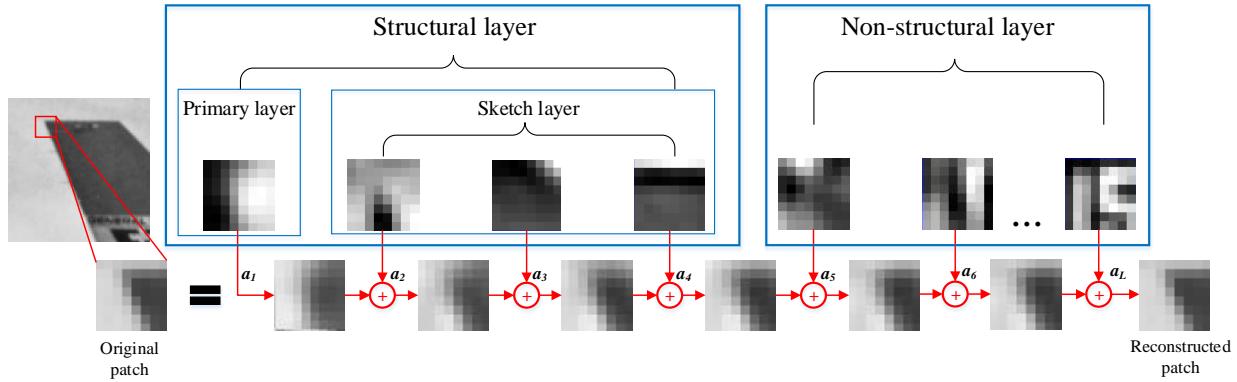


Fig. 8. Illustration of the hierarchical image representation and reconstruction process, that can be decomposed into the structural and non-structural layers.

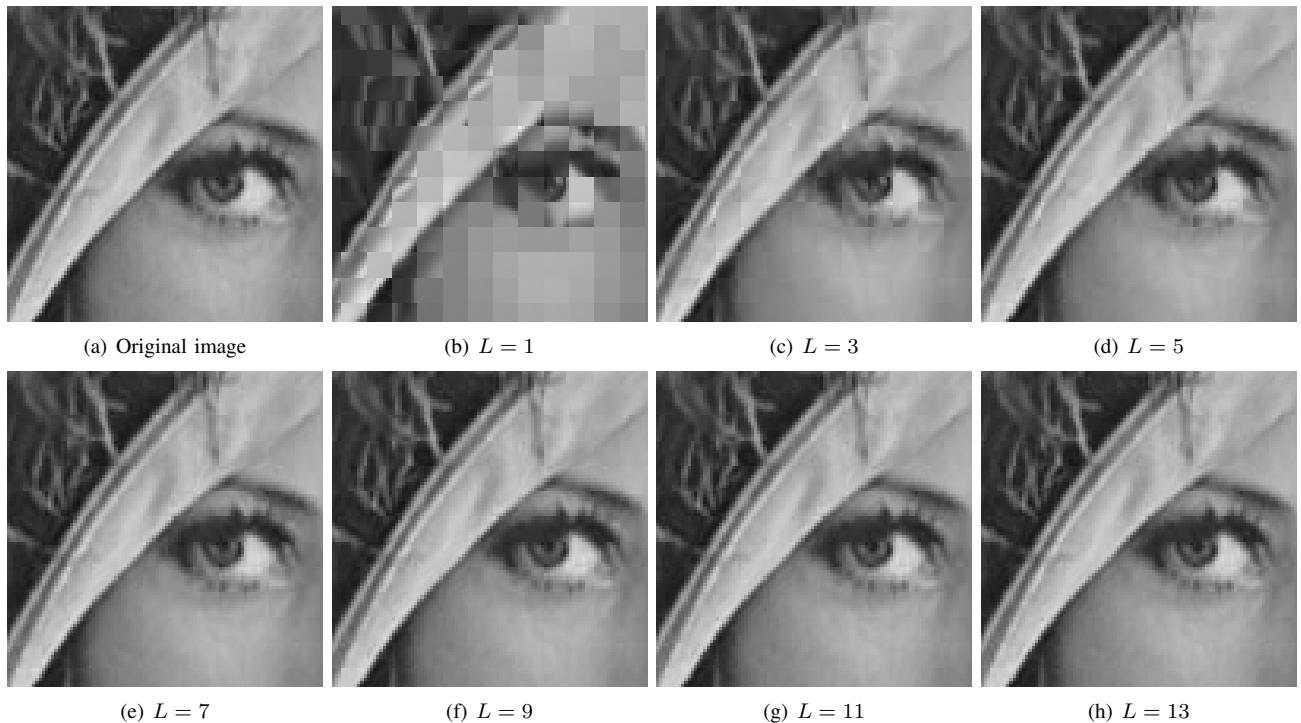


Fig. 9. Reconstructed *Lena* images with different layer numbers (i.e. L).

visual information can be quantified to evaluate the “surprise” level when encountering a visual scene. Consequently, the boundary between the structural and non-structural layers can be determined.

