Towards Accurate Visual Information Estimation with Entropy of Primitive

Xiang Zhang, Shiqi Wang, Siwei Ma, Ruiqin Xiong and Wen Gao
Institute of Digital Media, Peking University, Beijing 100871, China
Email: {x_zhang, sqwang, swma, rqxiong, wgao}@pku.edu.cn

***Abstract*—Recently, a novel concept referred to as *Entropy of Primitive* (*EoP*) has been proposed for evaluating the visual information of natural images. The idea originates from the sparse representation, which has been successfully applied in a wide variety of signal processing and analysis tasks. This is because of the high efficiency of sparse representation in dealing with rich, varied and directional information contained in the natural scene. In this paper, we further explore the *EoP* to bridge the sparse representation and visual perception. Sparse primitives are divided into three categories depending on their visual importance. Accordingly the visual signal is decomposed into structural and non-structural layers. It is found that the image sparse representation is highly relevant with the hierarchical visual information construction process in representing the natural scene. We evaluate the efficiency and robustness of the *EoP* in real applications, including surveillance video and shot boundary detection.**

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 revelent with the visual information.

Near-threshold method measures the distortions that the HVS couldn't perceive. Generally, this is referred to as Just-Noticeable Distortion (JND). The JND models have been studied for decades and successfully applied in many fields such as image/video coding [1] and quality assessment. These models take advantages of the characteristics of the HVS, including contrast sensitivity function, luminance adaptation, and texture masking. Therefore, these kinds of methodologies are "bottom-up" method that mimic the functionalities of HVS.

In supra-threshold models, the traditional Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are popular for simplicity, but they cannot correlate well with the subjective quality. Recently, it is found that the natural image is highly structured and HVS is very adapted to the structural information, and therefore the proposed Structure SIMilarity (SSIM) index [2] has drawn extensive attentions and has been applied in video coding techniques [3].

Sparse representation is an emerging and powerful method to describe signals based on the sparsity and redundancy of their representations and is efficient in dealing with rich, varied and directional information contained in natural scene. Based on the sparse theory, a novel concept of *Entropy of Primitive* (*EoP*) has been proposed to estimate the visual information, and has been successfully applied in the areas of image quality assessment [4] and JND model [5]. In this work, we gain some insights into the *EoP* by exploring the distribution of the sparse representation. In the literature [6], it is stated that the primitive has the properties of spatially localized, oriented and bandpass, which closely corresponds to the characteristics of receptive fields of simple cells. Thus we divide the sparse primitives into different categories depending on their visual importance and accordingly the visual construction process could be represented by hierarchical structure. Moreover, we discover the correlations between the EoP and the hierarchical visual representation, and apply the EoP in accurate visual information estimation of natural scene.

The rest of the paper is organized as follows. In Section II, the *EoP* is briefly reviewed and we give a deep analyses of the primitive distribution. In Section III, the visual perception is interpreted as a hierarchical signal decomposition based on the proposed primitive classification, and we bridge the gap between the hierarchical structure and the primitive distribution as well as *EoP*. In Section IV the *EoP* has shown its efficiency and robustness in estimating visual information for real applications, such as surveillance video and shot boundary detection. In Section V, we conclude this paper.

II. Entropy of Primitive

In this section, we briefly introduce the novel concept - the *Entropy of Primitive* (*EoP*) [5]. The image primitive coding assumes that each natural image signal $x(x \in \mathbb{R}^n)$ can be approximated by a linear combination of 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 [7] is employed in this work. The input for training is the non-overlapped 8×8 image patches. After training process, the typical *orthogonal matching pursuit (OMP)* algorithm [8] is applied to solve the sparse representation problem. The *OMP* method works in a greedy fashion that choosing the primitives most similar with the residual at each iteration. Note that the "residual" at first iteration is the original patch itself.

Then the original signal is subtracted by the chosen primitive to update the residual.

Let the n_j^i indicates the number of the j^{th} primitive selected in the i^{th} iteration in the *OMP* method. For instance 100 image patches vote for the first primitive in the first iteration, thus we have $n_1^1 = 100$. And N_j^i represents the total number of the j^{th} primitive selected in previous i iterations, which can be calculated as $N_j^i = \sum_{t=1}^{i} n_j^t$. So that the corresponding probability density function (PDF) is given by $P^i(j) = \frac{N_j^i}{\sum\limits_t N_t^i}$, which represents the cumulative distribution of primitives in previous i iterations. Based on the *Shannon* Theory, the *EoP* is defined as follows,

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

where k is the number of the primitives. Interestingly, in the visual construction process, the *EoP* value monotonously increases with the iterations to approach a constant, while the reconstructed image reaches the state without noticeable visual distortion.

