Deep Learning Based Tangut Character Recognition

Guangwei Zhang
School of History and Civilization
Shaanxi Normal University
Xi'an, Shaanxi, China

Xiaomang Han
School of History and Civilization
Shaanxi Normal University
Xi'an, Shaanxi, China

Abstract—**The Tangut script, a logographic writing system, was used for writing the extinct Tangut language of the West Xia Dynasty. The huge amount of Tangut historical documents are being mainly recognized by Tangut experts manually, because the Tangut language has not been used since 16th century and it was impossible to recognized automatically in the past. With the help of deep learning, we build an end-to-end Tangut character recognition system to reduce the labor of Tangut experts. The high accuracy of a deep learning system for character recognition is essentially guaranteed by a large training dataset of well-labeled data. We construct a training dataset containing more than 100,000 labeled Tangut images, which is used for training a deep convolutional neural network (DCNN) to recognize Tangut characters. The Tangut images in the training dataset are from Tangut historical documents and they are labeled in a cluster-and-label way to reduce the human efforts. Based on the training dataset, the validation accuracy of the DCNN is more than 94% according to our experiments. We will release the training dataset for further study and construct an OCR system for transcribing Tangut historical documents automatically in the future.**

I. Introduction

The Tangut script, a logographic writing system, was used for writing the extinct Tangut language of the West Xia Dynasty (also known as the Xi Xia Empire). A considerable number of Tangut historical documents, including secular literature and Buddhist scriptures, have been published in recent years. These historical documents containing more than ten million Tangut characters are the foundation for studying this period of Chinese history, and the Tangut script is the key. Because the Tangut language has not been used since the 16th century, the Tangut historical documents can only be interpreted by the Tangut experts now. The Tangut scholars around the world have been working very hard to recognize and analyze the Tangut historical documents. Most of the research is done by hand, which is very time-consuming. New technologies are increasingly introduced into this field but still very limited. Nowadays, Tangut characters can be input into computers and 6,125 characters of the Tangut script were included in Unicode version 9.0 in June 2016 in the Tangut block. In the past, the coding of Tangut characters is not consistent among several Tangut input methods and the Tangut characters can only be recognized by Tangut experts, therefore it is difficult to set up a complete Tangut database efficiently. As we know OCR has been widely used in transcribing paper documents into computers, which greatly reduce the researchers' labor. The character recognition is the core function of the OCR and it determines the accuracy of the documents we get, therefore a Tangut character recognition system can help to implement an OCR system for transcribing Tangut historical documents.

It is necessary to utilize machine learning techniques to automatically recognize the Tangut script instead of by hand. The deep convolutional neural network (DCNN) achieves great success in Chinese character recognition. Because Tangut is similar to Chinese characters in appearance, the techniques used in Chinese handwritten character recognition can be applied and verified in recognizing Tangut characters. The Tangut script is also similar with Chinese characters in the vocabulary size. In this work, we focus on recognizing the first 1,000 high frequency Tangut characters containing in the Tangut secular documents. As we know, the labeled dataset is fundamental for a DCNN recognizing images, but there doesn't exist such a training dataset for Tangut character recognition, therefore, we collect more than 100,000 images from historical Tangut documents, each of which contains a Tangut character. We annotate these images and train a DCNN to recognize these high frequency Tangut characters. In this paper:

- We build the training dataset with a cluster-and-label method which reduces the labor of Tangut experts in great degree. We first cluster the collected Tangut images into groups. After discarding the wrongly clustered images in each group, the Tangut experts give each group a correct label;
- We train a DCNN with the training dataset to recognize high frequency Tangut characters on a GPU cluster. The recognition accuracy is more than 94% according to our experiments.

This paper is organized as follows: section II briefly introduces the DCNN for character recognition; section III describes the cluster-and-label method and the procedure for labeling the Tangut images sliced from scanned historical Tangut documents; section IV presents the architecture of the DCNN for recognizing Tangut characters; section V shows the experiment and results of the Tangut character recognition system; section VI reviews the related work.