A. Entropy of Primitive

The second law of thermodynamics states that the entropy of an isolated system never decreases and isolated systems always evolve toward thermodynamic equilibrium, a state with maximum entropy. However for the biological agents such as HVS, this maximum entropy state could never be reached as their internal states are limited to a relative low entropy level for keeping themselves within some physiological bounds. This bound is determined by the level of “surprise” in a particular visual scene which is known as the “free energy

principle” [14]. Analogous to this law, we propose a novel concept to measure the state in the image representation system when responding to a “surprise” environment, i.e. the entropy of primitive (EoP). It is inspired by the observation in Fig. 7, wherein the three curves have the similar tendency of converging to a stable state. This state could be related to the upper bound of the visual system.

Before introducing the EoP, useful mathematical notations should be firstly defined. Let n_j^i indicate the number of the j^{th} primitive selected in the i^{th} iteration during the *OMP* algorithm. For instance, 100 patches select the first primitive in the first iteration, such that $n_1^1 = 100$. N_j^i represents the total number of the j^{th} primitive used in the previous i iterations, which can be calculated as follows,

$$N_j^i = \sum_{t=1}^i n_j^t. \quad (5)$$

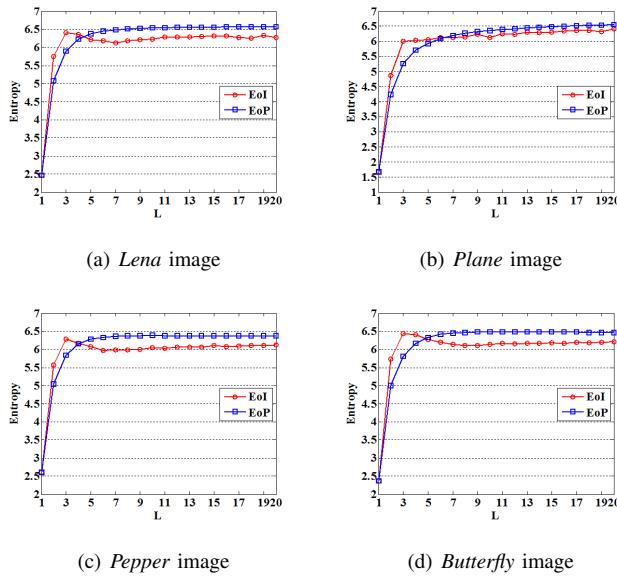


Fig. 10. The EoI and EoP curves in terms of the iteration L .

Then two probability density functions (PDF) can be given by

$$p^i(j) \triangleq \frac{n_j^i}{\sum_t n_t^i}, \quad (6)$$

$$P^i(j) \triangleq \frac{N_j^i}{\sum_t N_t^i}, \quad (7)$$

which represent the distributions of primitives in the i^{th} iteration and in the previous i iterations, respectively. They can also be interpreted as the increment distribution and cumulative distribution of the primitives during the *OMP* algorithm. According to the relationship between the receptive fields of simple cells and primitives, we evaluate how surprise the receptive field reflects on the encountered input scene by estimating the information of primitives. Specifically, based on the *Shannon* theory, two entropy values have been defined as follows,

$$EoI_i \triangleq -\sum_{j=1}^k p^i(j) \log p^i(j), \quad (8)$$

$$EoP_i \triangleq -\sum_{j=1}^k P^i(j) \log P^i(j), \quad (9)$$

where k is the number of the primitives. The *Entropy of Increment* (EoI_i) and the *Entropy of Primitive* (EoP_i) indicate the entropy of the i^{th} distribution and the entropy of the cumulative distribution of previous i iterations, respectively. Note that $EoP_1 = EoI_1$ because they correspond to the same distribution in the first iteration.

The EoI and EoP curves in terms of the iteration L have been shown in Fig. 10, where the horizontal axis is the iteration L . With the increasing L , the EoP_L monotonously rises and converges to a constant level. It is also interesting to observe that the value of EoI_L always keeps in a relatively high level except when $L = 1$.

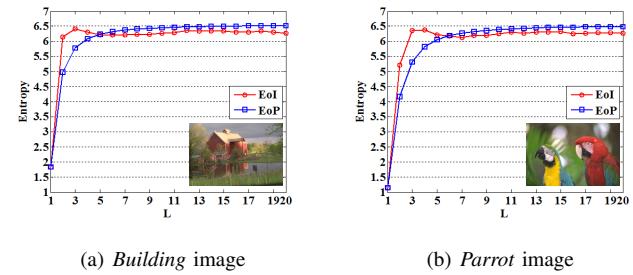


Fig. 11. The *EoI* and *EoP* curves using the specific dictionary trained from *Lena* image. The thumbnail of test image is given in the right corner.

B. Robustness Verification of EoP

To verify the robustness of EoP, we conduct several experiments considering the impacts from both the dictionary universality and the patch size.

1) *Universal Dictionary vs. Adaptive Dictionary*: First, we train a dictionary by *Lena* image and apply it to other natural images. The corresponding *EoI* and *EoP* curves are depicted in Fig. 11, where the similar statistics have been achieved compared to that in Fig. 10, though the visual content of test images is completely distinct to the training image.

Subsequently, a universal dictionary is trained over the 25 original images in TID2008 dataset [41], and is applied to all the test images in Fig. 1. The K-SVD algorithm [21] is a typical method for training the content-adaptive sparse dictionary, which is performed in an iterative batch way by accessing the whole training set at each iteration. To achieve this we have to efficiently handle very large training data, an advanced and efficient online training algorithm [42], based on stochastic approximation, is employed for obtaining the universal dictionary in this work. The trained dictionaries using adaptive and universal approaches are demonstrated in Fig. 12, where it can be found that the primitives in universal dictionary contain less structural contents comparing with those in adaptive dictionary especially for texture and sketch primitives.

