Glioma Segmentation with Cascaded UNet

Dmitry Lachinov[1,2](✉), Evgeny Vasiliev[1], and Vadim Turlapov[1]

[1] Lobachevsky State University, Gagarina ave. 23, 603950
Nizhny Novgorod, Russian Federation
dlachinov@gmail.com, eugene.unn@gmail.com, vadim.turlapov@gmail.com

[2] Intel, Nizhny Novgorod, Russian Federation
dmitry.lachinov@intel.com

Abstract. MRI analysis takes central position in brain tumor diagnosis and treatment, thus its precise evaluation is crucially important. However, its 3D nature imposes several challenges, so the analysis is often performed on 2D projections that reduces the complexity, but increases bias. On the other hand, time consuming 3D evaluation, like segmentation, is able to provide precise estimation of a number of valuable spatial characteristics, giving us understanding about the course of the disease.

Recent studies focusing on the segmentation task, report superior performance of Deep Learning methods compared to classical computer vision algorithms. But still, it remains a challenging problem. In this paper we present deep cascaded approach for automatic brain tumor segmentation. Similar to recent methods for object detection, our implementation is based on neural networks; we propose modifications to the 3D UNet architecture and augmentation strategy to efficiently handle multimodal MRI input, besides this we introduce approach to enhance segmentation quality with context obtained from models of the same topology operating on downscaled data. We evaluate presented approach on BraTS 2018 dataset and achieve promising results on test dataset with 14th place and Dice score of 0.720/0.878/0.785 for enhancing tumor, whole tumor and tumor core segmentation respectively.

Keywords: Segmentation · BraTS · UNet · Cascaded UNet · Multiple encoders

1 Introduction

Multimodal magnetic resonance imaging (MRI) is a powerful tool for studying human brain. Among it's different applications, it is mainly used for disease diagnosis and treatment planning. Accurate assessment of MRI results is critical throughout all these steps. Since MRI scans are the set of multiple three dimensional arrays, it's manual analysis and evaluation is a non-trivial procedure and requires time, attention and expertise. Lack of these resources can lead to unsatisfying results. Typically, these scans are analyzed by clinical experts using two

© Springer Nature Switzerland AG 2019
A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11384, pp. 189–198, 2019.
https://doi.org/10.1007/978-3-030-11726-9_17

dimensional cut and projection planes. It limits the amount of data taken into account for decision making, thus it adds bias to the resulting evaluation. On the other hand, accurate segmentation and 3D reconstruction is able to provide more insights on disease progression and help a therapist to plan the treatment better. However these methods are not widely used due to unreasonable amount of time needed for manual labeling.

Denoting the problem of automatic glioma segmentation Brain Tumor Segmentation (BraTS) challenge [1,11] was created and became an annual competition allowing participants to evaluate and compare their state of the art methods using unified framework. Participants are called to develop their algorithms and produce segmentation labels of the different glioma sub-regions: "enhancing tumor" (ET), "tumor core" (TC) and "whole tumor" (WT). The training data [2,3] consists of 210 high grade and 75 low grade glioma MRIs manually labeled by experts in the field. Testing data is split into two parts: **validation set** that can be used for evaluation throughout the challenge and **test set** for final evaluation. Performance of the methods is measured using Dice coefficient, Sensitivity, Specificity and Hausdorff distance.

Above-named challenge made a significant impact on the evolution of computational approaches for tumor segmentation. In the last few years, a variety of algorithms were proposed to solve this problem. Compared with other methods, convolutional neural networks have been showing the best state of the art performance for computer vision tasks in general and for biomedical image processing tasks in particular.

In this paper we present cascaded variant of the popular UNet network [6,12] that iteratively refines segmentation results of it's previous stages. We employ this approach for brain tumor segmentation task in the scope of BRATS 2018 challenge and evaluate it's performance. We also compare regular 3D UNet [6] with it's cascaded counterpart.

2 Method

In this study we propose neural networks based approach for brain tumor segmentation. Our method can be represented as a chain of multiple classifiers C_i of the same topology F refining segmentation output of previous iterations. Every classifier C_i shares the same topology but has it's own set of parameters W_i that is subject to optimization during training. Y_i - the result of the i-th step can be represented as $Y_i = F(X_i, Y_{i-1}, Y_{i-2}, W_i)$, where X_i is the i-th input.

Described approach is illustrated in Fig. 1. Each of the basic blocks C_i is a UNet network modified with respect to the task of glioma segmentation. Compared to the original UNet architecture described in [12] and extended for 3D case in [6], we employ multiple encoders separately handling input modalities and introduce the way to merge their output. In this paper we describe UNet modification with multiple encoders first. Then we propose ensembling strategy to efficiently merge segmentation results obtained on different scales.