II. DCNN for Character Recognition

A DCNN for character recognition is usually composed of many convolution and pooling layers after the input layer and a fully connected layer before the output layer. It can model the complex non-linear relationship between input and

978-1-5386-1107-4/17/$31.00 ©2017 IEEE

output from the self-extracted features from raw data without hand-engineered features, such that it makes an end-to-end character recognition system possible [1] and the performance is even better than human beings. The DCNN has been successfully used for recognizing many kinds of non-latin handwritten characters, for example, Chinese [2][3], Thai [4] and Nepali [5], etc.

The training of neural networks is generally time-consuming because there are a large number of parameters to optimize. The convolution operations, shared weights in convolutional layers and the pooling layers in the DCNN largely reduce the parameters compared with regular neural networks. However, it requires a lot of computation to train a DCNN especially for recognizing a language with a large vocabulary such as Chinese and Tangut. GPUs can largely reduce the training time of a DCNN because of its massive parallelism and high memory bandwidth, therefore most deep learning applications are commonly trained on GPUs. We train the Tangut recognition DCNN on a server equipped with a GTX 1080Ti GPU. The experiment is described in section V.

Besides GPUs, a training dataset of accurately labeled images is also important to the success of the DCNN in character recognition. For example, NIST dataset for English recognition, MNIST for handwritten digit recognition, CASIA Chinese handwritten dataset for Chinese handwritten character recognition [6], they all significantly benefit the research of recognizing the corresponding language. Therefore, a training dataset is indispensable for recognizing Tangut characters with a DCNN. However, to the best of our knowledge we are the first to try training a DCNN to recognize Tangut characters, so that we have to construct a complete training dataset with correctly labeled Tangut images by ourselves. It is well known that it is the most time-consuming and expensive step to label images for a deep learning system. We construct the training dataset for Tangut characters as stated in section III.

The DCNN is the core of the Tangut recognition system. Its hyper parameters influence the recognition accuracy and the execution time, including the number of hidden layers, the convolution kernel size, the pooling function and pooling size, learning rate, number of training epochs, etc. We describe the architecture and the tuning of hyper parameters of the DCNN for recognizing Tangut characters in section IV.

III. Labeling the Tangut Images

The labeled data is the foundation of a deep learning system. It is well known that the public datasets with large amounts of well-labeled images such as ImageNet help DCNN make tremendous success on many challenging computer vision tasks [7]. Data labeling is usually time-consuming, error-prone and expensive. Because there does not exist a labeled Tangut character dataset, we have to construct a training dataset in order to recognize Tangut characters by a DCNN. There are more than 6,000 different Tangut characters in total that have been recognized from historical Tangut documents by Tangut experts. We focus on recognizing the first 1,000 high frequency Tangut characters in this paper, and we are going to recognize all the Tangut characters in near future. To recognize all these characters, we have to provide each character with hundreds of characters written in different styles. All the Tangut images are sliced from scanned historical Tangut documents. A part of a page from the Mahāratnakūṭa Sūtra written in Tangut is shown in figure 1. An example of different styles of a Tangut character we collect are shown in figure 2(It means ”that”). We need to label about one million characters to recognize all the 6,000+ Tangut characters, which is difficult if not impossible to achieve manually. At the very first, we labeled the Tangut images by hand, as the early Tangut researchers conducted their research. It was so time-consuming and error-prone that we designed a mixed method to accelerate the labeling process.

