https://doi.org/10.24867/GRID-2018-p56 Original scientific paper

CONTENT-AWARE IMAGE COMPRESSION WITH CONVOLUTIONAL NEURAL NETWORKS

Alena Selimović, Aleš Hladnik
University of Ljubljana, Faculty of Natural Sciences and Engineering,
Department of Textiles, Graphic Arts and Design, Ljubljana, Slovenia

Abstract: *Traditional image compression algorithms treat all image regions equally, regardless of their content, often resulting in reconstructed images that do not correlate well with human perception. Content-aware compression, on the other hand, prioritizes image regions that are more relevant to the interpretation of an image and encodes them at a higher bitrate, i.e. without loss or with less loss, than the rest of the image. Our paper explores the multi-structure region of interest (MS-ROI) model, a convolutional neural network, which enables the localization of multiple regions of interest (ROIs) in an image. The localization is expressed as a corresponding saliency map, which identifies the relevance of individual image regions and provides a saliency value for each pixel of the given image. This information is then used to guide the compression. The saliency values are discretized into multiple levels and more important levels are encoded with a higher quality factor Q than the less important ones, allowing for most of the reduction in image resolution to occur in non-salient image regions. Because the generated saliency maps produce soft boundaries between salient and non-salient image regions, smooth transitions between these regions are achieved. The obtained image is then encoded further using the standard JPEG algorithm with a uniform Q factor, resulting in the final image of the standard JPEG format. Our model was trained on the Caltech-101 image dataset and its performance was tested on two other image datasets. Presented are the obtained saliency maps for several images, as well as the results of content-aware compression, which are compared to the standard JPEG compression at different Q factors. For an objective comparison and evaluation of the quality of the obtained images, various standard quality metrics were used, i.e. mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and multi-scale structural similarity index (MS-SSIM).*

Key words: convolutional neural networks, image compression, JPEG, saliency maps, MS-ROI

1. INTRODUCTION

The primary objective of image compression is to reduce the amount of bits that is required for the image data to be stored or transmitted. Traditional lossy image compression algorithms, such as the widely adopted JPEG standard, take advantage of repetitive, redundant or imperceptible image data, and of the limitations of the human visual system, to encode the data more efficiently and reduce the file size. While the goal is to preserve the perceptual quality of an image, the approximation of the represented content results in the development of some unavoidable visual artifacts, which worsen the interpretability of the reconstructed images. Blocking effects, ringing artifacts and blurring, which are most characteristic for JPEG compression, appear as a consequence of the discontinuities between adjacent 8x8 pixel blocks and the elimination of high frequencies. Because the algorithm treats all image regions equally, regardless of their content, the compression artifacts are equally visible in the image background as well as in foreground objects. In order to minimize the visibility of these unwanted artifacts in the decoded images, numerous approaches that focus on obtaining a more accurate reconstruction of the original signal have been proposed (Dong et al, 2015; Dong et al, 2016a; Tao et al, 2017).

Content-aware compression methods, on the other hand, encode the content in a way that corresponds more to the manner in which the human eye interprets the image. Because the degree of human interest in different image regions varies according to what we perceive as more relevant for understanding the image, content-aware compression prioritizes the more important image regions, i.e. the regions of interest (ROIs), and enables them to be preserved with less loss than the rest of the image.

Before the encoding process can be accomplished, a saliency map, corresponding to the image content, needs to be obtained. A saliency map partitions the image into several categories, depending on the image regions they contain, and therefore serves as the means for quantifying the relevance of individual regions. The provided contextual information about the image is then integrated into a compression scheme. Because the selected, more important image regions, are encoded at a higher bitrate than the

image background, the compression artifacts in the ROIs of the reconstructed images are less noticeable than those in the background.

1.1 Saliency maps

Saliency maps can be obtained using different techniques. In recent years convolutional neural networks (CNNs) have been used successfully in a variety of image processing and computer vision tasks and have enabled a more accurate localization, detection and segmentation of objects and ROIs in images. For the purpose of obtaining saliency maps in order to guide the image compression, these approaches are inappropriate due to the following drawbacks:

i) The use of typical methods aimed at object localization results in an object's position being represented within a rectangular window, which does not capture the object's silhouette.

ii) The image segmentation techniques subdivide an image into its constituent regions by classifying each pixel as either a part of a foreground object or a part of the background. The resulting saliency maps produce sharp boundaries between different regions, which is not needed for the purpose of image compression.