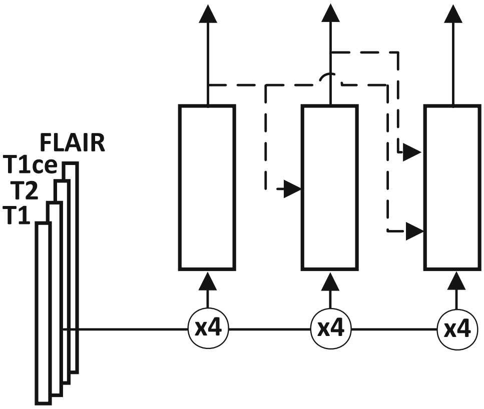

Fig. 1. Schematic representation of approach employed in this paper. T1, T2, T1ce, FLAIR stands for input MRI modalities. x4, x2 indicate downsampling factor for the network input. Dotted arrows indicate connections between networks C_i that are illustrated as basic blocks.

2.1 Multiple Encoders UNet

Traditional UNet architecture [12] extended for handling volumetric input [6] has two stages: encoder part where network learns feature representations on different scales and aggregates contextual information, and decoder part where network extracts information from observed context and previously learned features. Skip connections employed between corresponding encoder and decoder layers enable efficient training of the deep parts of the network and comparison of identically scaled features with different receptive fields.

This method allows to handle multimodal MRI input, however, it mixes and processes signals of different types identically. In contrast, we propose approach that learns feature representations for every modality separately and combines them at later stages. This is achieved by employing grouped convolutions in the encoder path with number of groups equals to the number of input modalities. Resulting features are calculated as a maximum of the feature maps produced by encoders. In order to preserve feature maps' sizes we employ point-wise convolution right after max operation. Similar to the original UNet, the number of filters is doubled with every downsampling operation and reduced by half with every upsampling operation, ReLU is used as activation function after every convolution layer. Described architecture is illustrated in Fig. 2.

The network is built of basic pre-activation residual blocks [7] that consist of two instance normalization layers, two relu activation layers and two convolutions with kernel size 3. This basic building block is illustrated in Fig. 3.

The motivation behind this architecture is to encourage model to extract features separately for every modality. In combination with feature maps merging strategy and channel-out augmentation it allows to build more robust model that can process data with one or more corrupted modalities (Fig. 4).

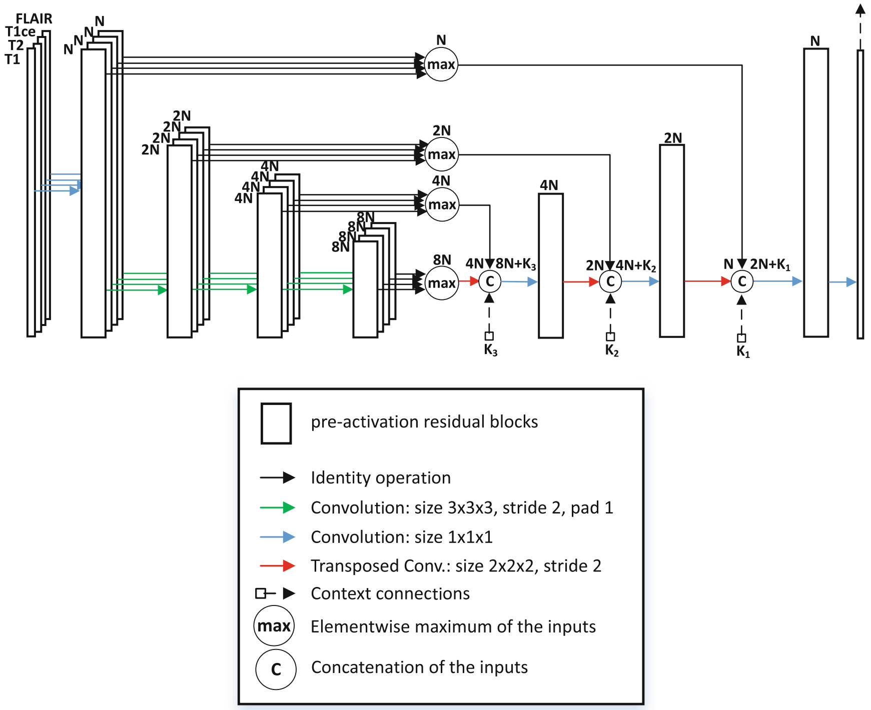
FLAIR
T1ce
T2
T1
N
N
N
N
2N
2N
2N
2N
4N
4N
4N
4N
8N
8N
8N
8N
N
max
2N
max
4N
max
8N
max
4N
8N+K3
C
K3
4N
2N
4N+K2
C
K2
2N
N
2N+K1
C
K1
N
pre-activation residual blocks
Identity operation
Convolution: size 3x3x3, stride 2, pad 1
Convolution: size 1x1x1
Transposed Conv.: size 2x2x2, stride 2
Context connections
Elementwise maximum of the inputs
Concatenation of the inputs