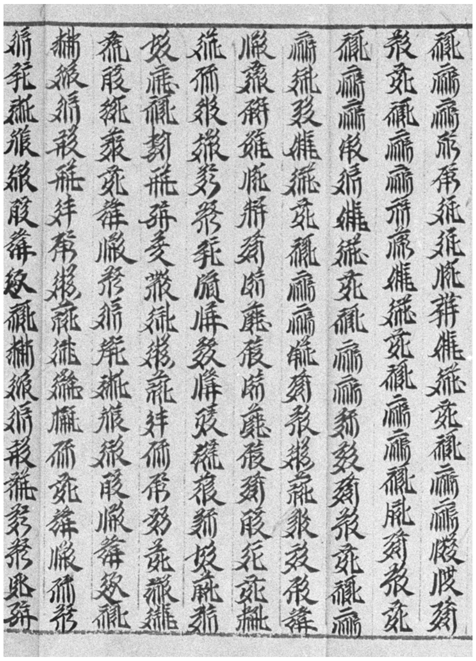

Fig. 1. A part of a page from the Mahāratnakūṭa Sūtra written in Tangut

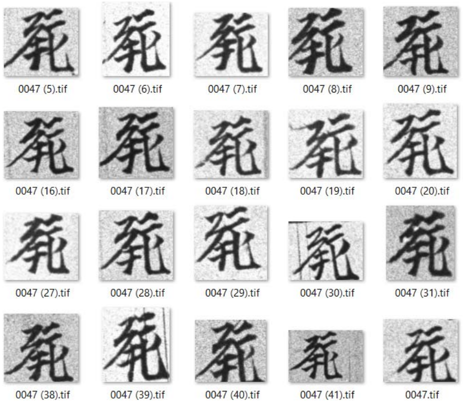

Fig. 2. An example of different styles of a Tangut Character

Semi-supervised methods have been used in data labeling, which can significantly reduce the human efforts. We label the Tangut images in a cluster-and-label way using K-means and Principal Component Analysis (PCA). The clustering accuracy is not accurate enough (nearly 60% according to our experiments), however, the roughly grouped images greatly reduce the labor of labeling. We refine the Tangut groups by discarding the wrongly grouped images, after which the Tangut experts label the Tangut characters by group and all the labeled characters are added into the training dataset. A clustering result is shown in figure 3.

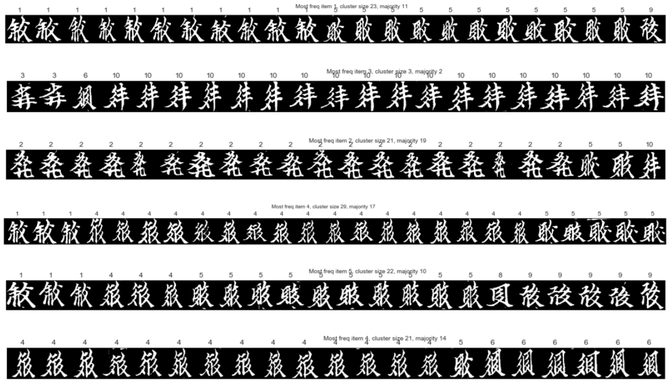
Most freq item 1, cluster size 23, majority 11
Most freq item 3, cluster size 3, majority 2
Most freq item 2, cluster size 21, majority 19
Most freq item 4, cluster size 29, majority 17
Most freq item 5, cluster size 22, majority 10
Most freq item 4, cluster size 21, majority 14

Fig. 3. Clustering Tangut images

The slicing and labeling steps are done in a pipeline pattern as follows: After 2 to 3 pages are sliced, we can get about 1,000 Tangut images. We use the cluster-and-label method to label the images. The labeled images are added to the training dataset the discarded images are added to the new sliced images for further clustering and labeling. Therefore, we can set a relatively small k for K-means in the cluster-and-label process, because there are limited individual Tangut characters in a 2 to 3 pages. We found $k = [10, 20]$ works for our project in most cases. We have labeled about 100,000 Tangut images with the cluster-and-label method. The whole labeled dataset is divided into training and testing dataset, which is used for training and validating the DCNN respectively.

IV. DCNN for Tangut Character Recognition

We build an end-to-end Tangut recognition system based on a DCNN that learns features itself from the training dataset we build. The workflow of the construction of the DCNN is as follows: (1) Image preprocessing; (2) Training the DCNN; (3) Tuning the hyper parameters; (4) Inferencing with the trained DCNN.

A. Tangut Images Preprocessing

The Tangut images sliced from scanned historical Tangut documents are of different sizes, as shown in figure 2, while the DCNN requires the same size of input images. Therefore, all the images are denoised, binarized and resized to 128×128 after padding the shorter side with 0. After that, all the labeled images are fed into the DCNN.

