THERMALNET: A DEEP CONVOLUTIONAL NETWORK FOR SYNTHETIC THERMAL IMAGE GENERATION

V. V. Kniaz[a,b]*, V. S. Gorbatsevich[a], V. A. Mizginov[a]

[a] State Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia - (vl.kniaz, gvs, vl.mizginov)@gosniias.ru
[b] Moscow Institute of Physics and Technology (MIPT), Russia

Commission II, WG II/5

KEY WORDS: infrared images, augmented reality, object recognition, deep convolutional neural networks

ABSTRACT:

Deep convolutional neural networks have dramatically changed the landscape of the modern computer vision. Nowadays methods based on deep neural networks show the best performance among image recognition and object detection algorithms. While polishing of network architectures received a lot of scholar attention, from the practical point of view the preparation of a large image dataset for a successful training of a neural network became one of major challenges. This challenge is particularly profound for image recognition in wavelengths lying outside the visible spectrum. For example no infrared or radar image datasets large enough for successful training of a deep neural network are available to date in public domain. Recent advances of deep neural networks prove that they are also capable to do arbitrary image transformations such as super-resolution image generation, grayscale image colorisation and imitation of style of a given artist. Thus a natural question arise: how could be deep neural networks used for augmentation of existing large image datasets? This paper is focused on the development of the Thermalnet deep convolutional neural network for augmentation of existing large visible image datasets with synthetic thermal images. The Thermalnet network architecture is inspired by colorisation deep neural networks.

1. INTRODUCTION

Thermal cameras provide a robust solution for object detection and scene understanding. As the thermal vision is robust against degraded visual environments such as fog, dust or night time it is widely used for driver support systems such as an enhanced vision system. Thermal imaging is also widely exploited in the field of autonomous driving, where it helps to improve object detection rates significantly. However the appearance of objects in thermal images could change greatly for different weather conditions. Thus a powerful object detection algorithm is required to detect and recognise objects in thermal images.

Deep convolutional neural networks (CNN) have significantly changed the landscape of the modern computer vision. Nowadays methods based on deep neural networks show the best performance among image recognition and object detection algorithms. CNN also provide flexible solution for object detection in multispectral images. For a successful learning of a CNN a large training dataset with thousands of images is required. An intensive scholar attention to the field of CNN stimulated the development of extremely large image datasets with ground truth labelling of tens million of images (Deng et al., 2009, Lin et al., 2014, Everingham et al., 2015). Most of datasets that are available online include only images captured in visible spectrum and could not be used for training of multispectral CNN. Existing multispectral datasets are either not available in public domain(Weber and Penn, 2005) or have imperfect geometrical alignment (Hwang et al., 2015) (figure 1).

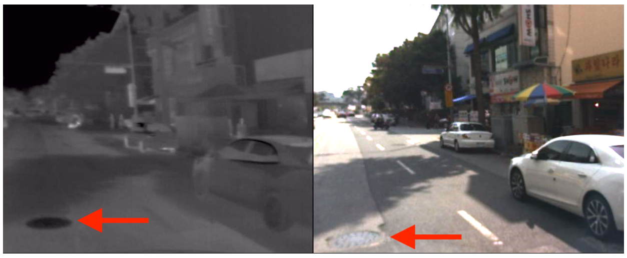


Figure 1. An example of imperfect geometrical alignment in 'Multispectral Pedestrian Dataset'. Note the position of the manhole on the infrared frame (left) and the visible frame (right)

The key step in the development of the new generation of multispectral object detection algorithms using CNN is the generation of large multispectral datasets. The direct creation of such datasets using experiments is time consuming and hardly could guarantee the required variety of images and object classes. 3D modelling provide a flexible solution for synthetic thermal image generation (Kniaz et al., 2016). The main drawback of this technique is a highly time consuming 3D model generation step that is done manualy. Also noise and distortion of real sensors is absent on synthetic images (figure 2).

