Combining Textual and Visual Information for Typed and Handwritten Text Separation in Legal Documents

Alessandro TORRISI, Robert BEVAN, Katie ATKINSON, Danushka BOLLEGALA and Frans COENEN

Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
e-mail: {alessandro.torrisi, robert.bevan, k.m.atkinson, danushka, coenen}@liverpool.ac.uk

Abstract. A paginated legal bundle is an indexed version of all the evidence documents considered relevant to a court case. The pagination process requires all documents to be analysed by an expert and sorted accordingly. This is a time consuming and expensive task. Automated pagination is complicated by the fact that the constituent documents can contain both typed and handwritten texts. A successful auto-pagination system must recognise the different text types, and treat them accordingly. In this paper we compare methods for determining the type of text data contained within paginated bundle pages. Specifically, we classify pages as containing typed data only, handwritten data only, or a mixture of the two. For this purpose, we compare text classification methods, image classification methods, and ensemble methods using both textual and visual information. We find the text and image based approaches provide complimentary information, and that combining the two produces a powerful document classifier.

Keywords. Pagination of Legal Bundles, Image Classification, Text Classification

1. Introduction

Legal document pagination [1] is an important process that is typically performed prior to a court hearing. The purpose of pagination is to produce an indexed court bundle containing all of the information and evidence related to a case. Processing legal documents in this way improves an advocate's ability to present a case during a hearing. During pagination, a domain expert must filter large volumes of information, meticulously sorting documents according to subject and often chronology. For example, in the medical negligence domain [2], medical records represent an important source of evidence and can easily contain hundreds or thousands of pages; during pagination any sections that are relevant to the medical negligence case need to be extracted from this vast amount of information. Pagination is further complicated in this instance due to the often non-contiguous distribution of evidence contained within medical records. In addition to evidence of any negligent acts, a patient's medical history may be relevant to the case, as well as any negative outcomes they experienced as a result of the negligence, which may occur months or years after the initial negligent act.

Regardless of the complexity of the case under examination, the pagination of medical records is generally a time-consuming and expensive task. The expense involved in

pagination calls for the development of automated methods to help speed up the process. Legal documents typically contain both typed and handwritten texts. Identification of typed and handwritten texts is a necessary precursor to building an effective automatic pagination tool. For example, if a system can determine that a page contains only typed data, the page can be analysed using classical Optical Character Recognition (OCR) approaches, whereas if the page contains handwritten data, handwriting recognition will need to be applied. A further advantage of identifying the type of text contained within a page is that it can help with the task of page categorisation. For example, in the medical negligence domain, consultation notes and records of correspondence between parties often contain typed text only, whereas pages containing handwritten text only, and those containing a mixture of typed and handwritten data, often correspond to doctors' notes, consent forms and/or laboratory examination reports respectively.

In this paper we compare different methods of classifying paginated bundle pages into three categories: (a) typed text only pages (**typed**), (b) handwritten text only pages (**handwritten**), and (c) pages containing both types (**mixed**). We experimented with two different approaches. For the first approach the problem was treated as a text classification task. For the second approach it was treated as an image classification task. We compare different methods of visual feature extraction including construction of visual keypoints using classical feature extraction methods, and feature extraction using pre-trained convolutional neural networks (CNNs) [3]. We also experimented with fine-tuning a CNN pre-trained on a large dataset [4]. Finally we combined the separate methods in an ensemble, and observed that the two different approaches provided complimentary information, with a best accuracy of over 95%.

2. Proposed Approach

Law firms typically receive medical records in hard copy. These are then scanned and converted into an electronic format using OCR software. Modern OCR software performs very well, but typically typed text is handled better than handwritten text. OCR applied to handwritten text can be error prone. This discrepancy in the quality of texts produced when applying OCR to the different text types motivated our text classification approach. We treat the problem as a standard text classification task. Specifically, we built a dictionary of unigrams, bigrams, and trigrams using a distinct set of medical records from those used in our experiments. Documents were converted into a machine readable format using a Bag of Words (BoW) model. A logistic regression classifier was trained to classify the documents, optimized using a grid-search approach.

Humans are easily able to distinguish typed and handwritten text by eye. The purpose of the image classification approach was to exploit visual features. Additional motivation was the fact that OCR software will sometimes fail to identify any text in pages that do in fact contain text, limiting the effectiveness of the text classification approach. Image classification relies on the use of visual words related to small parts of a page (converted to an image) which carry information related to features such as colour, shape or texture. In the Computer Vision community, a number of local feature operators have been presented [5]. After the advent of the well known Scale Invariant Feature Transform (SIFT) [6], different alternatives were proposed which satisfy a more efficient and effective calculation. Two examples are ORB [7] and BRISK [8] which are considered in this paper. Both algorithms detect keypoints inside an image, and assign a feature vector to each keypoint, which is of dimension 32 and 64 for ORB and BRISK, respectively.