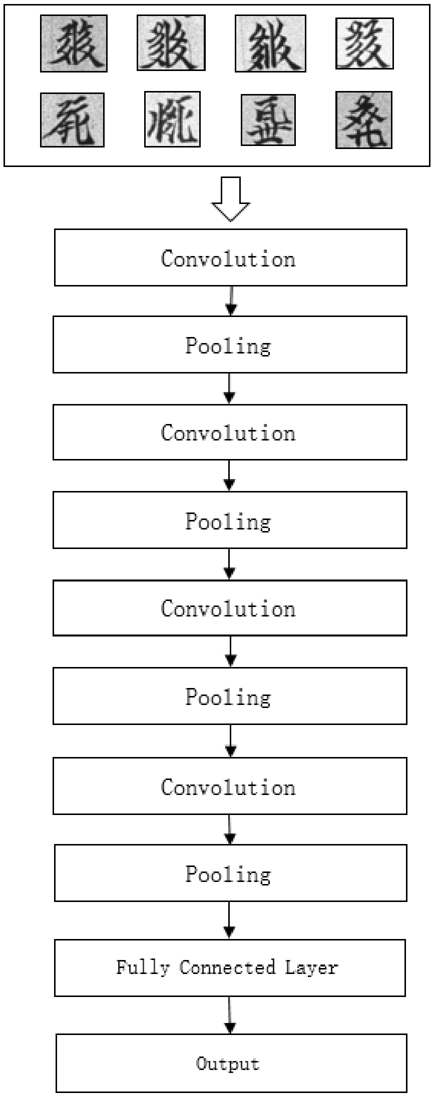

Fig. 4. The architecture of the DCNN

B. The Architecture of the DCNN

The DCNN transforms the original image through each layer in the architecture to generate a class score. The DCNN for Tangut character recognition contains three types of layers: the convolutional layers, the pooling layers and fully connected layers. The architecture of our DCNN is shown in figure 4. The convolutional layers are responsible for extracting features of the input data, which are the core of the DCNN conducting most of the computation. The lower convolutional layers in figure 4 get more abstract features that can be used for character classification. After each convolutional layer, we insert a RELU activation layer. Next, a max pooling layer is inserted to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control the overfitting problem.

C. Tuning the Hyper Parameters

It is critical to find the optimized hyper parameters for the accuracy and execution time of deep learning. We use an empirical tuning method to tune the hyper parameters. At first, the parameters are set with some commonly used values in some famous character recognition system such as digital recognition using MNIST. We train the DCNN every time when we finish labeling about 1,000 Tangut characters with previously used hyper parameters. If the accuracy is down, we will change some parameters until the accuracy is good enough. That means the hyper parameters are tuned many times with the training dataset growing. In our GPU server, the optimized DCNN for Tangut character recognition includes 4 convolutional layers, 4 RELU activation layers and 4 pooling layers. The pooling size is 2×2. The dropout rate is 0.5. The size of the convolution kernel is 8×8. The epoch number is 30.

As we stated in section III, the training dataset is generated gradually in the cluster-and-label procedure. Therefore, the DCNN is trained iteratively, when the training dataset size is small we tune the hyper parameters empirically. When we adding new training data, the hyper parameters are validated. During the construction of the training dataset and continuous validation, we found the hyper parameters not changing dramatically.

The DCNN is also validated with different ratio between training and testing dataset. Different ratio of training and testing dataset is tested such as 5:5, 6:4, 7:3, 8:2 and 9:1. After trying all the configurations, we find the validation accuracy of our DCNN is more than 90%. That means the hyper parameters adapt well in our project.

V. Experiment and Results

In this work, our aim is to classify the first 1,000 high frequency Tangut characters, so that the output of the DCNN has 1,000 categories. It means the DCNN will find a mapping from more than 100,000 Tangut images to the 1,000 categories after the training procedure.

We construct the Tangut recognition system using Keras with Tensorflow as backend. It is trained on a server quiped with a GTX 1080Ti GPU. We conduct 30 epochs to train the DCNN. The training process needs about 6 hours. The training accuracy is up to 98% as shown in figure 5 and the validation accuracy is more than 94% as shown in figure 6. The experiment results show that the recognition accuracy is good enough for practical usage in an OCR system.