This paper is focused on the development of a CNN for transformation of visible range images into infrared images. The developed CNN is based on the SqweezeNet CNN. (Forrest N. Iandola, 2016). The CNN was trained using NVIDIA DIGITS. (NVIDIA, 2016)

2. RELATED WORK

The first mention of the use of deep convolutional neural networks for image generation appeared in 2013. (Zeiler and Fergus, 2013). In this article, the authors proposed a method for visualising a trained network based on the selection of an image that gives the maximum response of a given filter. It is commonly

*Corresponding author

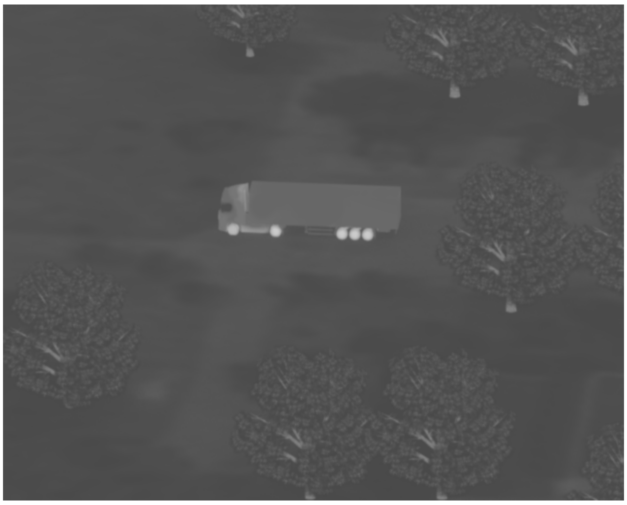


Figure 2. An example of synthetic thermal image generated using 3D modelling

known, that CNN is most often used for image recognition. However (Zhang et al., 2016) showed that they can also be used to colorise monochrome images (figure 3).

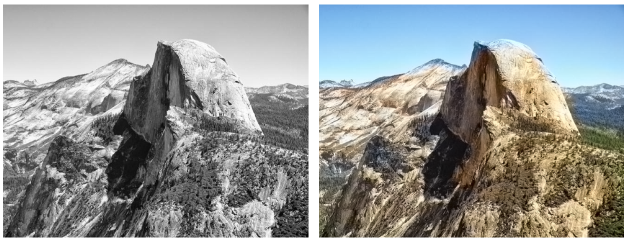


Figure 3. Creating a color image from a monochrome using CNN

Recently a network has been developed capable of simulating various artistic styles (Zhang et al., 2016, Gatys et al., 2015, Deshpande et al., 2015) An image containing the reference style, and the source image is given as an input for the network. After that, the image is generated using the stochastic gradient descent (figure 4). The style is taken from outputs of the filters for a given reference image, and the content is taken from outputs of filters for the original image. By varying network parameters, you can adjust the predominance of either the style or the content in the new image.

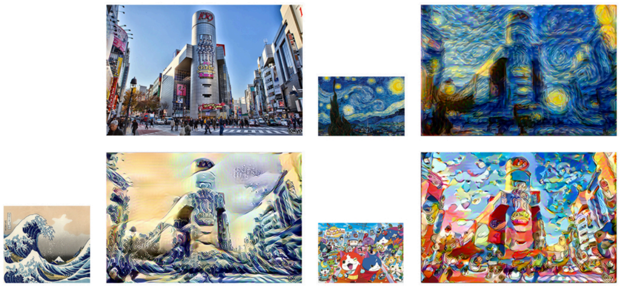


Figure 4. Imitation of different artistic styles

The question arises: is it possible to use a CNN to transform images from one spectral to another? In (Limmer and Lensch, 2016), a method is proposed for converting near-infrared images to visible images (figure 5). This paper presents the deep CNN for transformation of visible images to infrared images.