The EoP curve comparisons between universal and adaptive dictionary are demonstrated in Fig. 13. Note that both of the two dictionaries contain 256 primitives with the size of 8×8 . It reveals that the universal dictionary leads to larger convergence EoP value, which approaches the extreme entropy of uniform distribution over 256 bins. This can be accounted by the instability of the universal dictionary. However the converging tendencies are similar regardless of which dictionary is used.

2) *The Influence of Patch Size*: To analyze the impact of various patch sizes on *EoP*, we change it from 8×8 , 12×12 to 16×16 and the corresponding dictionary is adaptively trained for each input image. For a fair comparison, the dictionaries with different size should have identical completeness γ . The completeness here is defined as $\gamma \triangleq \frac{k}{n}$, considering the dictionary $\Psi \in \mathbb{R}^{n \times k}$. In this work we set $\gamma = 4$, i.e. the number of primitives is 256 for 8×8 patches, and so on. The experimental results are reported in Fig. 14, where the similar *EoP* tendency has been found for different patch sizes. However the convergence values are distinct, because they are

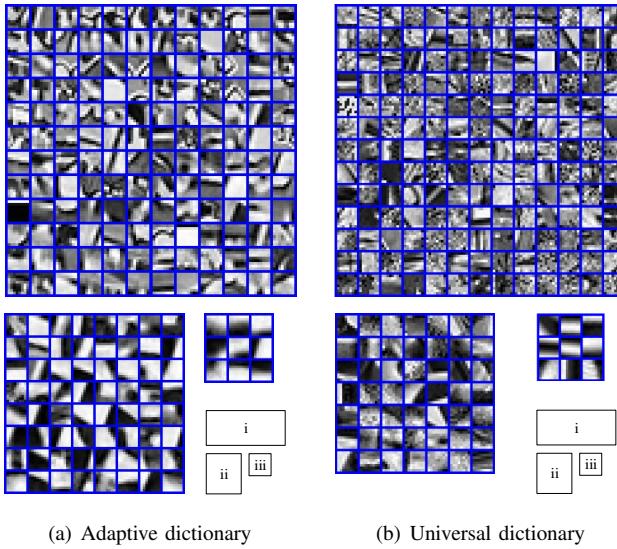


Fig. 12. Visualization of adaptive and universal dictionaries. (a) Adaptive dictionary trained using *Plane* image; (b) Universal dictionary trained using all images in TID2008 dataset [41]. Primitives are classified into different categories including: (i) texture primitives, (ii) sketch primitives and (iii) primary primitives.

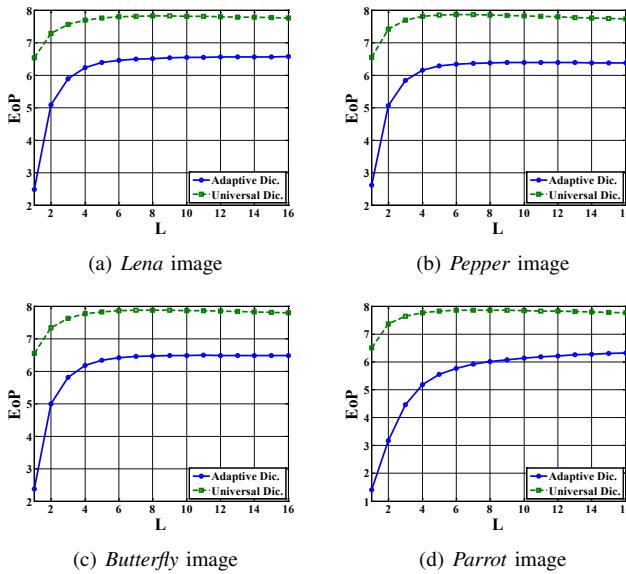


Fig. 13. *EoP* curve comparison between adaptive and universal dictionaries.

with different number of primitives resulting in different limit values.

C. A Closer Look at the *EoP*

To further explore the *EoP*, extensive mathematical analysis as well as experimental verifications have been presented in this subsection.

It has been found in Fig. 10 that the *EoI* tends to be constant when $L \geq 3$, and interestingly the *EoP* has the similar trend approaching the *EoI* curve after this threshold. It should be noted that the EoP_i is not simply calculated by adding EoP_{i-1} and EoI_i , but is determined by the summation of the

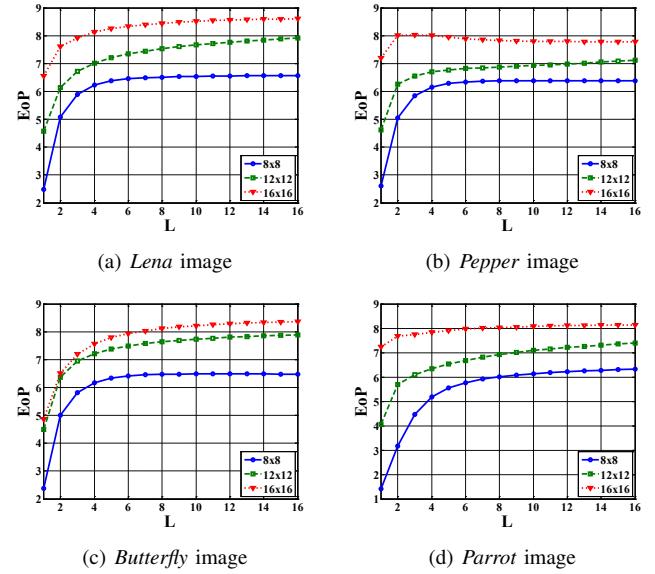


Fig. 14. *EoP* curve comparison among different Patch sizes, including 8×8 , 12×12 and 16×16 .

