A perceptual quantization strategy for HEVC based on a convolutional neural network trained on natural images

Md Mushfiqul Alam\textsuperscript{a}, Tuan D. Nguyen\textsuperscript{a}, Martin T. Hagan\textsuperscript{a}, and Damon M. Chandler\textsuperscript{b}

\textsuperscript{a}School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK, USA; \textsuperscript{b}Dept. of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Japan

ABSTRACT

Fast prediction models of local distortion visibility and local quality can potentially make modern spatiotemporally adaptive coding schemes feasible for real-time applications. In this paper, a fast convolutional-neural-network based quantization strategy for HEVC is proposed. Local artifact visibility is predicted via a network trained on data derived from our improved contrast gain control model. The contrast gain control model was trained on our recent database of local distortion visibility in natural scenes [Alam et al. JOV 2014]. Furthermore, a structural facilitation model was proposed to capture effects of recognizable structures on distortion visibility via the contrast gain control model. Our results provide on average 11% improvements in compression efficiency for spatial luma channel of HEVC while requiring almost one hundredth of the computational time of an equivalent gain control model. Our work opens the doors for similar techniques which may work for different forthcoming compression standards.

Keywords: HEVC, distortion visibility, convolutional neural network, natural images, visual masking

1. INTRODUCTION

Most digital videos have undergone some form of lossy coding, a technology which has enabled the transmission and storage of vast amounts of videos with increasingly higher resolutions. A key attribute of modern lossy coding is the ability to trade-off the resulting bit-rate with the resulting visual quality. The latest coding standard, high efficiency video coding (HEVC),\textsuperscript{1} has emerged as an effective successor to H.264, particularly for high-quality videos. In this high-quality regime, an often-sought goal is to code each spatiotemporal region with the minimum number of bits required to keep the resulting distortions imperceptible. Achieving such a goal, however, requires the ability to accurately and efficiently predict the local visibility of coding artifacts, a goal which currently remains elusive.

The most common technique of estimating distortion visibility is to employ a model of just noticeable distortion (JND);\textsuperscript{2–4} and such JND-type models have been used to design improved quantization strategies for use in image coding.\textsuperscript{5–8} However, these approaches are based largely on findings from highly controlled psychophysical studies using unnatural videos. For example, Peterson \textit{et al.}\textsuperscript{9} performed an experiment to determine visual sensitivity to DCT basis functions in the RGB and YCrCb color spaces; and, these results have been used to develop distortion visibility models.\textsuperscript{5,10–12} However, the presence of a video could induce visual masking, thus changing the visual sensitivity to the DCT basis functions.

Several researchers have also employed distortion visibility models specifically for HEVC and H.264. Naccari and Pereira proposed a just-noticeable-difference (JND) model for H.264 by considering visual masking in luminance, frequency, pattern, and temporal domains.\textsuperscript{13} However, the coding scheme was based on Foley and Boynton’s\textsuperscript{14} contrast gain-control model which did not consider the gain control in V1 neurons. In another study, Naccari \textit{et al.} proposed a quantization scheme for HEVC to capture luminance masking based on measurements of pixel intensities.\textsuperscript{15} Blasi \textit{et al.} proposed a quantization scheme for HEVC which took into account visual masking by selectively discarding frequency components of the residual signal depending on the input patterns.\textsuperscript{16} However, the scheme of frequency-channel discarding was not based on psychophysical measurements. Other researchers, for example Yu \textit{et al.}\textsuperscript{17} and Li \textit{et al.},\textsuperscript{18} have used saliency information, rather than visual masking to achieve higher compression for HEVC.
Thus, one potential area for improved coding is to develop and employ a visibility model which is not just designed for naturalistic masks (natural videos), but also designed to adapt to individual video characteristics (see, e.g., Refs. 19–23, which aim toward this adaptive approach for images). Another area for improvement is in regards to runtime performance. Operating fuller distortion visibility models in a block-based fashion, so as to achieve local predictions, is often impractical due to prohibitive computational and memory demands. Thus, predicting local distortion visibility in a manner which is simultaneously more accurate and more efficient than existing models remains an open research challenge.

Here, we propose an HEVC-based quantization scheme based on a fast model of local artifact visibility designed specifically for natural videos. The model uses a convolutional-neural-network (CNN) architecture for predicting local artifact visibility and quality of each spatiotemporal region of a video. We trained the CNN model using our recently published database of local masking and quality in natural scenes.24,25 When coupled with an HEVC encoder, our results provided 11% more compression over baseline HEVC when matched in visual quality, while requiring one hundredth the computational time of an equivalent gain control model. Our work opens the doors for similar techniques which may work for different forthcoming compression standards.