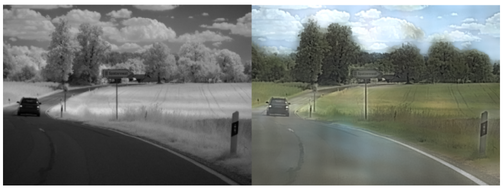


Figure 5. Converting an infrared image to a visible range image

3. APPROACH

The proposed method of image transformation is based on the use of a CNN for semantic image segmentation. A great number of CNN architectures were developed for image classification. Semantic image segmentation requires significant changes in CNN architectures. Such architectures are commonly known as 'fully convolutional' networks (Long et al., 2015) with no fully-connected layers (Long et al., 2016). In addition, the deconvolution layers (Hyeonwoo et al., 2015) are widely used to solve the problem of semantic segmentation.

3.1 Objective function

Given an input color image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, our objective is to learn a mapping $\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{X})$ to thermal emission $\mathbf{Y} \in \mathbb{R}^{H \times W}$, where H, W are image dimentions. We use multinomial cross entropy loss L_cl, defined as

$$L_cl(\hat{\mathbf{Y}}, \mathbf{Y}) = -\sum_{h,w} v(\mathbf{Y}_{h,w}) \sum_{q} \mathbf{Y}_{h,w,q} \log(\hat{\mathbf{Y}}_{h,w,q}) \quad (1)$$

The per-pixel, unnormalised softmax loss provides a good performance for segmenting images of various sizes into disjoint classes. The key idea of the softmax operation is the competition between classes. The competition promotes the most confident prediction. Another option is to train the network with the sigmoid cross-entropy loss. In (Zhang et al., 2016) it was shown that the sigmoid cross-entropy loss gives similar results, even though it normalises each class prediction independently.

3.2 Deep CNNs

The SqweezeNet CNN was developed in 2016. According to the authors, while preserving the accuracy of AlexNet (Krizhevsky et al., 2012), its a performance is 50 times higher. This became possible due to the replacement of the convolution with 3×3 filters by the convolution with 1×1 filters. Such replacement reduces the number of parameters by a factor of 9. The input of the remaining 3×3 filters is sampled only by a small number of channels. The size reduction is done as late as possible so that the convolution layers have a large activation area. These three strategies have led to the creation of the so-called 'fire module'. The entire network is constructed using this modules. The architecture of the Thermalnet network (figure 6) has the following contributions. Firstly, two deconvolution layers were added to restore the spatial resolution of the input image. Secondly, a global avgpool layer was removed to reduce the number of parameters.

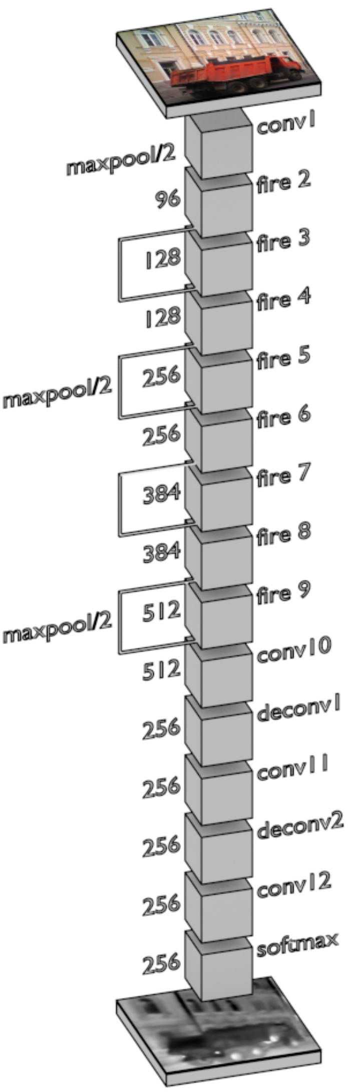


Figure 6. The architecture of the Thermalnet neural network

3.3 Framework

CNN training was done using the NVIDIA Digits environment. It is an open source software designed to perform research on the design and training of deep neural networks. DIGITS provides capabilities for training on large data sets locally and from the remote computer. The DIGITS also provides visualisation of neural networks and manages previously obtained results and models for comparison with new ones.