After the DCNN is trained, we have tried using it to inference on new sliced Tangut images. Because some new Tangut characters are contained, we find the DCNN we trained can classify the recognized Tangut at a similar accuracy with our validation step. And another interesting thing is that though the characters not belonging to the labeled dataset when the DCNN is trained are classified wrongly, the same characters as well as some similar characters are classified into the same category. We are using this method for labeling the rest Tangut charcters, because we find it is an efficient method for collecting unlabeled data to construct a training dataset gradually.

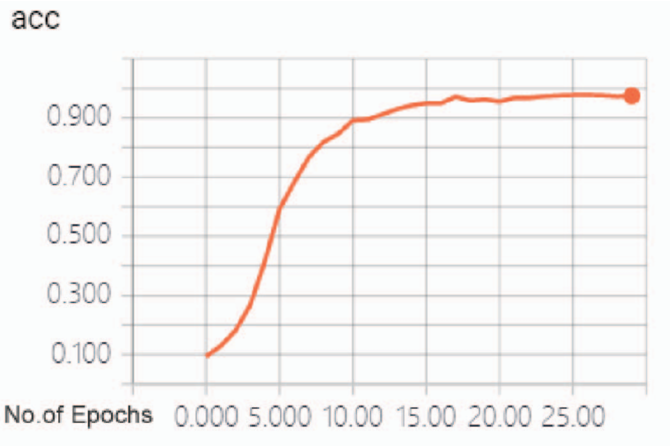

Fig. 5. Training accuracy

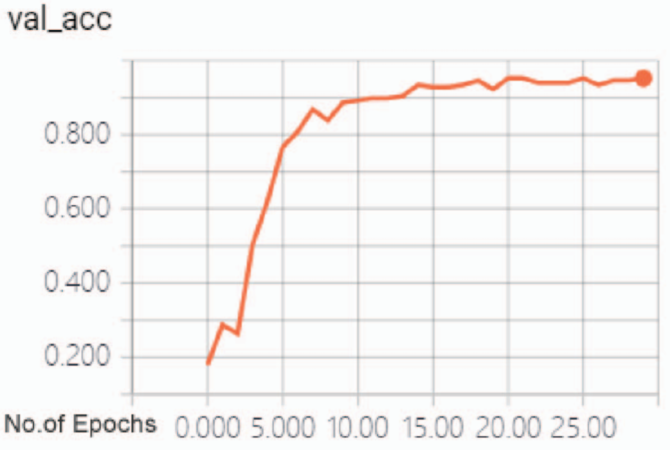

Fig. 6. Validation accuracy

VI. Related Work

There are huge amounts of historical archives in different languages around the world that need to be transcribed and indexed for better usage. In the past, it was hard to recognize automatically. Because deep learning has achieved great success in character recognition, deep learning based automatic transcription of historical archives has been studied [8][9].

Deep convolutional neural networks can learn features directly from images for classification, and performs the recognition task even better than human beings. Different kinds of DCNNs have been successfully used in many image recognition tasks, of which character recognition is an important area. LeCun's work in 1990 was successfully improving the recognition accuracy on the MNIST dataset by using a DCNN [10]. It is reported in ICDAR 2012 the testing error rate of a DCNN on MNIST was down to 0.23%. DCNNs have also been successfully used to recognize printed and handwritten characters of different languages [1][4][5]. Though the tasks of recognizing characters for different languages vary widely, the architecture and the training procedure of the DCNNs are similar. These character recognition tasks are much more difficult than recognizing handwritten digits, because they usually have a much larger vocabulary and the characters are much more complex. Chinese is thought to be one of the most difficult languages in the world. It was believed difficult if not impossible to recognize Chinese characters by machine especially for the handwritten Chinese characters. Researchers had paid huge efforts to recognize Chinese characters but the recognition accuracy had been low until deep learning was employed recently. GoogLeNet is used to recognize handwritten Chinese characters in [3] and their recognition accuracy is 96.74%. Tangut is similar to Chinese in both appearance and the vocabulary size. Machine learning based Tangut character recognition has been studied in recent years, but most of them use traditional methods and cannot recognize handwritten Tangut characters well[11]. Therefore, we build a GoogLeNet-based DCNN to recognize Tangut characters.