corresponding two distributions, i.e.

$$N_j^i = N_j^{i-1} + n_j^i. \quad (10)$$

Accordingly we assume that the EoI_L corresponds to an identical distribution when L is larger than the threshold L^* , and we denote this identical distribution as B_j . With this definition, we have

$$n_j^i \triangleq B_j \quad (i > L^*), \quad (11)$$

and the corresponding PDF of B_j can be written as

$$p_B(j) \triangleq \frac{B_j}{\sum_j B_j}. \quad (12)$$

Then let the A_j represent the summation of the previous L^* distributions, i.e.

$$A_j \triangleq N_j^{L^*}. \quad (13)$$

Such that the PDF of the distribution A_j is given by

$$P_A(j) \triangleq \frac{A_j}{\sum_j A_j}. \quad (14)$$

Obviously we get that $EoP_{L^*} = -\sum_{j=1}^k P_A(j) \log P_A(j)$. At this point, we are focusing on the value of EoP_{L^*+n} , where $n \in \mathbb{Z}$. Substituting Eqn. (10) (11) (13) into the Eqn. (7), we have,

$$P^{L^*+n}(j) = \frac{A_j + nB_j}{\sum_j A_j + n \sum_j B_j}. \quad (15)$$

Consequently,

$$EoP_{L^*+n} = -\sum_{j=1}^k \frac{A_j + nB_j}{S_A + nS_B} \log \frac{A_j + nB_j}{S_A + nS_B}. \quad (16)$$

Concise, we simplify the definitions as $S_A \triangleq \sum_j A_j$ and $S_B \triangleq \sum_j B_j$. Considering the limit of Equ. (16), we have

$$\lim_{n \rightarrow \infty} EoP_{L^*+n} = -\sum_{j=1}^k \frac{B_j}{S_B} \log \frac{B_j}{S_B}. \quad (17)$$

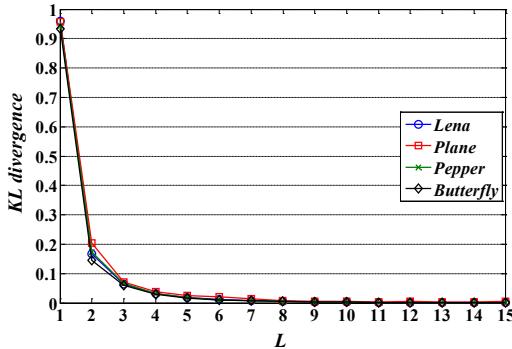


Fig. 15. *KL divergence* of adjacent increment distributions in terms of L .

Note that the right side of the limitation is exactly the entropy of the assumed identical distribution B_j in Equ. (11), indicating that the limit value of EoP is dominated by the *image-independent identical* distribution. It explains why the EoP curves of different images increase to approach the EoI curves and converge to an approximately equal value. This limitation value in experiments (about 6.5) is much higher than the initial entropy EoP_1 , but is close to the maximum entropy of the uniform distribution (8, because the number of primitives is 256). This corresponds well to the upper bound of entropy state in visual system as stated in the free energy principle.

To further demonstrate the reasonability of the hypotheses of the *image-independent identical* distribution, another verification experiment is conducted, in which the *Kullback-Leibler divergence (KL divergence)* is employed to measure the difference between adjacent increment distributions $p^i(j)$ and $p^{i+1}(j)$. The *KL divergence* defined in Equ. (18) satisfies $KL\{p^i(j)||p^{i+1}(j)\} \geq 0$ with equality if, and only if $p^i(j) = p^{i+1}(j)$.

$$KL\{p^i(j)||p^{i+1}(j)\} \triangleq - \sum_{t=1}^k p^i(t) \log \left\{ \frac{p^{i+1}(t)}{p^i(t)} \right\}. \quad (18)$$

The results have been reported in Fig. 15, where the increment distribution beyond a threshold (i.e. n_j^i or $p^i(j)$ when $i > L^*$) has no significant changes since the *KL divergence* of these adjacent distributions achieves almost 0. The results accord well with our hypothesis that the *image-independent identical* distribution exists in the visual perception system, and finally creates the convergence of the EoP curve as well as the saturation of visual information.