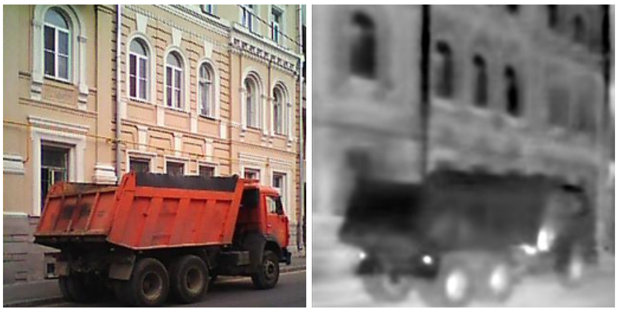
3.4 Postprocessing

After the infrared images were generated, the postprocessing was performed using the algorithm proposed in (Gatys et al., 2015). For this, a trained VGG-16 (Simonyan and Zisserman, 2014) network was used which was used. The network serves as a measure of similarity during the iterative generation of the output image that must match a given reference image. The network was operated using the Torch7 (Collobert R., 2011) library. The method of imitating a style using a deep convolutional network is based on an iterative selection of the required image, in which the network acts as a measure of the similarity of the 'style'. The generation of a new image with the matching style is done using a gradient descent. The initial image is initialised with a Gaussian white noise. After that the initial image changes until it produces the same response in a specific layer of the network as the original image. The post-processing operation was performed to reduce the quadratic loss error between the resulting image and the reference image.

4. DATASET

4.1 Dataset design

The training dataset consists of 1000 pairs of geometrically aligned pairs of television and thermal imaging images of various objects (figure 7). The choice of scenes and objects was due to the presence of a significant thermal contrast between the object of interest and the environment. It is obvious that multiply correct thermal images could be produced for a given colour image if the temperature of the object will be changed. However in the scope of training image dataset augmentation problem any thermal image that could possibly correspond to the given colour image will be the correct solution. Moreover, if there will be more different (but physically possible) thermal images in the training dataset the performance of the trained recognition algorithm will improve. Visible range images were used as inputs to the CNN. Infrared images served as the ground truth.



(a) TV image (b) IR image

Figure 7. Example of a pair of images from the training sample

4.2 Dataset generation

To create a training sample, the FLIR One portable thermal imaging camera was used. Its technical specification are presented in table 1.

Parameter	Value
Visible range resolution	640x480
Infrared resolution	160x120
Field of view	35x45
Temperature range	-20…120 C
The spectral range	8 – 14 μm
Pixel size	12 μm

Table 1. FLIR One camera parameters

FLIR One is a portable device that has a visible range and thermal imaging camera and connects to a smartphone. The thermal imager has a built-in battery, which allows it to work up to 45 minutes. The database was both indoors and outdoors under the same weather and temperature conditions.

5. EXPERIMENTS

The network was trained using a Titan X PASCAL captured GPU and was 1000 epochs. In figure 8 the graph of a validation loss is presented.

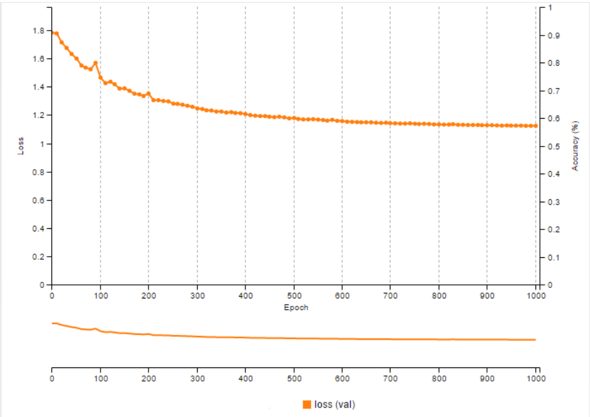


Figure 8. Validation loss for the training dataset

5.1 CNN evaluation

The testing of the trained network was carried out using the test dataset including 100 image pairs. The degree of similarity between the synthetic images and the original infrared images was estimated using the test dataset. The root mean square error (RMS) metric was used to measure the difference between real infrared and synthesised images. The RMS is given by

$$E_{\text{rms}} = \sqrt{\frac{\sum_{h,w}(\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w})^2}{h \cdot w}}, \tag{2}$$