The success of deep learning is largely attributed to the large labeled dataset. A DCNN requires a large well-labeled character images to get high accuracy for recognizing the handwritten characters of a language. It is much easier to get Chinese characters and label them but the Tangut characters we can collect are limited, because Chinese are still widely used today, but Tangut has not been used for centuries. That is one reason why our Tangut recognition accuracy is not as high as Chinese recognition. Because labeling data is usually time-consuming and expensive, some unsupervised or semi-supervised machine learning methods are used to label data, such as the work [8][9][12]where clustering of different representations and active learning methods are used to label the data. In our experiments, we found the partially trained DCNN could be used to cluster unlabeled Tangut images, with which we are improving the cluster-and-label method.

VII. Conclusion and Future Work

In this paper, we develop a deep learning based Tangut character recognition system, which is trained for recognizing the first 1,000 high frequency Tangut characters based on the Tangut training dataset. The Tangut images in the training dataset are sliced from scanned historical Tangut documents, and they are labeled in a cluster-and-label way, which greatly reduce the human efforts. The validation accuracy of the Tangut character recognition system is more than 94% according to our experiments. The training dataset needs to be extended to help the DCNN recognize all the Tangut characters in the future. We are going to improve the image clustering work to accelerate the labeling process. We will release the training dataset that could be used for studying Tangut script recognition. Based on the Tangut character recognition system, a Tangut OCR system could be developed to help Tangut scholars easily transcribe the historical documents into computers and construct the Tangut database.

Acknowledgment

The authors would like to thank the anonymous reviewers. This work is supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.17YJCZH239).

References

[1] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 3304–3308.

[2] C. Cheng, X. Y. Zhang, X. H. Shao, and X. D. Zhou, "Handwritten chinese character recognition by joint classification and similarity ranking," in *International Conference on Frontiers in Handwriting Recognition*, 2017, pp. 507–511.

[3] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten chinese character recognition using GoogLeNet and directional feature maps," in *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, ser. ICDAR '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 846–850.

[4] T. Kobchaisawat and T. H. Chalidabhongse, "Thai text localization in natural scene images using convolutional neural network," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–7.

[5] S. Acharya, A. K. Pant, and P. K. Gyawali, "Deep learning based large scale handwritten devanagari character recognition," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Dec 2015, pp. 1–6.

[6] C. L. Liu, F. Yin, D. H. Wang, and Q. F. Wang, "Casia online and offline chinese handwriting databases," in *2011 International Conference on Document Analysis and Recognition*, Sept 2011, pp. 37–41.

[7] X. Wang, L. Lu, H. C. Shin, L. Kim, M. Bagheri, I. Nogues, J. Yao, and R. M. Summers, "Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition," pp. 998–1007, March 2017.

[8] J. Richarz, S. Vajda, R. Grzeszick, and G. A. Fink, "Semi-supervised learning for character recognition in historical archive documents," *Pattern Recogn.*, vol. 47, no. 3, pp. 1011–1020, Mar. 2014.

[9] M. Villegas, A. H. Toselli, V. Romero, and E. Vidal, "Exploiting existing modern transcripts for historical handwritten text recognition," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 66–71.

[10] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson, "Advances in neural information processing systems 2," D. S. Touretzky, Ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Handwritten Digit Recognition with a Back-propagation Network, pp. 396–404.

[11] C. Liu, "On tangut historical documents recognition," *Physics Procedia*, vol. 33, no. Supplement C, pp. 1212 – 1216, 2012, 2012 International Conference on Medical Physics and Biomedical Engineering (ICMPBE2012).

[12] J. Richarz, S. Vajda, and G. A. Fink, "Annotating handwritten characters with minimal human involvement in a semi-supervised learning strategy," in *2012 International Conference on Frontiers in Handwriting Recognition*, Sept 2012, pp. 23–28.