To gain more insights into this problem, the geometrical interpretation of the *OMP* method is further demonstrated in Fig. 16. We treat the primitives as n -dimensional vectors in a n -dimensional space, denoted by ψ_i . Given a n -dimensional signal X , *OMP* algorithm finds a linear combination of the primitives to approximate X in a greedy fashion. The first chosen primitive ψ_1 is the most similar one with the given X . The non-local similarity contained in natural scene will decrease the instability of the sparse system resulting in relatively small entropy. However, in the following iterations the “randomly” distributed residual signals will lead to in-

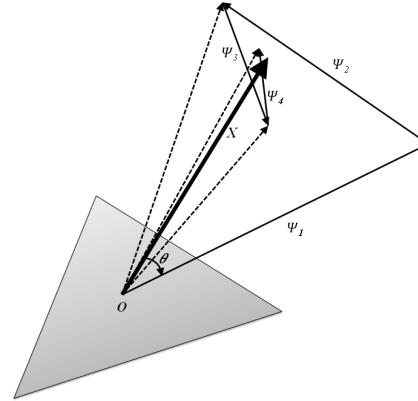


Fig. 16. Geometrical illustration of the *OMP* method. The bold solid line indicates the original signal X , the other solid lines denoted by $\psi_1 \dots \psi_4$ are primitives selected by the *OMP*. The dashed lines represent the reconstructed signals in each iteration, which are gradually approaching to X .

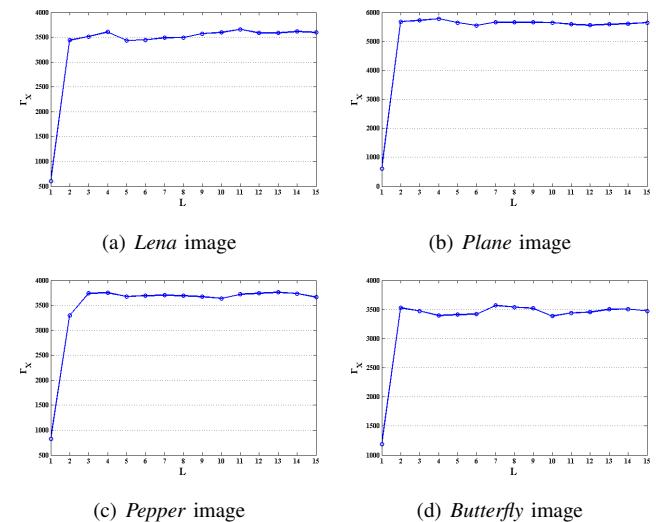


Fig. 17. Angular disparity plots in terms of L .

creasing and convergent entropy value. It gives another reliable interpretation of the high convergent value of the EoP and EoI.

To quantify the randomness level of an image, the angular disparity is defined as follows,

$$\Gamma_X \triangleq \sum_{i=1}^N \sum_{j=1}^N \frac{\langle x_i, x_j \rangle}{|x_i| \cdot |x_j|}, \quad (19)$$

where X represents the original or residual image, and x_i is a patch in it. $\langle x_i, x_j \rangle$ denotes the inner product of the two vectors. The larger value of Γ_X corresponds to higher disorder level of the image system. The Γ_X of the residual images during the sparse reconstruction process has been plotted in Fig. 17, indicating that residual signals ($L > 1$) are much more randomly distributed than the original signal ($L = 1$). It can be inferred that the significant increment of disorder level causes the dramatically increasing and finally converging EoP curve.

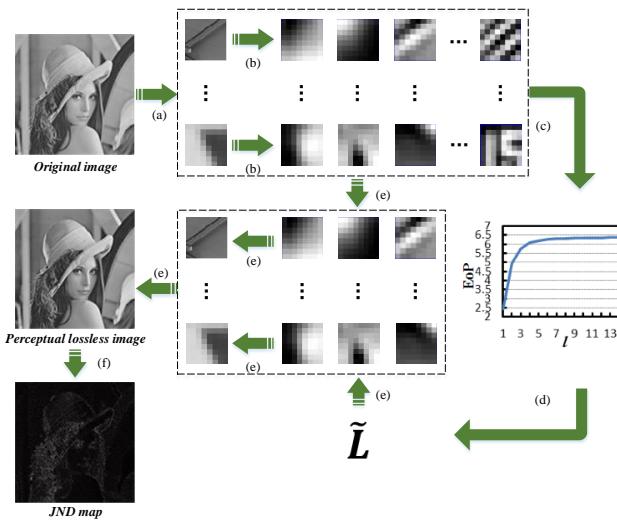


Fig. 18. The framework of the EoP based JND scheme. (a) Image partitioning and dictionary learning, (b) Sparse decomposition using *OMP* method, (c) EoP curve, (d) Computing the threshold \tilde{L} , (e) Image reconstruction with \tilde{L} primitives, (f) Generating the JND map by subtracting the reconstructed image from the original image.

D. Entropy of Primitive and Visual Information

It has been stated in Section II that sparse representation corresponds to a process of hierarchical signal decomposition, in which the high level sparse primitives of non-structural layer give less information to the HVS. Also it is widely acknowledged that the entropy is an effective measurement of information according to *Shannon's* theory. Consequently we propose to apply EoP in evaluating the visual information. It can be supported by the fact that the quality of reconstructed image and the corresponding EoP value have the similar converging tendency, implying the saturated visual information.

Consequently, the threshold \tilde{L} between the structural and non-structural layers, as mentioned in Section II, can be defined as follows,

$$\tilde{L} = \arg \min_i i, s.t. \frac{EoP_i - EoP_{i-1}}{\max_j (EoP_j) - \min_j (EoP_j)} < \varepsilon, \quad (20)$$

where ε is the control parameter that determines the convergence condition and is set as 0.01 in the implementation. This threshold also indicates the saturation of perceptual information, beyond which the non-structured visual elements are negligible that can be hardly perceived by the HVS.