2. DISTORTION VISIBILITY PREDICTION MODEL-1: CONTRAST GAIN CONTROL WITH STRUCTURAL FACILITATION (CGC+SF)

This paper describes two separate models for predicting distortion visibility: (1) A model of masking which operates by simulating V1 neural responses with contrast gain control (CGC); and (2) a model which employs a convolutional neural network (CNN). The training data of the CNN model are derived from the CGC model.

Contrast masking has been widely used for predicting distortion visibility in images and videos.26 Among the many existing models of contrast masking, those which simulate the contrast gain-control response properties of V1 neurons are most widely used. Although several gain control models have been proposed in previous studies (e.g., Refs. 26, 28–31), in most cases, the model parameters are selected based on results obtained using either unnatural masks only or a very limited number of natural images.20,22 Here, we have used the contrast masking model proposed by Watson and Solomon.31 This model is well-suited for use a generic distortion visibility predictor because: (1) the model mimics V1 neurons and is therefore theoretically distortion-type-agnostic; (2) the model parameters can be tuned to fall within biologically plausible ranges; (3) the model can directly take images as inputs, rather than features of images as inputs; and (4) the model can operate in a spatially localized fashion.

We improved upon the CGC model in two ways. First, we optimized the CGC model by using our recently developed dataset of local masking in natural scenes (currently, the largest dataset of its kind). Second, we incorporated into the CGC model a structural facilitation model which better captures the reduced masking observed in structured regions.

2.1 Watson-Solomon contrast gain control (CGC) model

This subsection briefly describes the Watson and Solomon contrast gain-control model.31 Figure 1 shows the Watson and Solomon model flow. Both the mask (image) and mask+target (distorted image) go through the same stages: a spatial filter to simulate the contrast sensitivity function (CSF), a log-Gabor decomposition, excitatory and inhibitory nonlinearities, inhibitory pooling, and division.31 After division, the subtracted responses of the mask and mask+target are pooled via Minkowski pooling.31 The target contrast is changed iteratively until the Minkowski-pooled response equals a pre-defined “at threshold” decision (d), at which point, the contrast of the target is deemed the final threshold.

The Watson and Solomon model is a model of V1 simple-cell responses (after “Divide” in Figure 1). Let \( z(x_0, y_0, f_0, \theta_0) \) denote the initial linear modeled neural response (log-Gabor filter output) at location \((x_0, y_0)\), center radial frequency \( f_0 \), and orientation \( \theta_0 \). In the model, the response of a neuron (after division) tuned to these parameters is given by

\[
r(x_0, y_0, f_0, \theta_0) = g \cdot \frac{z(x_0, y_0, f_0, \theta_0)^p}{b^q + \sum_{(x,y,f,\theta) \in I_N} (z(x_0, y_0, f_0, \theta_0))^q},
\]

(1)
where $g$ is the output gain, superscript $p$ provides the point-wise excitatory nonlinearity, $q$ provides the point-wise inhibitory nonlinearity from the neurons in the inhibitory pool, $I_N$ indicates the set of neurons which are included in the inhibitory pool, and $b$ represents the semi-saturation constant.

Watson and Solomon optimized their model parameters using results of Foley and Boynton’s masking experiment using sine-wave gratings and Gabor patterns. However, as mentioned previously, we have recently developed a large dataset of local masking in natural scenes. Here, this dataset was used with a brute-force search to determine the optimum model parameters within biologically plausible ranges. The parameters of the optimized gain control model are given in Table 1.

Table 1. Parameters of the optimized gain control model. Parameters with (*) superscript were varied. Values of the parameters without (*) superscript were found from Watson and Solomon’s work and Chandler et al.’s work. Detailed description of the functions can be found in our previous work.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{CSF}$</td>
<td>Peak frequency</td>
<td>6.0 c/deg</td>
</tr>
<tr>
<td>$BW_{CSF}$</td>
<td>Contrast sensitivity filter bandwidth</td>
<td>1.43</td>
</tr>
<tr>
<td>$N_f^*$</td>
<td>Number of frequency bands</td>
<td>6</td>
</tr>
<tr>
<td>$BW_f^*$</td>
<td>Frequency channel bandwidth</td>
<td>2.75 octave</td>
</tr>
<tr>
<td>$f_{rc}^*$</td>
<td>Center radial frequencies of the bands</td>
<td>0.3, 0.61, 1.35, 3.22, 7.83, 16.1 c/deg</td>
</tr>
<tr>
<td>$N_\theta$</td>
<td>Number of orientation channels</td>
<td>6</td>
</tr>
<tr>
<td>$BW_\theta$</td>
<td>Orientation channel bandwidth</td>
<td>30°</td>
</tr>
<tr>
<td>$\theta_{rc}$</td>
<td>Center angles of orientation channels</td>
<td>0°, ±30°, ±60°, 90°</td>
</tr>
</tbody>
</table>