To provide a robust estimate of the best feature operator, a balanced dataset of 15M keypoints was extracted from training data using both BRISK and ORB. The extracted keypoints were clustered using a standard k-means with the aim of grouping together all the keypoints related to similar objects. At the end of the clustering, a dictionary was obtained composed of the centroids of the resulting clusters (sets of visual words).

A second image classification was performed using Convolutional Neural Networks (CNN). CNNs produce state-of-the-art image classification performance when trained with very large datasets. In our application, the dataset was very small. Fortunately, it is possible to leverage the power of CNNs without a large training set through transfer learning [3]. There are two main approaches to transfer learning for image classification: fine-tuning and feature extraction combined with a linear classifier. In the first the CNN's output layer, and any fully-connected layers at the top of the network, are replaced with the number of units in the output layer equal to the number of classes in the problem; model training is resumed with the new dataset. Due to the hierarchical structure of the features extracted from the different network layers, typically, only a subset of the layer weights towards the top of the network are updated during training, as features extracted in the lower layers are less specialized and therefore more likely to be useful. In the case of CNNs for extracting features, typically the fully-connected layers at the top of the network are replaced with a single pooling layer and a softmax classifier. The resulting network is trained to classify the new dataset, with only the softmax layer weights updated. It is also possible to truncate the network at a lower level prior to adding the pooling and softmax layers, which can lead to superior classification performance due to the hierarchical feature structure.

Both feature extraction and fine-tuning approaches were considered. First, we trained linear classifiers using features extracted with the following pre-trained networks: Xception, ResNet152V2, InceptionV3, InceptionResNetV2, MobileNet, DenseNet201, and NASNetLarge [9, 10, 11, 12, 13, 14, 15]. In each instance the fully-connected layers at the top of the network were replaced with an average pooling layer and softmax layer containing three units corresponding to the three classes. In addition, we experimented with extracting features using different sub-architectures of the InceptionResNetV2 network (the network was truncated at different levels prior to feature extraction). Next, we fine-tuned the pre-trained MobileNet network [13], replacing the fully-connected layers at the top of the network with three dense layers with ReLu activation function and a softmax output layer with three units. MobileNet was selected for fine-tuning in the belief that it was the least likely to overfit due to its relatively low parameter count. Each network was optimized using Adam [16].

3. Evaluation Data

To evaluate the proposed approach, 50 pre-paginated medical bundles were used of the form that might be used in accident claims litigation. 30 bundles were randomly selected as the training data. 3000 different pages were extracted, 1000 for each category. The last 20 bundles were used as test data. A total of 1800 pages were extracted (600 pages for each category). Creation of ground truth information was conducted using two different domain experts. Selection of candidate samples was undertaken to include as many handwritten writing styles as possible. All the collected documents were in PDF format. Text was at first extracted and then pages were converted to images to perform image

classification. Consent was obtained from clients to use their medical data for this research. Non-anonymous sensitive information was included and this prohibits us from making this data publicly available.

4. Experimental results

For the evaluation of both text and image classifications, the metrics used were Precision, Recall and the F1 measure. Results of the CNN feature extraction experiments are shown in Figure 1. Each of the CNNs produced useful classification features: the worst performing classifier, trained using features extracted with DenseNet201, achieved a class-averaged F1 score of greater than 80%. The best performing classifier, trained using features extracted with InceptionResNetV2, achieved an F1 score of over 89%. We found that extracting features at an earlier stage of the InceptionResNetV2 network improved classification performance by ~2% (Figure 1). This may be because the features extracted at the top of the network are more specialized to the original training task.

Each network was optimized using Adam ($\eta = 0.001$; $\beta_1 = 0.9$; $\beta_2 = 0.999$). Prior to training, 20% of the training data was randomly selected for use as a validation set. Fine-tuning was performed for 50 epochs with early stopping according to validation loss. In the feature extraction setting, classifiers were trained for 500 epochs without early stopping. In both settings, model checkpoints were saved at epochs where the validation performance exceeded the previous best, and the best performing model was selected for use in the evaluation. Each experiment was repeated 5 times, and the best performing models in each trial were combined in an ensemble for the evaluation in order to minimize the effect of model initialisation.