IV. APPLICATION: JUST NOTICEABLE DISTORTION

A. EoP based JND profile

In this section, the EoP is employed in generating the JND map, motivated by the convergence characteristic of the visual information from the EoP perspective. The framework of the proposed method is illustrated in Fig. 18. Given an original image X , firstly it is partitioned into non-overlapped patches and the K-SVD algorithm is applied for training the primitive dictionary. Then each patch is decomposed into a linear combination of a few primitives by the *OMP* approach. The primitive

distribution in each iteration can generate corresponding EoP value by its definition. The threshold \tilde{L} can be calculated using Eqn. (20). Subsequently, the image can be reconstructed by the first \tilde{L} primitives, while the other primitives are discarded. The reconstructed image denoted as \tilde{X} has been shown to be visually equal with the original image, because the negligible visual information is highly non-structured. Therefore, the image \tilde{X} can be treated as perceptual lossless. At last the JND map is generated by subtracting the perceptual lossless image from the original image, which can be defined as follows,

$$JND \triangleq |\tilde{X} - X|, \quad (21)$$

where the notation $|\bullet|$ indicates the absolute operator.

B. Comparison with the state-of-the-art JNDs

The performance of the JND model can be evaluated by its effectiveness in concealing distortions in images. The JND noised image is generated by injecting the JND map into the original image as follows,

$$\tilde{X}(i, j) = X(i, j) + \eta \cdot S^{random}(i, j) \cdot M(i, j), \quad (22)$$

where i and j are coordinates in pixel domain, $X(i, j)$ and $M(i, j)$ represent a pixel in the original image and the corresponding JND map, respectively. \tilde{X} is the noised image by JND. $S^{random}(i, j)$ is randomly set as -1 or +1. The parameter η should be adjusted to guarantee that the noised images by different JND models have the identical error energy, i.e. the same PSNR or MSE. With the same distortion level, the better perceptual quality the noise-injected image has, the more accurate the JND model is. Namely, with a same level of perceptual visual quality, a more accurate JND model can shape more noises in an image [4].

Perceptual visual quality of the JND-guided noised images can be evaluated using subjective viewing tests. In this work, the two-alternative forced choice (2AFC) evaluation method [43] is conducted, which is widely employed in image and video oriented applications. In the 2AFC experiment, a pair of images is presented to the subject at the same time, one of the pair is the original image while the other is the noised version. Then the subject is forced to choose one with pleasing quality from the pair. After all pair-wise tests, the percentage ω of correctly choosing the original image over all test pairs is recorded. If the ω is close to 0.5, the subject can hardly distinguish the original and noised image, indicating a more accurate JND model.

In this work, 14 subjects with half experts and half non-experts were invited to participate in the subjective experiment. Each observer was preliminarily instructed before the actual experiments. All the 50 images in Fig. 1 were tested by subjects. The image pairs were presented in a LCD monitor with 1920×1080 resolution while keeping constant viewing conditions. Each pair allows up to 5 seconds to response relying on the intuition for the testers. The EoP based JND was compared to the state-of-the-art JND models including [44]–[46]. Notice that all the noised images are subjected to the identical distortion level (PSNR values equal to 35.0 dB in the experiments). Fig. 19 shows the subjective results indicating

the EoP based JND outperforms other JND models as the averaged percentage ω of EoP based JND is much closer to 0.5. It also shows that the HVS can hardly notice the noises injected by EoP based JND. Furthermore, the noised images by different JNDs are demonstrated in Fig. 20. It is obvious that the images noised by EoP based JND have better perceptual quality in comparison with others, though they have the same PSNR. Consequently, the proposed JND profile outperforms the state-of-the-art JND models, demonstrating that the EoP is an effective approach in evaluating visual information.

V. CONCLUSION AND FUTURE WORK

In this paper, we bridge the sparse representation and visual information evaluation with the concept of EoP. Firstly, we interpret the visual perception as a hierarchical representation process in the sparse representation perspective. The sparse primitives are classified into different groups according to their inherent properties, in the meanwhile the visual perception can be partitioned into structural (primary and sketch) and non-structural (texture) layers. It has also been found that the structural layer can shape most of the visual information, while the non-structural layer is negligible for HVS. Accordingly the concept of the entropy of primitive (EoP) is introduced to estimate the visual information and also to determine the boundary between structural and non-structural layers. It is further demonstrated that the EoP is highly relevant with visual information, supported by sufficient mathematical analysis and experimental verifications. We also verify the robustness of the EoP according to different patch sizes and dictionary universality. Moreover the EoP is applied in generating the JND map, which has been verified to outperform the state-of-the-art JND schemes in the subjective experiments.

The effective of EoP may be understood as a successful combination of the psychological process in human visual system and sparse representation. Based on the good correlation with visual information, the EoP based applications such as image quality assessment and image/video coding will be further studied. In addition, we will further explore the relationship between sparse representation and visual information representation by considering: 1) adaptively selecting the patch size to better characterize the visual contents, 2) applying different weights to primitives in terms of their visual importance, 3) the impacts of viewing conditions on visual perception, such as viewing distance and image resolution.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviews for their valuable comments that significantly helped us in improving the presentation of the paper.