Fig. 2. Architecture of multiple encoders UNet. T1, T2, T1CE, FLAIR stand for input modalities. N is a base number of filters, K is a number of filters in context feature map obtained from lower scale models.

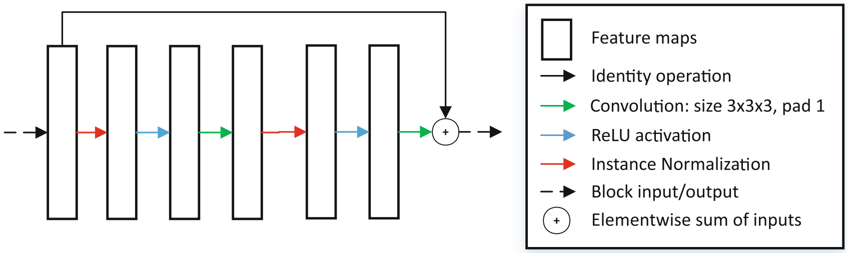

Fig. 3. Design of the residual block

Cascaded UNet. Proposed network is illustrated in Fig. 1 and consists of three basic blocks. Each block by itself is a modified UNet network with it's own loss function at the end. Every next block takes downsampled volume as an input and produces segmentation of the corresponding size. Similar to DeepMedic [10], this architecture simultaneously processes the input image at multiple scales and extracts scale-specific features. The feature map before the last convolutional layer in every block is concatenated to the corresponding feature map of higher-scale block. It enables the context information flow between networks with different scales.

In UNet architecture decoder output at each scale i depends on encoder output at the same scale (skip connections) and decoder output of the previous scale: $d_i^t = f(e_i^t, d_{i-1}^t)$, where d_i^t is decoder output, and e_i^t is encoder output at scale i, and t is the index of the network. Expanding the first convolution of f we get $d_i^t = g(W_{i,e}^t e_i^t + W_{i,d}^t d_{i-1}^t)$, where W are trainable parameters. Here we propose to incorporate context of the lower scale networks by concatenating corresponding network output y^t (see Fig. 2, illustrated as dotted arrows) so d_i^t becomes $d_i^t = g(W_{i,e}^t e_i^t + W_{i,d}^t d_{i-1}^t + W_{i,y}^t y^{t-i})$. This approach fuses multiple networks operating at different scales together and encourages model to iteratevily refine results of previous iterations.

The connections between networks are illustrated as dotted arrows in Fig. 1. Each basic UNet network produces two outputs: feature map (dotted arrows) and softmax operation over this feature map (straight arrows). The resulting probability tensor can be further used for ensembling, yet, we are interested in a final feature map. Since it has the most meaningful information about segmentation on the given scale, we want to propagate this feature map to higher resolution networks. To achieve the flow of the context between classifiers of different scale we propose to concatenate their output feature map to corresponding feature map of the higher scale network (see Fig. 2, illustrated as dotted arrows).

By employing following ensembling strategy we are building quite deep convolutional neural network. Compared to standard approach of doubling the number of feature channels after each pooling operator, out method takes less parameters and introduces bottlenecks between networks. Having same number of parameters, presented approach performs better than models with the same depth or the same number of parameters.

2.2 Data

In this paper we are focusing at brain tumor segmentation with deep neural networks. For training and evaluation purposes we are using BraTS 2018 [1–3,11] dataset. It contains clinically acquired preoperative multimodal MRI scans of glioblastoma and lower grade glioma obtained in different institutions with different protocols. These multimodal scans contain native T1, post-contrast T1-weighted, T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes, and co-registered to the same anatomical template, interpolated to the same resolution ($1mm^3$) and skull-stripped. These MRI scans were manually annotated by one to four raters, and approved by experienced radiologist. Segmentation labels describe different glioma sub-regions: “enhancing tumor” (ET), “tumor core” (TC) and “whole tumor” (WT). In total, dataset has 285 MRIs for training (210 high grade and 75 low grade glioma images), 67 validation and 192 testing MRIs.