Saliency maps can also be based on visual saliency models. While these models accurately capture the fixations of the human eye, the fixations themselves do not encompass the object's edges, which prevents the models from capturing the complete extent of the object. As shown in (Yu et al, 2009), the use of saliency maps based on human fixations for the task of image compression results in blurred edges and a soft focus of the objects in the obtained image.

1.2 The MS-ROI model

The multi-structure region of interest (MS-ROI) model, proposed in (Prakash et al, 2017), is a CNN model that enables the localization of multiple ROIs. The architecture of the model consists of convolutional layers, which are followed by a nonlinear activation function and a max-pooling operation. Fully connected layers, typically added on top of the traditional CNN with the aim of producing the predicted categorical output, are removed and replaced with a GAP layer that applies a global average pooling. The GAP operation calculates the spatial average of each feature map (a three-dimensional tensor) from the convolutional layer preceding the GAP layer, reducing each feature map to a single value. The resulting vector is fed directly into the final, Softmax layer, which outputs the model's prediction. The weights connecting the GAP layer to the output layer encode the contribution of each feature map to the predicted class – the bigger the contribution of a specific detected visual pattern, the more weight it is given. A saliency map is obtained by mapping the weights of the final layer back to the last convolutional layer and calculating a weighted sum of the feature maps. Rather than picking only the most probable class (the highest activation), the activations are sorted from the index of the element with the lowest value to the index of the element with the highest value and a weighted sum of the five highest-scoring classes is taken, while the less probable classes are discarded. By applying a colourmap consisting of a range of cold and warm colours over the obtained greyscale saliency map, the final localization is expressed as a heatmap, which highlights the discriminative ROIs specific to the predicted classes. The most important image regions are represented with the red colour, whereas the least important image regions are represented with the dark blue colour.

The two significant modifications made to the architecture of a traditional deep convolutional neural network - the removal of the fully connected layers and the inclusion of the GAP layer – in addition to choosing the five most probable classes instead of only one, enable the three important advantages of the MS-ROI model in comparison with the approaches of obtaining saliency maps, which are based on localization, segmentation or human fixation.

Firstly, the removal of fully connected layers allows for the model to retain the ability of convolutional layers to behave not only as feature extractors but also as object detectors, despite being trained only on image-level labels with no additional annotation provided (e.g. bounding box supervision or pixel annotation).

Secondly, an addition of the GAP layer provides the means for the network to determine the full extent of the object. If instead of the GAP operation the global max-pooling (GMP) was applied, it would force the model to discard all of the values except the highest one, enabling the model to identify only one discriminative part of the object in an image, but preventing it from highlighting the whole object. The employment of the GMP would therefore present a drawback similar to the usage of a visual saliency model.

Lastly, by using a weighted sum of the top five predictions instead of only the highest scoring class, the obtained heatmap enables the detection of multiple regions of interest, which can include objects belonging to different classes. The localization of multiple ROIs and the production of soft boundaries between different image regions create coherent heatmaps and enable smooth transitions from salient to non-salient regions in the reconstructed images.

1.3 The compression process

The JPEG compression algorithm treats all areas of the image with equal importance, using the same quality factor Q to encode them evenly. The heatmaps obtained by the MS-ROI model, on the other hand, enable a content-aware encoding with a variable factor Q. The generated heatmap provides a saliency value, located in the interval between 0 and 1, for each pixel of the given image. These values are then discretized into a range of levels, varying in their relevance. Pixels with a saliency value of 1 are categorized as the most important, while pixels with a saliency value of 0 are categorized as the least important. Finally, a range of JPEG quality levels from Q_l to Q_h, corresponding to the levels of importance, is chosen.
The compression process involves two encoding passes. In the first encoding pass, the less important levels, indicated by the cold colours on the heatmap, are encoded with lower Q factors (meaning a higher compression), whereas the more important levels, indicated by the warm colours on the heatmap, are encoded with higher Q factors (meaning a lower compression). The areas encoded at higher Q factors are therefore capable of preserving more information about the original image.
The second encoding pass employs a uniform Q factor, Q_{final}, to encode all regions equally. Because the final image is encoded using the standard JPEG encoder, decoding the image can be done by the standard JPEG decoder.