**Divisive gain control**

| $\alpha$ | Excitatory and inhibitory exponents | 2.4, 2.35 |
| $b$ | Semi-saturation constant | 0.035 |
| $g^*$ | Output gain | 0.1 |
| $s_{\nu^*}$ | Inhibitory pool kernel: space | $3 \times 3$ Gaussian, zero mean, 0.5 standard deviation |
| $s_f^*$ | Inhibitory pool kernel: frequency | ±0.7 octave bandwidth with equal weights |
| $s_\theta$ | Inhibitory pool kernel: orientation | ±60° bandwidth with equal weights |
| $\beta_\nu, \beta_f, \beta_\theta$ | Minkowski exponents: space, frequency, and orientation | 2.0, 1.5, 1.5 |
| $d$ | “At threshold” decision | 1.0 |

### 2.2 Structural facilitation (SF) model

With the abovementioned parameter optimizations, the Watson and Solomon gain control model performed fairly well in predicting the detection thresholds from our local masking database; a Pearson correlation coefficient of 0.83 was observed between the ground-truth thresholds and the model predictions. However, the model consistently overestimated the thresholds for (underestimated the visibilities of) distortions in regions containing recognizable structures. For example, Figure 2 shows how the optimized Watson and Solomon model overestimates the thresholds near the top of the gecko’s body, and in the child’s face.
2.2.1 Facilitation via reduced inhibition

Figure 2 suggests that recognizable structures within the local regions of natural scenes facilitate (rather than mask) the distortion visibility. This observation also occurred in several other images in our dataset (see Ref. 24 for a description of the observed facilitation). Here, to model this “structural facilitation,” we employ an inhibition multiplication factor ($\lambda_s$) in the gain control equation:

$$r(x_0, y_0, f_0, \theta_0) = g \cdot \frac{z(x_0, y_0, f_0, \theta_0)^p}{b^q + \lambda_s \sum_{(x, y, f, \theta) \in \Omega_N} (z(x_0, y_0, f_0, \theta_0))^q}.$$  \hspace{1cm} (2)

The value of $\lambda_s$ varies depending on the strength of structure within an image. Figure 3 shows the change of

![Figure 3](image_url)

Figure 3. The inhibition multiplier $\lambda_s$ varies depending on structure strength. Strong structures give rise to lower inhibition to facilitate the distortion visibility.
inhibition in V1 simple cells depending on the strength of a structure. Recent cognitive studies\textsuperscript{36} have shown active feedback connections to V1 neurons coming from higher level cortices. Our assumption is that recognition of a structure results in lower inhibition in V1 simple cells via increased feedback from higher levels. For the $i^{th}$ block of an image, $\lambda_{si}$ is calculated via the following sigmoid function:

$$
\lambda_{si} = \begin{cases} 
1 - \frac{80}{1} \sum_{x,y=1}^{M,N} \left[ \frac{1}{1 + \exp \left( - \frac{S_i(x,y) - \mu(S,80)}{0.005} \right)} \right], & \text{max}(S) > 0.04 \& \text{Kurt}(S) > 3.5 \\
1, & \text{otherwise,}
\end{cases}
$$

(3)

where $S$ denotes a structure map of an image, $S_i$ denotes the $i^{th}$ block of $S$, $M$ and $N$ denote the block width and height, and $p(S,80)$ denotes the 80\textsuperscript{th} percentile of $S$. The maximum value and kurtosis of the structure map $S$ are used to determine whether there is localized strong structure within the image. The values of the parameters of equation 3 were determined empirically.