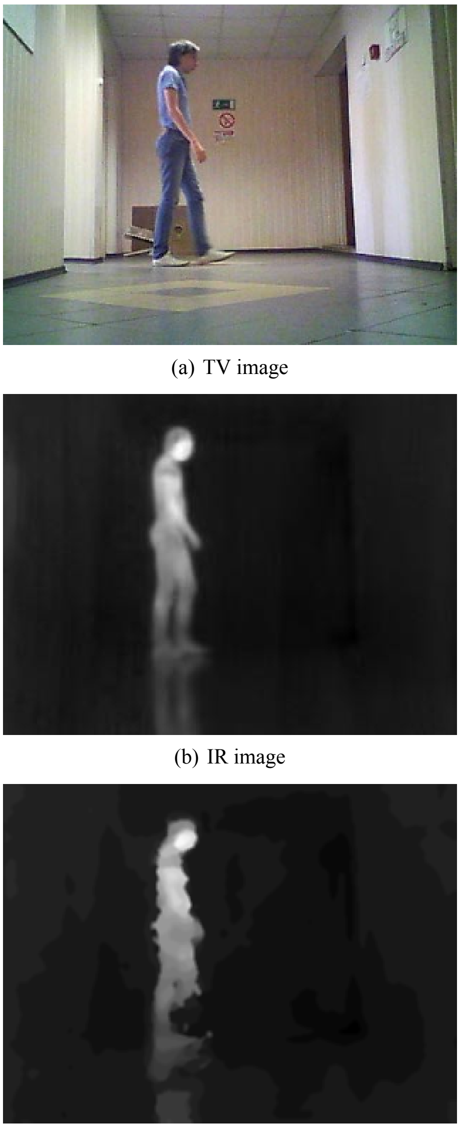
where $\mathbf{Y}$ is brightness of the source infrared image, $\hat{\mathbf{Y}}$ - brightness of the syntetic infrared image, h, w are image dimensions. The average RMS for the training dataset is present in table 2. The network was capable to corrected predict the thermal emission of the objects. However a significant presence of high frequency components that were absent on original thermal image. Examples of generated images are shown in figures 9-10.

5.2 Postprocessing evaluation

The evaluation of the post processing step was performed using the same test dataset. For each synthetic thermal image a new image was generated using the style reconstruction. An original thermal image was used as the source for the style. A RMS was computed using equation (2) to evaluate the performance of the postprocessing. An evaluation have shown that the style reconstruction failed to imitate the smoothness and distortion present in original thermal images. The results of evaluation are summarised in table 2.

Method Object	TN	TN + PP	TN + GS
Human	9.1	31.9	7.5
Microwave oven	37.8	60.4	29.2
Computer	16.7	36.5	13.5

Table 2. Average RMS. TN – Thermalnet, TN + PP – Thermalnet + postprocessing, TN+ GS – Thermalnet + gaussian smooth 3x3