III. From Hierarchical Signal Representation to Visual Information Estimation

In this section, we further explore the properties of *EoP* and build the correlations between sparse representation and visual perception.

A. Entropy of Increment

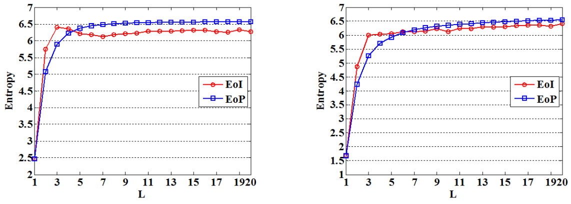
The *EoP* corresponds to the cumulative distribution of primitives in previous i iterations. Similarly we define the PDF in the i^{th} iteration as $p^i(j) = \frac{n_j^i}{\sum\limits_t n_t^i}$, which could be regarded as the increment distribution. Such that the *Entropy of Increment* (*EoI*) is given by,

$$EoI_i = -\sum_{j=1}^{k} p^i(j) \log p^i(j). \tag{2}$$

Note that $EoP_1 = EoI_1$ because they correspond to the same distribution in the first iteration.

The curves of EoI and EoP of two test images are shown in Fig. 1. Here the horizontal axis denotes the l^{th} iteration, with which the EoP_l increases monotonously and converges to a constant (around 6.5). It is also interesting to observe that the value of EoI_l approaches a relatively stable level when $l > 3$. Thus we make the hypothesis that the increment distribution $p^i(j)$ corresponds to an *image-independent identical distribution (IIID)* and leads the *EoP* to a constant value. The *IIID* hypothesis has twofold interpretation. First it is *identical* meaning that the distributions $p^i(j)$ are identical when j is beyond a threshold, such that the EoI and EoP curves could converge. Besides, the *image-independent* hypothesis accounts for the fact that *EoP* curves of different images converge to a similar value.

To verify the hypothesis, we conduct several experiments where the *Kullback-Leibler divergence* (*KL divergence*) is employed to measure the similarity of adjacent increment distributions $p^i(j)$ and $p^{i+1}(j)$. The *KL divergence* satisfies $KL\left\{p^i(j) \| p^{i+1}(j)\right\} \geq 0$ with equality if, and only if $p^i(j) = p^{i+1}(j)$. The results shown in Fig. 2 indicate that the increment distributions are almost *identical* at iterations when $l > 5$, as the *KL divergence* approaches 0. The results accord well with our hypothesis that the *IIID* actually exists in the sparse representation system and finally leads the *EoP* curve to converge.



(a) Lena (b) Plane

Fig. 1. *EoI* and *EoP* curves in terms of the iteration l.

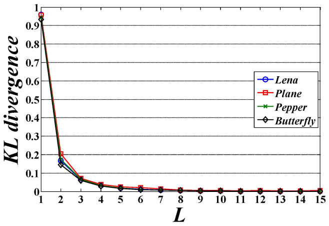


Fig. 2. *KL divergence* of adjacent increment distributions in terms of l.

B. Classification of Sparse Primitive

To deeply analyze the *EoP* and *EoI*, we gain some insights into the characteristics of primitives. A simple solution for primitives classification is provided. We apply k-means cluster algorithm to classify the primitives using extracted features. Specifically, we extract two features from the DCT domain as well as one feature from spatial domain, and employ these features to adaptively cluster all primitives into three categories, named as *primary*, *sketch* and *texture*, respectively [9].

The two DCT domain features are defined as follows,

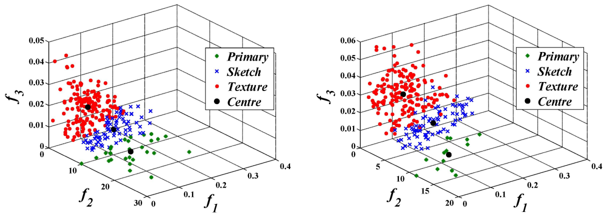
$$f_1 = \bar{L}, \tag{3}$$

$$f_2 = \bar{L} / \left(\bar{M} + \bar{H}\right), \tag{4}$$

where $\bar{L}$, $\bar{M}$ and $\bar{H}$ refer to the mean values of low-frequency, middle-frequency and high-frequency coefficients respectively.

The *Laplacian* operator is applied to obtain another feature in the spatial domain for its high efficiency in dealing with gradient detection applications. The third feature f_3 is the mean value of the *Laplacian* map of a primitive.