2. METHODS

The code for the MS-ROI model and the image compression was written in Python 3. The processing of the images was performed using the CUDA software on the Nvidia GPU with 8 GB of memory.
The model was trained on the Caltech-101 image dataset (Fei-Fei et al, 2006), which consists of 9144 monochromatic, greyscale and RGB images (representing the independent variables) belonging to 102 classes (representing the dependent variables). The last 15 images of each class were used for the validation set, whereas the remaining images were used for the training set. The training set and the validation set initially contained 7614 images (83.3%) and 1530 (16.7%) respectively, but during the training phase the size of the trainset was increased using real-time image augmentation. The original images were rotated up to 30 degrees, centrally scaled up to 30% and flipped horizontally and vertically. Every iteration produced a number of transformed images from each class that was approximately equal to the number of images that the class originally contained. Since the transformations were perfomed randomly, each iteration included different variations of the input images.
The implementation of the model was based on the pretrained VGG16 model (Simonyan et al, 2014), the architecture of which was modified by removing the fully connected layers at the top of the model and replacing them with three additional convolutional layers and a GAP layer. The 1000 nodes comprising the Softmax layer of the standard VGG16 model were replaced with 102 nodes, corresponding to 102 classes included in the Caltech-101 image dataset. The input images were resized to a fixed input size of 224x224 pixels. 3x3 pixel convolution filters with a stride of 1 and 2x2 pixel pooling windows with a stride of 2 were used for convolution and max-pooling. In total, the neural network consisted of 23 layers – 16 convolutional layers, each followed by a ReLU activation function, 5 max-pooling layers, following each of the 5 blocks of convolutional layers, a GAP layer and a final, Softmax layer. The combination of removing the fully connected layers, thereby decreasing the number of parameters, and adding a GAP layer, which in itself serves as a regularizer, reduced the risk of overfitting the model to the training data largely enough that the dropout was not needed.
Three different methods were used for weight initialization. In the unmodified layers of the VGG16 model the pretrained weights were initialized to the constant numbers. In the convolutional layers, added on top of the VGG16 model, the weights were initialized from the truncated Gaussian distribution using the standard deviation of 0.1, whereas the weights in the GAP layer were initialized using the Gaussian distribution. The reason for using the truncated normal distribution in each of the three additional convolutional layers was to reduce the risk of neuron saturation.

The cost was calculated using the cross entropy function and minimized using the Adam optimization algorithm with a learning rate of 0.0001 and the default values of the parameters beta 1, beta 2 and epsilon. The learning process included 100 iterations using a batch of 32 images.
In order for the model to be able to identify multiple ROIs, the matrix of sorted activations for each input image needed to be obtained. Instead of the more commonly applied argmax method, which finds the index of the element with the maximum activation, the argsort method was used to obtain and sort all the activations. The five highest scoring activations were picked for every input image and their weighted sum was taken. Matplotlib's Jet colourmap was used to generate the colour scheme.
The performance of the content-aware compression method based on the MS-ROI model was assessed on JPEG images from the Salicon dataset (Yu et al, 2015) and on uncompressed BMP images from the General-100 dataset (Dong et al, 2016b). The results of the MS-ROI based compression were compared to the standard JPEG compression at Q factors of 50, 30 and 70. For an objective comparison and evaluation of the quality of the obtained images, the mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and multi-scale structural similarity index (MS-SSIM) were used. Higher PSNR (measured in dB), SSIM and MS-SSIM values refer to a higher image quality, wherease a higher MSE indicates a bigger error in the reconstructed image. For all of the experiments, the chosen Q values for the first encoding pass of the MS-ROI compression method ranged from Q_l = 30 to Q_h = 70. The Q_{final} factor of the second encoding pass depended on the selected Q factor of the JPEG compression and the file size of the original image. When comparing the MS-ROI compression to the standard JPEG compression at Q = 50, the average Q_{final} was 57, at Q = 30 the average Q_{final} was 31, and at Q = 70 the average Q_{final} was 73. The maximum difference between the file sizes of the standard JPEG images and the images obtained using the MS-ROI model was 1%.