### 2.2.2 Structure detection

The structure map $S$ of an image is generated via the following equation which uses different feature maps:

$$
S = L_n \times Sh_n \times E_n \times (1 - D_{\mu_n})^2 \times (1 - D_{\sigma_n})^2.
$$

(4)

$L_n$ denotes a map of local luminance. $Sh_n$ denotes a map of local sharpness.\textsuperscript{37} $E_n$ denotes a map of local Shannon (first-order) entropy. $D_{\mu_n}$ and $D_{\sigma_n}$ denote maps of the average and standard deviation of, respectively, fractal texture features\textsuperscript{38} computed for each local region. Each of these features was measured for 32 × 32-pixel blocks with 50\% overlap between neighboring blocks. Each map was then globally normalized so as to bring all values within the range 0 to 1, and then resized by using bicubic interpolation to match the dimensions of the input image. Examples of structure maps computed for two images are shown in Figure 4.

![Structure maps of two example images.](image)

Figure 4. Structure maps of two example images. The colorbar at right denotes the structure strength at each spatial location of the structure map.

### 2.2.3 Performance of the combined (CGC+SF) model

By incorporating the SF model with the optimized CGC model (as specified in Equation (2)), we were able to improve the model’s predictions for regions of images containing recognizable structures, while not adversely affecting the predictions for other regions. Figure 5 shows some results of using the combined (CGC+SF) model. Observe that the structural facilitation improved the prediction performance in local regions of images containing recognizable structures. For example, in the geckos image, near the top gecko’s skin, the predictions using the CGC+SF model better match the ground-truth thresholds as compared to using only the CGC model. Furthermore, the Pearson correlation coefficients between the CGC+SF model predictions and ground-truth thresholds also improved as compared to using the CGC model alone.

### 3. DISTORTION VISIBILITY PREDICTION MODEL-2: CONVOLUTIONAL NEURAL NETWORK (CNN) MODEL

In terms of accuracy, the CGC+SF model performs quite well at predicting local distortion visibilities. However, this prediction accuracy comes at the cost of heavy computational and memory demands. In particular, the
Figure 5. Structural facilitation improves the distortion visibility predictions in local regions of images containing recognizable structures. Pearson correlation coefficient (PCC) of each map with the experiment map is shown below the map.

CGC+SF model uses a filterbank consisting of 20 log-Gabor filters yielding complex-valued responses to mimic on-center and off-center neural responses. This filtering is quite expensive in terms of memory requirements, which easily exceed the cache on modern processors. In addition, the computation of the CGC responses is expensive, particularly considering that an iterative search procedure is required for the CGC+SF model to predict the thresholds for each block.

One way to overcome these issues, which is the approach adopted in this paper, is to use a model based on a convolutional neural network (CNN). The key advantages of a CNN model are two-fold. First, a CNN model does not require the distorted (reconstructed) image as an input; the predictions can be made based solely on the original image. The distorted images are required only when training the CNN, and it is during this stage that characteristics of particular compression artifacts are learned by the model. The second advantage is that a CNN model does not require a search procedure for its prediction.

3.1 CNN Network architecture

In a recent work we have proposed a CNN architecture (VNet-1) to predict local visibility thresholds for Gabor noise distortions in natural scenes. For VNet-1 the input patch dimension was $85 \times 85$ pixels (approximately $1.9^\circ$ of visual angle), and the training data came from our aforementioned masking dataset.

In this paper, we propose a modified CNN architecture (VNet-2) for predicting local visibility thresholds for HEVC distortions for individual frames of a video. Figure 6 shows the network architecture. The network contains three layers, C1: Convolution, S2: Subsampling, and F3: Full connection. Each layer contains one feature map. The convolution kernel size of the C1 layer is $19 \times 19$. The subsampling factor in layer S2 is 2. The dimensions of the input patch and feature maps of VNet-2 are $M = N = 64$, $d_1^r = d_1^c = 46$, $d_2^r = d_2^c = 23$. 

Proc. of SPIE Vol. 9599  959918-6
3.2 CNN training parameters

VNet-2 contains a total of 894 trainable parameters: The C1 layer contains 362 trainable parameters (19×19 kernel + 1 bias). The S2 layer contains 2 (1 weight + 1 bias) trainable parameters. The F3 layer contains 530 (23×23 weight + 1 bias) trainable parameters.

For training VNet-2, the ground-truth dataset was created using the outputs of the CGC+SF model described in Section 2. Specifically, the CGC+SF model was used to predict a visibility threshold for each 64×64 block of the 30 natural scenes from the CSIQ dataset, and these predicted thresholds were then used as ground-truth for training the CNN model. Clearly, additional prediction error is imposed by using the CGC+SF model to train the CNN model. However, our masking dataset from Ref. 24 contained thresholds for 85×85 blocks, whereas thresholds for 64×64 blocks are needed for use in HEVC. Our CNN should thus be regarded as a fast approximation of the CGC+SF model.