2.3 Preprocessing and Data Augmentation

We have found data preprocessing employed in [8] to be especially effective. Like in [8], we perform z-score normalization on non-zero (brain) voxels. After that we are eliminating outliers and noise by clamping all values to the range from –5 to 5. At the final step we shift brain voxels to the range [0;10] and assign zeros to background.

For offline data augmentation we artificially increase number of samples by employing b-spline transformation to the original data. It has been done with ITK implementation [9].

During training we randomly flip input image along sagittal plane and "mute" input modalities with predefined probability. Without this augmentation the network was only considering one of the input modalities while making a prediction and not taking others into account. To deal with this issue we are randomly filling input channels with Gaussian noise. We introduce probability to apply this augmentation for every channel and set it to 0.1, so there is 34% chance to mute at least one out of four modalities. This also helps to aggregate information allover input data and to deal with noisy or corrupted input images like illustrated in the Fig. 4.

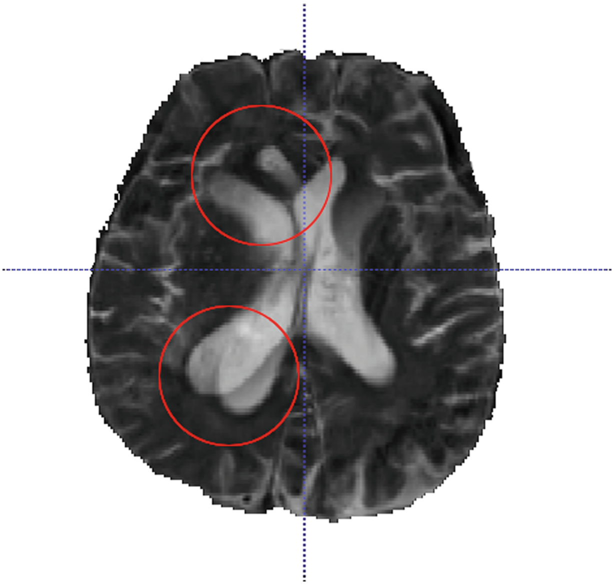

Fig. 4. Example of the registration artifacts found in the training dataset. This series contain one corrupted modality (shown) and three correct ones. Overlapping structures of the brain are marked with red circles. Visualization is done with ITK-SNAP [13]. (Color figure online)

2.4 Training

The training procedure is conducted on brain regions resampled to $128 \times 128 \times 128$ voxels. We are operating with downsampled data to preserve the context since we believe it plays important role for robust segmentation of multimodal MRI scans obtained from different institutions and scanners. We use Mean Dice loss $L_{mean\_dice}(g, p)$ where g is a ground truth, p is a model's prediction. We trained our network with stochastic gradient descent with initial learning rate of 0.1, exponential learning rate decay with rate 0.99 for every epoch, weight decay of 0.9 and minibatch size equal to 4 samples.

$$L_{mean\_dice}(g, p) = 1 - \frac{1}{|C|} \sum_{c \in C} \frac{\sum_{i_c} p_c^i g_c^i}{\sum_{i_c} p_c^i + g_c^i},$$

where C is a set of different classes.

This CNN was implemented in MXNet framework [5] and trained using four GTX 1080TI with batch size 4 to enable data parallelism. Training was performed for 500 epoches.

3 Results

In this section we report evaluation results obtained with online validation system provided by organizers. With intention to penalize model for relying on the one single modality we apply channel-out augmentation to the input data by randomly filling input modalities with Gaussian noise in addition to standard augmentations like mirroring and elastic transformations. Then we compare results obtained with this augmentation disabled (Table 1) and enabled (Table 2). The challenge validation data [2,3] contains 66 MRI scans obtained with different scanners and from different institutions. Results of evaluation on validation dataset are reported in Table 3; and on test dataset in Table 4.

Table 1. Evaluation of glioma segmentation without channel-out augmentation; Dice index is reported, WT stands for whole tumor, ET stands for enhancing tumor, TC stands for tumor core, ME UNet stands for Multiple Encoders UNet and C ME UNet stands for Cascaded Multiple Encoders UNet. Tested networks has the same number of parameters.

Method	WT	ET	TC
UNet	0.901	0.767	0.797
ME UNet	0.904	0.763	0.823
C ME UNet	0.906	0.772	0.836

Table 2. Evaluation of glioma segmentation with channel-out augmentation; Dice index is reported, WT stands for whole tumor, ET stands for enhancing tumor, TC stands for tumor core, ME UNet stands for Multiple Encoders UNet and C ME UNet stands for Cascaded Multiple Encoders UNet. Tested networks has the same number of parameters.