3. RESULTS

3.1 The accuracy and reproducibility of saliency maps

Because the quality of the final image is heavily dependent on the accuracy and reproducibility of the obtained heatmaps, the model was evaluated on input images representing different content and containing different semantic objects. Since the model's prediction (the matrix of activations) for the same input image is slightly different every time the image is passed through the network, different variations of the heatmaps are generated. To show that the model's prediction for the same input image varies only to some degree and to estimate the robustness of the model, as well as the reproducibility of the obtained heatmaps, some of the test images were passed through the network five times in order to obtain five different variations of the heatmap (shown in Figures 1 – 4).
As seen in Figures 1 and 2, the model is able to accurately identify one or a few clearly distinguishable salient image regions. Consequently, the similarity between the produced heatmaps is relatively high. In cases where the identification of the ROIs and the isolation of foreground objects from the background are more difficult (Figure 3) or where the input images do not contain any salient regions at all (Figure 4), the precision and the reproducibility of the generated heatmaps are lower.

Figure 1: Example of an input image with one salient region

Figure 2: Example of an input image with two salient regions

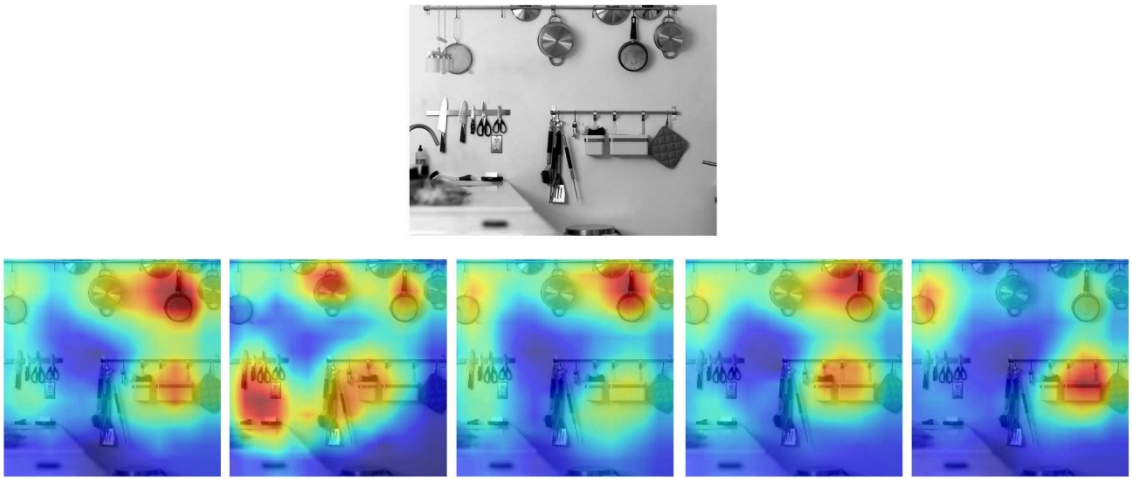

Figure 3: Example of an input image with several salient regions

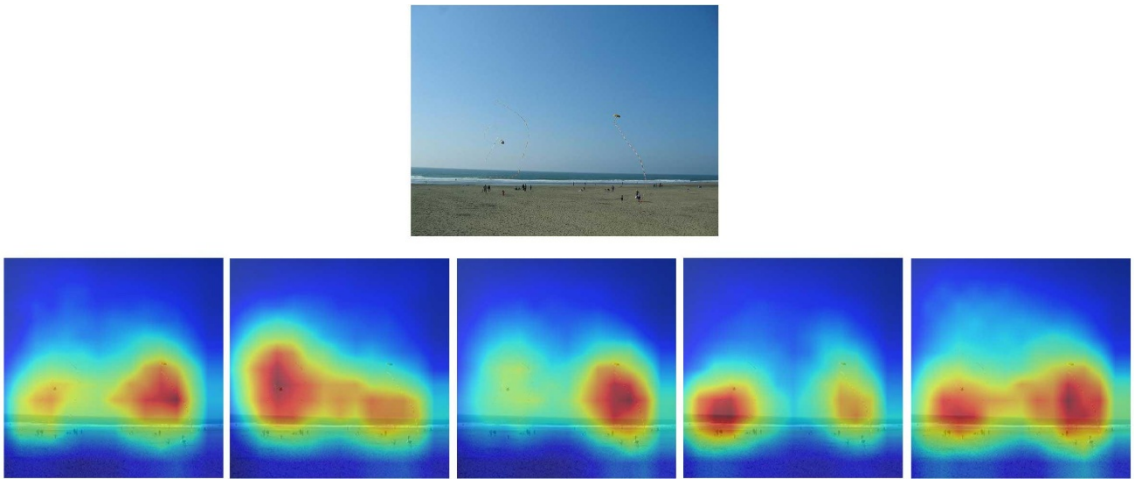

Figure 4: Example of an input image without salient regions

3.2 Objective evaluation of the quality of the reconstructed images

Table 1 presents the results of calculated quality metrics for the set of five images that contain distinct salient areas (Figure 5). The compression method based on the MS-ROI model outperforms the standard JPEG compression at Q = 50 for all five images, even though the file sizes of images that were compressed