(a) TV image

(b) IR image

(c) Syntetic IR image

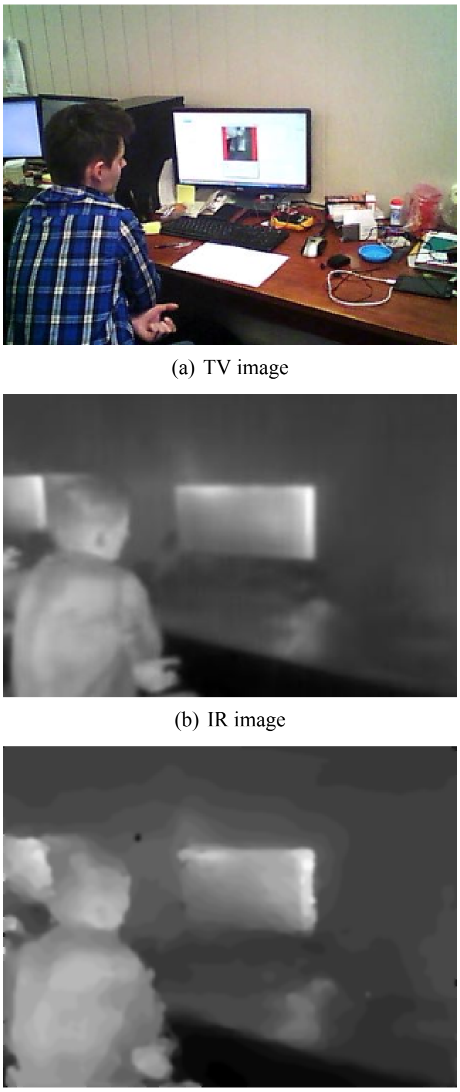
Figure 9. Examples of generated images

6. CONCLUSION

A deep convolutional network for synthetic thermal image generation was developed. The network is based on the SqueezeNet deep convolutional network. The original architecture was modified and supplemented with deconvolution layers. The network architecture for the NVIDIA DIGITS platform was written. To train the network a training dataset was collected using the FLIR ONE thermal camera. The training dataset consists of 1000 pairs of geometrically aligned pairs of visible spectrum and infrared images of various objects. The network was trained using the training dataset. The final loss during the training stage was equal to 0.5%. The network performance was evaluated using the test dataset including 100 image pairs. The evaluation have shown that the network is capable to correctly recover the thermal emission of the objects that were present in the training dataset.

7. ACKNOWLEDGEMENTS

The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project № 16-38-



(a) TV image

(b) IR image

(c) Syntetic IR image

Figure 10. Examples of generated images

00940 мол_а and by Russian Science Foundation (RSF) according to the research project № 16-11-00082.

REFERENCES

Collobert R., Kavukcuoglu K., F. C., 2011. Torch7: A Matlab-like Environment for Machine Learning. *NIPS Workshop*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, IEEE, pp. 248–255.

Deshpande, A., Rock, J. and Forsyth, D., 2015. Learning Large-Scale Automatic Image Colorization. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, pp. 567–575.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision* 111(1), pp. 98–136.

Forrest N. Iandola, Song Han, M. W. M., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size.

Gatys, L. A., Ecker, A. S. and Bethge, M., 2015. A Neural Algorithm of Artistic Style. *CoRR*.

Hwang, S., Park, J., Kim, N., Choi, Y. and Kweon, I. S., 2015. Multispectral pedestrian detection - Benchmark dataset and baseline. *CVPR* pp. 1037–1045.

Hyeonwoo, N., Seunghoon, H. and Bohyung, H., 2015. Learning Deconvolution Network for Semantic Segmentation. *Department of Computer Science and Engineering, POSTECH, Korea ICCV2015*.

Kniaz, V., Gorbatsevich, V. and Mizginov, V., 2016. Generation of synthetic infrared images and their visual quality estimation using deep convolutional neural networks. *Scientific Visualization* 8(4), pp. 67–79.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

Limmer, M. and Lensch, H. P. A., 2016. Infrared Colorization Using Deep Convolutional Neural Networks. *CoRR abs/1501.02565*.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P., 2014. Microsoft COCO: Common Objects in Context. *ArXiv e-prints*.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Long, J., Shelhamer, E. and Darrell, T., 2016. Fully Convolutional Models for Semantic Segmentation. *CVPR 2015, and PAMI 2016*.

NVIDIA, 2016. NVIDIA deep learning gpu training system. `https://developer.nvidia.com/digits`. Accessed: 2016-04-01.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*.

Weber, B. A. and Penn, J. A., 2005. Synthetic FLIR Signatures for Training and Testing Target Identification Classifiers. *Sensors and Electron Devices Directorate,ARL*.

Zeiler, M. D. and Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. *arXiv.org* p. arXiv:1311.2901.

Zhang, R., Isola, P. and Efros, A. A., 2016. Colorful Image Colorization. *ECCV* 9907(Chapter 40), pp. 649–666.