Only the luma channel of each block was distorted by changing the HEVC quantization parameter $QP$ from 0 to 51. The visibility threshold (contrast detection threshold $C_T$) and the corresponding quantization parameter threshold ($QP_T$) were calculated for each of the 1920 patches using the CGC+SF model. Moreover, the number of patches was increased by a factor of four (to a total of 7680 patches) by horizontally and/or vertically flipping the original patches. Flipping a patch would not, in principle, change the threshold for visually detecting distortions within that patch; however, in terms of the network, flipping a patch provides a different input pattern, and thus facilitates learning.

We used 70% of the data for training, 15% for validation, and 15% for testing. The data were randomly selected from each image to avoid image bias. A committee of 10 networks was trained via the Flexible-Image-Recognition-Software-Toolbox using the resilient backpropagation algorithm; the results from the 10 networks were averaged to predict the thresholds.

3.3 Prediction performance of CNN model

The CNN model provided reasonable predictions the thresholds from the CGC+SF model. Figure 7 shows the scatter plots between the CGC+SF thresholds and the CNN predictions. The Pearson correlation coefficients were 0.95 and 0.93 for training+validation and testing data, respectively. From Figure 7, observe that the scatter plots are saturating at both ends of the visibility threshold range. Below −40 dB the patches are very smooth, and above 0 dB the patches are very dark. In both cases, the patches are devoid of visible content, and thus the coding artifacts would be imperceptible if multiple blocks are coded similarly.

3.4 $QP$ prediction using neural network

The CNN model predicts the visibility threshold (detection threshold $C_T$) for each patch. To use the model in an HEVC encoder, we thus need to predict the HEVC quantization parameter $QP$ such that the resulting distortions exhibit a contrast at or below $C_T$. Note the quantization step ($Q_{step}$) in HEVC relates to $QP$ via

$$Q_{step} = (2^{1/6})^{Q_P - 4}. \quad (5)$$
Figure 7. Scatter plots between CGC+SF thresholds and CNN predictions. Example patches at various levels of distortion visibilities are also shown using arrows.

Figure 8. (a) The quantization step ($Q_{step}$) has an exponential relationship with the contrast ($C$) of the HEVC distortion; (b) $\log(Q_{step})$ has a quadratic relationship with $C$.

Figure 8 shows plots of (a) $Q_{step}$ and (b) $\log(Q_{step})$ as a function of the contrast ($C$) of the induced HEVC distortion. Note that $Q_{step}$ shows an exponential relationship with $C$, and $\log(Q_{step})$ shows a quadratic relationship with $C$. Thus, we use the following equation to predict $\log(Q_{step})$ from $C$:

$$\log(Q_{step}) = \alpha C^2 + \beta C + \gamma,$$

where $\alpha$, $\beta$, and $\gamma$ are model coefficients which change depending on patch features. The coefficients $\alpha$, $\beta$, and $\gamma$ can be reasonably well-modeled by using four simple features of a patch: (i) the Michelson luminance contrast, (ii) the RMS luminance contrast, (iii) the standard deviation of luminance, and (iv) the y-axis intercept of a plot of the patch’s log-magnitude spectrum. Descriptions of these features are available in the “Analysis” section of our masking dataset.

To predict $\alpha$, $\beta$, and $\gamma$ from the patch features, we designed three separate committees of neural networks. Each committee had a total of five two-layered feed-forward networks, and each network contained 10 neurons. Each of the networks were trained by using a scaled conjugate gradient algorithm with 70% training, 15% validation, and 15% test data.

Figure 9 shows the scatter plots of $\alpha$, $\beta$, and $\gamma$. In the x-axis values of $\alpha$, $\beta$, $\gamma$ found from fitting the equation 6 are shown, and in the y-axis $\alpha$, $\beta$, $\gamma$ predictions from the feed-forward networks are shown. Observe that the
Figure 9. Scatter plots of (a) $\alpha$, (b) $\beta$, and (c) $\gamma$. The x-axis shows $\alpha$, $\beta$, and $\gamma$ found by fitting Equation (6). The y-axis $\alpha$, $\beta$, and $\gamma$ are predictions from the feed-forward networks.

patch features predict $\alpha$, $\beta$, and $\gamma$ reasonably well, yielding Pearson correlation coefficients of 0.96, 0.97, and 0.97, respectively.