REFERENCES

- [1] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, "Just noticeable difference for images with decomposition model for separating edge and textured regions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1648–1652, 2010.
- [2] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 742–752, 2005.
- [3] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 820–829, 2006.
- [4] X. Yang, W. Lin, Z. Lu, E. P. Ong, and S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Processing: Image Communication*, vol. 20, no. 7, pp. 662–680, 2005.
- [5] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.
- [6] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in dct domain," *IEEE Trans. Circuits Syst. Video Technol.*
- [7] H. R. Wu and K. R. Rao, *Digital video image quality and perceptual coding*. CRC press, 2005.
- [8] A. B. Watson, "DCTune: A technique for visual optimization of dct quantization matrices for individual images," in *Sid International Symposium Digest of Technical Papers*, vol. 24. SOCIETY FOR INFORMATION DISPLAY, 1993, pp. 946–946.
- [9] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? a new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, 2009.
- [10] B. Girod, "What's wrong with mean-squared error?" in *Digital images and human vision*, 1993, pp. 207–220.
- [11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [12] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.
- [13] L. Zhang, D. Zhang, and X. Mou, "Fsim: a feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.
- [14] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, 2012.
- [15] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2. Ieee, 2003, pp. 1398–1402.
- [16] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.
- [17] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2387–2390, 2015.
- [18] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.
- [19] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer, 2009, vol. 39.
- [20] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
- [21] M. Aharon, M. Elad, and A. Bruckstein, "KSVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [22] J. Z. Salvatierra, *New sparse representation methods; application to image compression and indexing*. Human-Computer Interaction, 2010.
- [23] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [24] C. Hua-Wen, Y. Hua, G. Yong, and W. Ming-Hui, "Sparse feature fidelity for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4007–4018, 2013.
- [25] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1146–1153.
- [26] X. Zhang, S. Wang, S. Ma, S. Liu, and W. Gao, "Entropy of primitive: A top-down methodology for evaluating the perceptual visual information," in *Visual Communications and Image Processing (VCIP)*, 2013, pp. 1–6.
- [27] S. Wang, X. Zhang, S. Ma, and W. Gao, "Reduced reference image quality assessment using entropy of primitives," in *Picture Coding Symposium (PCS)*, 2013, pp. 193–196.
- [28] J. Zhang, S. Ma, R. Xiong, D. Zhao, and W. Gao, "Image primitive coding and visual quality assessment," in *Advances in Multimedia Information Processing–PCM*, 2012, pp. 674–685.

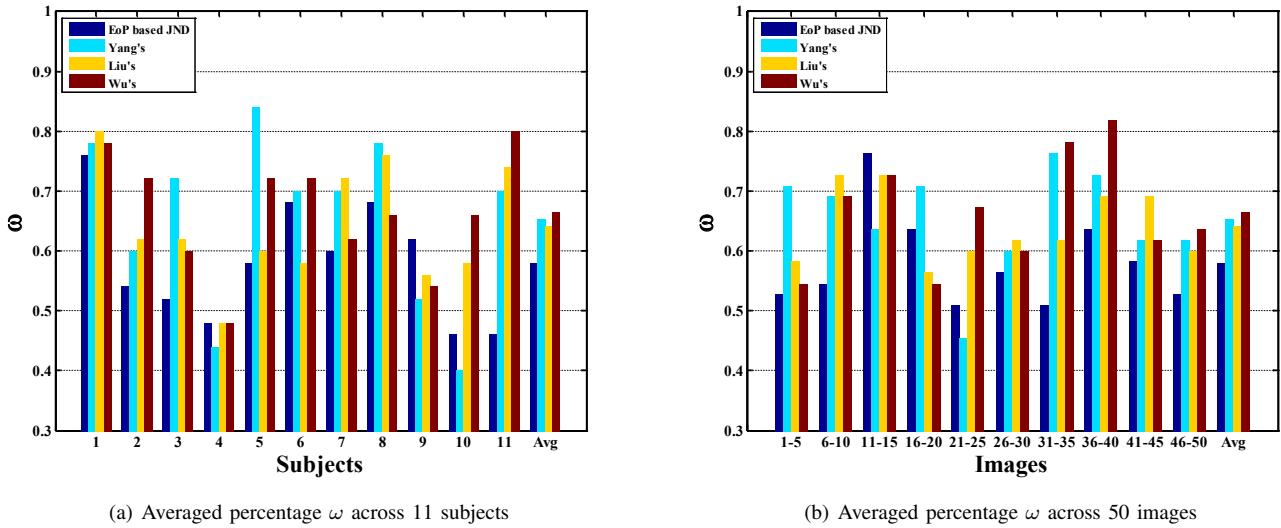


Fig. 19. Subjective results of comparison between EoP based JND and state-of-the-art JNDs including Yang's [44], Liu's [45] and Wu's [46]. The vertical axis indicates the ω , i.e. the probability of correct selection. (a) is averaged results across 11 subjects after excluding 3 outliers, while (b) is averaged across 50 images. Notice that the results of the 50 images are further divided into 10 groups and averaged in each group for viewing convenience. The last column in (a) & (b) is the ultimate averaged values across all the subjects and images.



Fig. 20. Images noised by different JND models including Yang's [44], Liu's [45], Wu's [46] and EoP based JNDs from left to right, respectively. Note that the PSNR of each noised images is exactly the same.