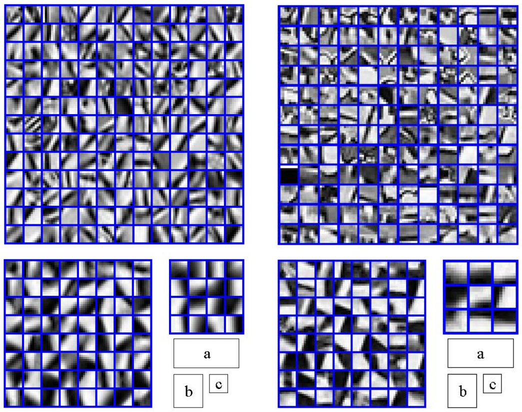
Finally, the three features are combined together as a feature vector notated by $F = \{f_1, f_2, f_3\}$. For each sparse primitive, we extract the feature vector F_i. The k-means algorithm is



(a) Lena

(b) Plane

Fig. 3. 3D plots of primitive classification by 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 centre of each classification.



(a) Lena

(b) Plane

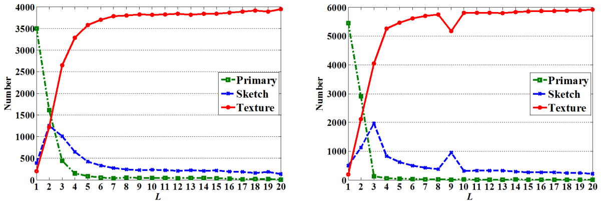
Fig. 4. Primitives classification results of *Lena* (left) and *Plane* (right) images. Three categories: (a) *texture*, (b) *sketch*, (c) *primary*.

then applied to divide all the feature vectors $\{F_i\}$ in the 3-D feature space into three parts, and each part represents a type of primitive. The classification results are shown in Fig. 3 & 4. *Primary* primitives with smooth changes have relatively larger value of f_1 and smaller value of f_2 as well as f_3. On the contrary, *texture* primitives with sharp contrast have relatively smaller value of f_1 and larger value of f_2 as well as f_3. And *sketch* primitives are located between these two kinds of primitives. Note that most of the primitives are classified into the *texture* type, while only small part of primitives belong to the *primary* type.

C. Hierarchical Visual Signal Representation

Base on the primitive classification, the numbers of each type of primitive selected in every iterations are recorded during the sparse reconstruction process (i.e. the *OMP* algorithm). The results are depicted in Fig. 5, from which we can find that: 1) The *primary* primitives dominate in the first iteration though it is the smallest portion of all primitives as shown in Fig. 4. 2) The number of *sketch* primitives is relatively small, and peaks at the 2^{nd} or 3^{rd} iteration. 3) The number of *texture* primitives is contrary to the *primary* type. It is smallest at first and goes up to achieve the maximum after $l = 6$.

It can be concluded that *OMP* scheme decomposes image signal into multiple layers, and these layers are naturally ordered by perceptual importance as illustrated in Fig. 6. The first layer ($l = 1$), which is mostly consisted of *primary* primitives, reconstructs the most of the visual information. Thus we call it *primary layer*. The second layer ($l \in [2 \sim \tilde{l}]$) with more *sketch* primitives recovers more detail visual information of the original signal, so it is named as *sketch layer*. The quantity $\tilde{l}$ corresponds to the threshold between the second and the third layer. The first two layers contain almost all visual information that can be captured by the HVS. Thus the two layers can be combined as the *structural layer*. In contrast with *structural layer*, the remaining *non-structural layer* ($l > \tilde{l}$) is negligible to HVS because there is little correlation with the perceptual experience. This observation highly coincides with the visual perception process, that the primary component (e.g. what object is it) is perceived before details (e.g. what does the object look alike).



(a) Lena

(b) Plane

Fig. 5. The number of three kinds of primitive selected by *OMP* algorithm at each iteration l.

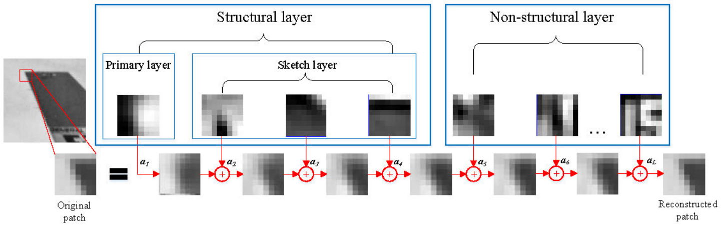


Fig. 6. Illustration of the hierarchical visual signal representation.

D. Visual Information Estimation

The hierarchical structure of visual perception is tightly relevant to the converging *EoP* curve and the *IIID* hypothesis. As the primitives in *IIID* are mostly the *texture* primitives belonging to the non-structural visual layer, these non-structural primitives are much “randomly” distributed with high *EoI* value. By iteratively add this distribution to the sparse primitive system, the total entropy of the build-up distribution (i.e. the *EoP*) tends to go up and finally converge. It is highly related to the generally saturated visual information.