Method	WT	ET	TC
UNet	0.901	0.779	0.837
ME UNet	0.907	0.784	0.827
C ME UNet	0.908	0.784	0.844

Table 3. Performance of proposed method on BraTS 2018 validation data, Dice index is reported.

	WT	ET	TC
Mean	0.908	0.784	0.844
StdDev	0.065	0.237	0.161
Median	0.926	0.858	0.906
25quantile	0.900	0.805	0.791
75quantile	0.943	0.897	0.947

Table 4. Performance of proposed method on BraTS 2018 test data, Dice index is reported.

	WT	ET	TC
Mean	0.878	0.720	0.795
StdDev	0.119	0.278	0.251
Median	0.913	0.818	0.901
25quantile	0.870	0.711	0.804
75quantile	0.940	0.877	0.936

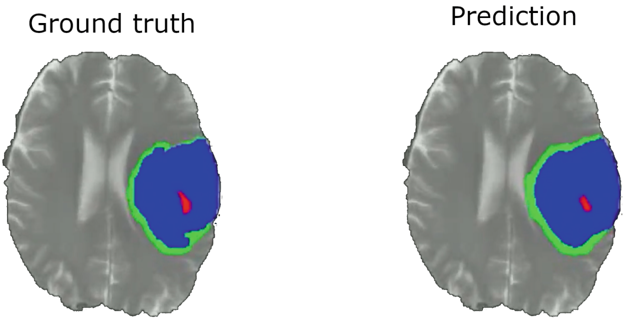

Fig. 5. Example of segmentation labels produces by proposed method in comparison with ground truth annotation.

4 Discussion and Conclusion

Analyzing the segmentation results provided by out model (Fig. 5) we noticed that it produces more smooth results compared to ground truth. According to BraTS 2018 challenge summarizing manuscript [4], out method took 14th place in final ranking. Analyzing the results we found out model to produce high number of inaccurate enhancing tumor segmentation labels (24th rank by DICE ET). This issue could be potentially overcame with learning ET, TC, WT labels instead of labels provided by annotation. However our model showed relatively high score for segmentation of Tumor Core (11th place by DICE TC) and Whole Tumor (10th place by DICE WT). Furthermore, it achieved ranks as high as 9th, 5th, 12th for segmentation of ET, TC, WT w.r.t. Hausdorff distance.

To sum it up, in this paper we presented automatic segmentation algorithm solving two main problem arising during brain tumor segmentation with multimodal scans: complex input consisting of multiple modalities and overconfidence of the classifier. Solving the problem of heterogeneous input we proposed to use multiple encoders, so that every individual input modality produces corresponding feature maps independently from others; and we introduced the way to merge encoded feature maps. Also we explored influence of channel-out augmentation on model's output quality and we showed that proposed architecture benefits from this aggressive augmentation. It encourages model to take into account whole input by implicitly penalizing classifiers that rely only on one single modality. As a result model becomes robust to the presence of noise and corrupted data that could be encountered in the training and validation datasets. Moreover we introduced the way to efficiently fuse multiple models operating on the different resolution that forms a cascade of classifiers. Every next classifiers takes results of previous ones and refines the segmentation for it's specific scale. It enables iterative result refinement with less parameters than in corresponding deep models. As a part of BraTS 2018 challenge [1,11] we implemented and evaluated our approach with online validation tools. As a result we achieved high mean score and notably high median score. The mean Dice score of 0.878/0.72/0.795 was reported on testing dataset for the Whole tumor, Enhancing tumor and Tumor core correspondingly

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5685212/, 28872634[pmid]
2. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
3. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF

4. Bakas, S., Reyes, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. CoRR abs/1811.02629 (2018). http://arxiv.org/abs/1811.02629
5. Chen, T., et al.: Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. CoRR abs/1512.01274 (2015). http://arxiv.org/abs/1512.01274
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, June 2016. https://doi.org/10.1109/CVPR.2016.90
8. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 287–297. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_25
9. Johnson, H.J., McCormick, M., Ibáñez, L., Consortium, T.I.S.: The ITK Software Guide. Kitware Inc, third edn. (2013, In press). http://www.itk.org/ItkSoftwareGuide.pdf
10. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. CoRR abs/1603.05959 (2016). http://arxiv.org/abs/1603.05959
11. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Yushkevich, P.A., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006)