using the MS-ROI method are all slightly smaller than those compressed using the JPEG algorithm. The most significant gain in PSNR, 1.61 dB, is achieved in the case of image 2 (snowboarder), the biggest improvement in SSIM, 0.0107, in the case of image 1 (bird), and the biggest gain in MS-SSIM, 0.0044 in the case of image 5 (elephant).

Figure 5: Five examples of images from the Salicon dataset and of their corresponding heatmaps

Table 1: Results of calculated quality metrics when comparing the MS-ROI compression to JPEG at Q = 50 – for the images with salient regions

	MSE	PSNR (dB)	SSIM	MS-SSIM	File size (B)
	Image 1 (bird)				
JPEG	35.134	32.674	0.9616	0.9925	47436
MS-ROI	26.932	33.828	0.9723	0.9947	47014
	Image 2 (snowboarder)				
JPEG	21.099	34.888	0.9590	0.9878	22171
MS-ROI	14.571	36.496	0.9669	0.9886	21951
	Image 3 (truck)				
JPEG	114.50	27.543	0.9201	0.9846	56867
MS-ROI	80.481	29.074	0.9290	0.9878	56362
	Image 4 (cathedral)				
JPEG	46.729	31.435	0.9233	0.9848	41895
MS-ROI	32.479	33.015	0.9261	0.9864	41669
	Image 5 (elephant)				
JPEG	52.656	30.916	0.9230	0.9824	42718
MS-ROI	37.492	32.391	0.9336	0.9868	42301

The results of calculated quality metrics for the set of five images, which do not contain any salient areas or where their identification is more ambiguous (Figure 6), are displayed in Table 2. Because content-aware compression is intended for encoding images that depict some salient objects, the model's predictions were expected to be much less accurate in cases where the input images contained patterns, shapes or textures. Nonetheless, the PSNR values of the MS-ROI compression method turned out to be

higher than those of the JPEG compression for all images, with the exception of the last one (corals). Unlike the PSNR value, the SSIM and MS-SSIM indexes were improved only for image 1 (stone wall) and image 3 (leaves).
It is worth mentioning that, even in the cases where any of the PSNR, SSIM or MS-SSIM values of the MS-ROI compression were higher than for the JPEG compression, the improvement of the MS-ROI method over the JPEG method is much less significant than in the case of the images that contain clearly identifiable salient objects.