Though the *image-independent* hypothesis of *IIID* leads the saturated visual information in different images to a similar value, the actual information contained in different images is generally distinct. Inspired by [5], we use the $EoP_{\tilde{l}}$ to estimate the visual information of an image, where the $\tilde{l}$ is the threshold between structural and non-structural layers which is given by,

$$\tilde{l} = \arg\min_i \; i, \quad s.t. \quad EoP_i > \overline{EoP}, \qquad (5)$$

where $\overline{EoP} = \frac{\sum EoP_i}{L}$, this mean value threshold can avoid instability caused by outliers and provides more accurate estimation of visual information.

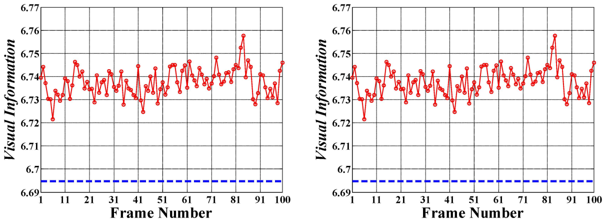
IV. Evaluation Experiments

A. Information Variance Detection for Surveillance Video

In this simulation, we train the background model using the method proposed in [10] for the given surveillance video. Then we evaluate the visual information contained in this background frame and the original frames. Note that the sparse dictionary is trained by the background picture and is used for all frames to make sure that all test frames have the identical basis. The results are given in Fig. 7. We can observe that the visual information provided by background frame (drawn by blue dashed line) is much less than that in original frames (drawn by red solid line), indicating more information in foreground objects. The fluctuation of visual information in *Crossroad* sequence is stronger than *Overbridge*, it is because more moving targets exist in *Crossroad*.

B. Video Shot Boundary Detection

In this subsection, we apply *EoP* for video shot boundary detection. We evaluate the visual information of each frame. A frame is marked as shot boundary when its visual information deviates away from the average of previous video scene. Note that the dictionary should be updated when new shot boundary is detected. The test sequence is a news video with totally 200 frames and 4 different scenes. We give the results in Fig. 8, where the curve of visual information is plotted with different colors and linetypes, indicating different video scenes. It is shown that all the 4 shot boundaries have been detected as the visual information is dramatically changed at shot boundary. Thus it can be concluded that the *EoP* is an efficient tool to accurately estimate the visual information.



(a) Overbridge

(b) Crossroad

Fig. 7. Estimated visual information of surveillance video frames. The dashed blue line represents the visual information of the trained background frame.

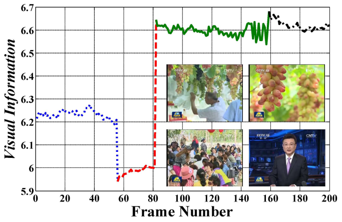


Fig. 8. Results of shot boundary detection. Different shots are with different colors and linetypes. The four different video scenes (the 1^{st}, 56^{th}, 82^{th} and 159^{th} frames respectively) are shown in the right corner.

V. Conclusion

In this paper, we bridge the sparse representation and visual information evaluation with the concept of *EoP*. By analyzing the primitive distribution *EoI*, we propose the *image-independent identical distribution* (*IIID*) hypothesis to interpret the converging *EoP*. Subsequently, by sparse primitive classification, the visual perception is decomposed into a hierarchical representation consisting of structural and non-structural layers. The hierarchical structure of visual perception is tightly relevant to the converging *EoP* curve and generally saturated visual information. Thus it motivates us to use *EoP* to estimate the visual information in natural images. The effectiveness and robustness of the *EoP* are verified in the applications of surveillance video and shot boundary detection.

Acknowledgment

This work was supported in part by the Major State Basic Research Development Program of China (2015CB351800) and in part by the National Science Foundation (61322106, 61390515 and 61210005). This work was also granted by Cooperative Medianet Innovation Center.

References

[1] 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 Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 6, pp. 742–752, 2005.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, “Perceptual video coding based on SSIM-inspired divisive normalization,” *Image Processing, IEEE Transactions on*, vol. 22, no. 4, pp. 1418–1429, April 2013.

[4] S. Wang, X. Zhang, S. Ma, and W. Gao, “Reduced reference image quality assessment using entropy of primitives,” in *Picture Coding Symposium (PCS), 2013*, Conference Proceedings, pp. 193–196.

[5] 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*, Conference Proceedings, pp. 1–6.

[6] 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.

[7] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[8] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[9] X. Zhang, R. Xiong, X. Fan, S. Ma, and W. Gao, “Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity,” *Image Processing, IEEE Transactions on*, vol. 22, no. 12, pp. 4613–4626, Dec 2013.

[10] X. Zhang, T. Huang, Y. Tian, and W. Gao, “Background-modeling-based adaptive prediction for surveillance video coding,” *Image Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 769–784, Feb 2014.