- [29] X. Zhang, S. Wang, S. Ma, and W. Gao, "Towards accurate visual information estimation with entropy of primitive," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1046–1049.
- [30] X. Zhang, S. Wang, K. Gu, T. Jiang, S. Ma, and W. Gao, "Sparse structural similarity for objective image quality assessment," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2015.
- [31] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [32] J. Zhang, C. Zhao, R. Xiong, S. Ma, and D. Zhao, "Image super-resolution via dual-dictionary learning and sparse representation," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2012, pp. 1688–1691.
- [33] J. Zhang, D. Zhao, and W. Gao, "Group-based sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3336–3351, 2014.
- [34] J. Zhang, C. Zhao, D. Zhao, and W. Gao, "Image compressive sensing recovery using adaptively learned sparsifying basis via l0 minimization," *Signal Processing*, vol. 103, pp. 114–126, 2014.
- [35] J. Zhang, D. Zhao, R. Xiong, S. Ma, and W. Gao, "Image restoration using joint statistical modeling in a space-transform domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 915–928, 2014.
- [36] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, 2010.
- [37] H. Xiong, Z. Pan, X. Ye, and C. W. Chen, "Sparse spatio-temporal representation with adaptive regularized dictionary learning for low bit-rate video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 710–728, April 2013.
- [38] J.-W. Kang, M. Gabbouj, and C.-C. Kuo, "Sparse/DCT (S/DCT) Two-layered representation of prediction residuals for video coding," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2711–2722, July 2013.
- [39] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Trans. Image Process.*, vol. 22,

- no. 1, pp. 43–54, 2013.
- [40] X. Zhang, W. Lin, and P. Xue, “Improved estimation for just-noticeable visual distortion,” *Signal Processing: Image Communication*, vol. 85, no. 4, pp. 795–808, 2005.
- [41] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008-a database for evaluation of full-reference visual quality assessment metrics,” *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.
- [42] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online Learning for Matrix Factorization and Sparse Coding,” *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.
- [43] C. K. Hiscock, J. D. Branham, and M. Hiscock, “Detection of feigned cognitive impairment: The two-alternative forced-choice method compared with selected conventional tests,” *Journal of Psychopathology and Behavioral Assessment*, vol. 16, no. 2, pp. 95–110, 1994.
- [44] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, “Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 742–752, June 2005.
- [45] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, “Just noticeable difference for images with decomposition model for separating edge and textured regions,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1648–1652, Nov 2010.
- [46] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, “Just noticeable difference estimation for images with free-energy principle,” *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1705–1710, Nov 2013.



Jian Zhang (M’14) received B.S. degree from Department of Mathematics, Harbin Institute of Technology (HIT), Harbin, China, in 2007, and received M.Eng. and Ph. D degrees from School of Computer Science and Technology, HIT, in 2009 and 2014, respectively. Currently, he is working as a post-doctoral fellow at National Engineering Laboratory for Video Technology (NELVT), Peking University (PKU), Beijing, China. His research interests include image/video compression and restoration, compressive sensing, sparse representation, and dictionary learning. He was the recipient of the Best Paper Award at the 2011 IEEE Visual Communication and Image Processing.



Huifang Sun (M’85-SM’93-F’00) graduated from Harbin Military Engineering Institute, Harbin, China, and received the Ph.D. from University of Ottawa, Canada. He was an Associate Professor in Fairleigh Dickinson University in 1990. He joined to Sarnoff Corporation in 1990 as a member of technical staff and was promoted to a Technology Leader of Digital Video Communication. In 1995, he joined Mitsubishi Electric Research Laboratories (MERL) and was promoted as Vice President and Deputy Director in 2003 and currently is a Fellow of MERL. He has co-authored two books and published more than 150 Journal and Conference papers. He holds more than 61 US patents. He obtained the Technical Achievement Award for optimization and specification of the Grand Alliance HDTV video compression algorithm in 1994 at Sarnoff Lab. He received the best paper award of 1992 IEEE Transactions on Consumer Electronics, the best paper award of 1996 ICCE and the best paper award of 2003 IEEE Transactions on CSVT. He was an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology and was the Chair of Visual Processing Technical Committee of IEEE Circuits and System Society. He is an IEEE Fellow.



Siwei Ma (S’03-M’12) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. From 2005 to 2007, he held a post-doctorate position with the University of Southern California, Los Angeles. Then, he joined the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing, where he is currently a Professor. He has published over 100 technical articles in refereed journals and proceedings in the areas of image and video coding, video processing, video streaming, and transmission.



Xiang Zhang received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He is currently working toward the Ph.D. degree. His research interests include image/video quality assessment, video compression and visual retrieval.



Wen Gao (M’92-SM’05-F’09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Japan, in 1991. He is a professor of computer science at Peking University, China. Before joining Peking University, he was a professor of computer science at Harbin Institute of Technology from 1991 to 1995, and a professor at the Institute of Computing Technology of Chinese Academy of Sciences. He has published extensively including five books and over 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. Dr. Gao served or serves on the editorial board for several journals, such as IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Transactions on Image Processing, IEEE Transactions on Autonomous Mental Development, EURASIP Journal of Image Communications, Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.



Shiqi Wang (M’15) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008, and the Ph.D. degree in computer application technology from the Peking University, in 2014. He is currently a Postdoc Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. From Apr. 2011 to Aug. 2011, he was with Microsoft Research Asia, Beijing, as an Intern. His current research interests include video compression and image/video quality assessment.