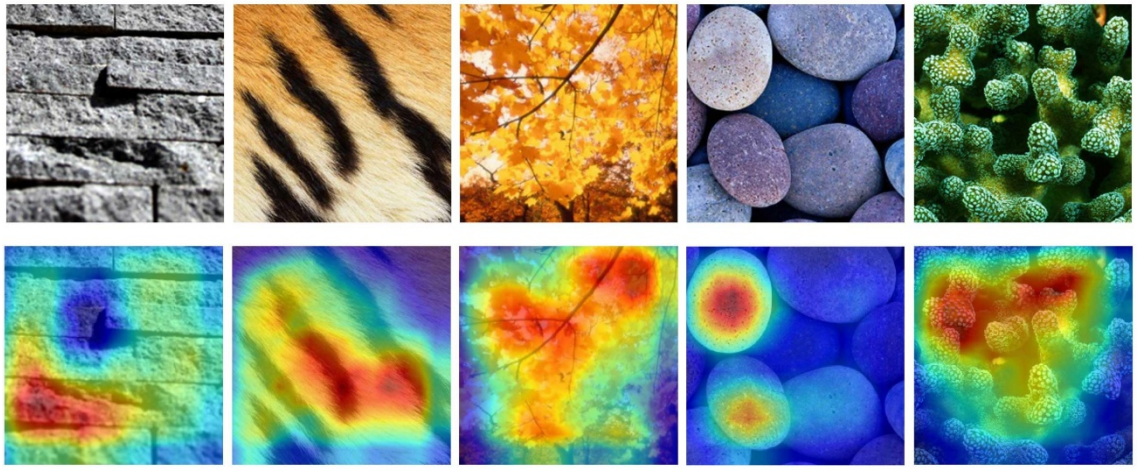

Figure 6: Five examples of images from the General-100 dataset and of their corresponding heatmaps

Table 2: Results of calculated quality metrics when comparing the MS-ROI compression to JPEG at Q = 50 – for the images without salient regions

	MSE	PSNR (dB)	SSIM	MS-SSIM	File size (B)
	Image 1 (stone wall)				
JPEG	94.637	28.370	0.9639	0.9962	14331
MS-ROI	94.425	28.380	0.9644	0.9963	14457
	Image 2 (tiger stripes)				
JPEG	96.061	28.305	0.8649	0.9812	24809
MS-ROI	94.273	28.387	0.8623	0.9808	24805
	Image 3 (leaves)				
JPEG	25.258	34.107	0.9516	0.9897	23126
MS-ROI	25.201	34.117	0.9518	0.9899	23261
	Image 4 (pebbles)				
JPEG	86.153	28.778	0.8672	0.9828	30722
MS-ROI	84.485	28.863	0.8657	0.9827	31003
	Image 5 (corals)				
JPEG	48.866	31.241	0.9584	0.9915	68008
MS-ROI	55.054	30.723	0.9571	0.9908	68278

The model was also evaluated on 200 randomly chosen images from the Salicon dataset, which contains a total of 20000 images, and on 50 randomly chosen images from the General-100 dataset, which contains

a total of 100 images. The average PSNR and SSIM values for the selected images from the Salicon and General-100 dataset are shown in Table 3 and Table 4, respectively.

Table 3: PSNR and SSIM of 200 images from the Salicon dataset

	PSNR	SSIM
JPEG (50)	32.670	0.9658
MS-ROI	33.763	0.9735
JPEG (30)	29.521	0.8991
MS-ROI	30.134	0.9021
JPEG (70)	43.445	0.9910
MS-ROI	44.416	0.9958

Table 4: PSNR and SSIM of 50 images from the General-100 dataset

	PSNR	SSIM
JPEG (50)	32.827	0.9501
MS-ROI	33.184	0.9537
JPEG (30)	30.637	0.9375
MS-ROI	30.641	0.9379
JPEG (70)	34.902	0.9741
MS-ROI	34.968	0.9759

Results show that the MS-ROI compression method performs better than the standard JPEG compression, since the former is characterized by a higher PSNR and a better visual quality of the obtained images. As seen in Table 3, compression based on the MS-ROI model achieves an average gain of 1.09 dB in PSNR and a 0.0077 gain in SSIM against the standard JPEG compression at Q = 50, and an average gain of 0.97 dB in PSNR and 0.0048 in SSIM at Q = 70. Meanwhile, the improvement of the MS-ROI compression, when compared to the JPEG compression at Q = 30, is not as substantial, though the MS-ROI method still performs better and on average improves the PSNR by 0.61 dB and the SSIM by 0.0030.

Similarly, the MS-ROI based compression generates better results in comparison with the standard JPEG compression at Q = 50 for the images from the General-100 dataset. On average, the MS-ROI compression gains 0.36 dB in PSNR and 0.0036 in SSIM. However, when compared to the JPEG compression at Q = 30 and Q = 70, the results of the MS-ROI compression are much less prominent. When compared to the JPEG compression at Q = 70, on average, the MS-ROI compression improves the PSNR by 0.07 dB and the SSIM by 0.0018, while, when compared to the JPEG compression at Q = 30, the average gain in PSNR is only 0.0040 dB and only 0.0004 in SSIM.