The prediction process of the quantization parameter $QP$ for an input image patch is thus summarized as follows:

- First, predict the distortion visibility ($C_T$) using the committee of VNet-2.
- Second, calculate four features of the patch: (i) Michelson contrast, (ii) RMS contrast, (iii) Standard deviation of patch luminance, and (iv) $y$-axis cut point of log-magnitude spectra.
- Third, using the four features as inputs predict $\alpha$, $\beta$, $\gamma$ from the committees of feed-forward networks.
- Fourth, calculate $\log(Q_{step})$ using equation 6.
- Finally, calculate $QP$ using equation:

$$QP = \max \left( \min \left( \text{round} \left( \frac{\log(Q_{step})}{\log(2^{1/6})} + 4 \right), 51 \right), 0 \right).$$

(7)

Furthermore, we have noticed that when $C_T < -40$ dB the patches are very smooth, and thus can be coded with higher $QP$. Thus, for both the CGC+SF model and CNN model, $QP$ was set to 30 for the patches with $C_T < -40$ dB.

4. RESULTS

In this section we briefly present our preliminary results.

4.1 $QP$ maps

Figure 10 shows the $QP$ maps for two $512 \times 512$ images. The $QP$ maps are generated from CGC+SF and CNN models. Note that the $QP$ maps show how the bits get allocation during encoding process. For example, the dark regions in the redwood tree (in redwood image) and the bush near the wood-log (in the log_seaside image) show higher $QP$ values, denoting those regions will be coded coarsely compared to other regions.
4.2 Compression performance

Both the CGC+SF and CNN models take the $64 \times 64$-pixels image patch as input, and predict the distortion visibility ($C_T$) and corresponding threshold $QP$. Preliminarily we have used the $QP$ maps to test the compression performance using sample images from the CSIQ dataset.

While coding a HEVC coding tree unit, we searched for the minimum $QP$ around the spatial location of that unit within the image, and assigned the minimum $QP$ to that unit. Figure 11 shows example images displaying the compression performance. Note that coding using $QP$ map essentially only add near or below-threshold distortions, and thus judging the quality of the images is quite difficult. However, for comparing with the reference HEVC, a visual matching experiment was performed by two experienced subjects. The purpose of the experiment was to find at which $QP$, the reference software coded image matches with the same quality of the CGC+SF model coded image. After the visually equivalent reference coded image was found, the corresponding bits-per-pixel (bpp) was recorded. In Figure 11 the first column shows the reference images. The second column shows the reference software coded images which are visually equivalent to the images coded using the CGC+SF map. The third column shows the images coded using CGC+SF $QP$ map, and the fourth column shows the images coded using CNN $QP$ map.

Although for near/below threshold distortions subjective experiments are more reliable for assessing the quality, we have reported the structural-similarity index (SSIM) between the reference images and the coded images. Note the qualities of the images are very close. The bpp and compression gains are also shown below the images. On average using $QP$ maps from the CGC+SF and CNN models yielded 15% and 11% gains, respectively. It should be noted that we generated the $QP$ maps mainly from contrast masking and structural facilitation. Thus, if any image does not contain enough local region to mask the distortions, using the $QP$ map yields no gain.

4.3 Execution-time of $QP$ map prediction

One of the reason for introducing a CNN model after the CGC+SF model was that CGC+SF model is too slow to be considered for a real-time application. The execution times of the two models were calculated using a computer having Intel(R) Core(TM)2 Quad CPU Q9400 2.67 GHz processor, 12 GB of RAM, NVIDIA Quadro K5000 graphics card, and running Matlab2014a. The execution times of the two models are calculated to generate $QP$ maps for $512 \times 512$ pixels images with $64 \times 64$ blocks. The execution times for the CGC+SF model and the CNN model were $239.83 \pm 30.564$ sec and $2.24 \pm 0.29$ sec, respectively. Thus, on average the CNN model was 106 times faster compared to the CGC+SF model.
5. CONCLUSIONS AND FUTURE WORKS

This paper described a fast convolutional-neural-network based quantization strategy for HEVC. Local artifact visibility was predicted via a network trained on data derived from an improved contrast gain control model. The
contrast gain control model was trained on our recent database of local distortion visibility in natural scenes. Furthermore, to incorporate the effects of recognizable structures on distortion visibility, a structural facilitation model was proposed which was incorporate with the contrast gain control model. Our results provide on average 11% improvements in compression efficiency for HEVC using the $QP$ map derived from the CNN model, while requiring almost one hundredths of the computational time of the CGC+SF model. Future avenue of our work includes computational time analysis for predicting distortion visibility in real-time videos.

REFERENCES