Overall, the MS-ROI model performs better on images from the Salicon dataset. This can be explained by the fact that the General-100 dataset contains more images depicting textures and patterns compared to the Salicon dataset, which consists mostly of images of natural indoor and outdoor scenery with various salient regions. Furthermore, the General-100 dataset contains images of smaller dimensions than the Salicon dataset, which is therefore better suited for a content-aware compression task.

The obtained results were interpreted using a one-way analysis of variance (Anova). The purpose of the test was to determine whether the MS-ROI compression and the JPEG compression are actually different in the measured characteristics. Since the improvement of the MS-ROI compression over the JPEG compression was more significant for the images from the Salicon dataset than for the images from the General-100 dataset, Anova was performed for the former dataset. An improvement of the MS-ROI model over the standard JPEG compression is statistically significant if the p-value is less than the

significance level of 0.05. The test yielded p-values that were lower than the significance level, for both the PSNR and SSIM values, when comparing the MS-ROI model to the JPEG compression at Q factors of 30, 50 and 70, thereby rejecting the null hypothesis in favour of the MS-ROI model.

4. CONCLUSIONS

This paper explores the content-aware compression based on saliency maps obtained using a convolutional neural network - the MS-ROI model. We showed that by varying the quantization of compression, the MS-ROI based encoding is capable of achieving a better visual quality of the reconstructed images compared to the standard JPEG compression. Since the accuracy of the obtained heatmaps is highly dependent on the content of the input image, the performance of the MS-ROI compression is especially superior when images with clear semantic regions are used. Because the model allows for the detection of multiple salient image regions and produces soft boundaries between them, the transitions from regions encoded at higher and lower bitrates are smooth. Further experimentation with the MS-ROI model based on different CNN architectures is required to better understand the effect of the model implementation on the generated saliency maps and the resulting quality of the content-aware compression.

5. REFERENCES

[1] Dong, C., Deng, Y., Loy, C. C., Tang, X.: "Compression artifacts reduction by a deep convolutional network", Proceedings of IEEE International Conference on Computer Vision (ICCV) 2015 (Santiago, Chile, 2015), pages 576–584. doi:10.1109/ICCV.2015.73.

[2] Dong, C., Loy, C. C., He, K., Tang, X.: "Image super-resolution using deep convolutional networks", IEEE Transactions on Pattern Analysis and Machine Intelligence 38(2), 295-307, 2016, doi:10.1109/TPAMI.2015.2439281.

[3] Dong, C., Loy, C. C., Tang, X.: "Accelerating the super-resolution convolutional neural network", Computer Vision – ECCV 2016 (Springer International Publishing, 2016), 391-407. doi: 10.1007/978-3-319-46475-6_25.

[4] Fei-Fei, L., Fergus, R., Perona, P.: "One-Shot learning of object categories", IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 594-611, 2006. doi: 10.1109/TPAMI.2006.79.

[5] Prakash, A., Moran, N., Garber, S., Dilillo, A., Storer, J.: "Semantic perceptual image compression using deep convolution networks", Proceedings of Data Compression Conference (DCC) 2017 (Snowbird, UT, USA, 2017), pages 250-259. doi: 10.1109/DCC.2017.56.

[6] Simonyan, K., Zisserman, A.: "Very deep convolutional networks for large-scale image recognition", Proceedings of International Conference on Learning Representations 2015 (ICLR , San Diego, CA, 2015), pages 1-14.

[7] Tao, W., Jiang, F., Zhang, S., Ren, J., Shi, W., Zuo, W., Guo, X., Zhao, D.: "An end-to-end compression framework based on convolutional neural networks", Proceedings of Data Compression Conference (DCC) 2017 (Snowbird, UT, USA, 2017), page 463. doi: 10.1109/DCC.2017.54.

[8] Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: "LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop", 2015, URL https://arxiv.org/abs/1506.03365 (last request: 2018-09-20).

[9] Yu, S. X., Lisin, D. A.: "Image compression based on visual saliency at individual scales", Proceedings of ISVC: International Symposium on Visual Computing, Advances in Visual Computing 2009 (5th International Symposium ISVC Las Vegas, NV, USA, 2009), pages 157-166. doi: 10.1007/978-3-642-10331-5_15.