Table 1 shows the evaluation of seven different classifiers, three of them consider an ensemble of textual and visual information. The best image classification (conducted considering a classical image classification approach) was achieved considering BRISK as feature operator and an Extra Tree Random Forest (ETRF) classification model. Seven different values of k in the range (50, 2000) were tested to find the optimal size of the code-word representing each page of a medical bundle. The achieved results are statistically comparable but a best F1 measure equal to 90.3% was registered considering BoVW vectors composed by 750 features (see second row in Table 1). Class probabilities obtained conducting a text classification improved the results of image classification in all the conducted experiments. F1 measures equal to 93.5% and 95.7% were achieved when text classification is combined with the image classification conducted through traditional approaches and using a fine-tuned CNN such a MobileNet, respectively. A fine-tuned MobileNet improved the class-averaged F1 score by 12% when compared with the classifier trained with features extracted using MobileNet, without any fine-tuning.

Use of visual information was useful in this case to resolve labels for pages not containing any text. 70% (28 out of 40) of empty pages were correctly classified using MobileNet. Another advantage of using a CNN instead of a classical image classification approach is related to time performance. It takes about a second to classify a document image considering classical approaches. It includes the extraction of keypoints, their conversion to a BoVW vector and classification. Using a fine-tuned CNN such as MobileNet drastically reduces the computational time, ensuring a classification of up to five images per second.

There are two main sources of error produced by the proposed approach. The first one is related to the straight lines which might be eventually included in pages of medical

CNN	Prec	Rec	F1
DenseNet201	84.9	82.1	82.3
ResNet152V2	85.5	84.3	84.4
MobileNet	85.3	84.5	84.6
Xception	86.4	85.7	85.8
InceptionV3	88.9	88.2	88.3
NASNetLarge	89.0	88.4	88.4
InceptionResNetV2	**89.7**	**89.3**	**89.4**

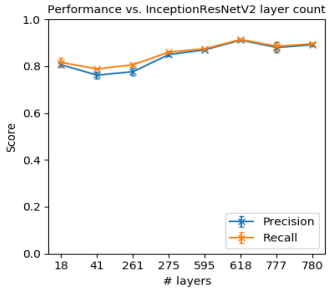

Figure 1. Left: Comparison of models trained using different CNNs as feature extractors. **Right:** Classifier performance for models trained using features extracted from various points of the InceptionResNetV2 network. All metrics were macro-averaged across classes.

Classifier	Prec	Rec	F1
Text Classification	86.6	86.5	86.5
ETRF	90.3	90.3	90.3
Text Classification + ETRF	93.6	93.5	93.5
MobileNet	94.7	94.7	94.7
Text Classification + MobileNet	**95.7**	**95.7**	**95.7**
InceptionResNetV2	91.4	91.1	91.2
Text Classification + InceptionResNetV2	94.5	94.4	94.4

Table 1. Evaluation of seven different decision systems.

records. As most of the pages containing straight lines are represented by medical forms, charts and tables, the current classifier is more prone to associate this information to pages belonging to the class “mixture”. A few other errors were discovered in pages not containing enough contrast between background and foreground pixels, making the keypoints extraction using BRISK more difficult. These errors suggest to us to conduct an image pre-processing step prior to classification in order to remove noise or any other artefact which can alter the classification verdict.

4.1. Discussion

The pagination process requires that a complete set of documents is at first allocated to analysts with expertise in a particular legal area. In the medical domain, pages of medical records are sorted and collated according to the instructions of a referring solicitor. Usually, this process requires the extraction of three main sub-bundles composed by correspondence, clinical information and General Practitioner (GP) records, respectively. The machine learning approach proposed in Torrisi et al. [1] to assist the pagination was originally implemented for such purposes but it considered only text information. This means that it becomes less feasible in the presence of a badly scanned document or when an OCR device does not provide enough accuracy. The approach proposed in this paper is intended to resolve such eventualities and it can be combined with the functionalities already implemented in [1] for providing further data categorisation. A further advantage of using the proposed approach is as a precursor step for developing bespoke OCR solutions for handwriting recognition, which is one of our ongoing steps.

5. Conclusion

In this paper an approach for separating typed from handwritten data was proposed for assisting the pagination process in litigation claims. Classification was conducted considering textual information, visual information and an unweighted combination of both types of data. Two different image classification approaches were tested: one considered feature extraction and classification using standard machine learning models; a second image classification was achieved considering the use of pre-trained neural networks. Best classification performance was conducted considering the combination of text classification with MobileNet, which resulted in a final F1 measure equal to 95.7% on test data composed of 1800 samples. Given the promising performance, our current target is increasing the size of the employed dataset, and also including documents from different sources so that a multi-context machine printed / handwritten text separator can be achieved. We are also investigating further alternatives to sort the documents according to subject. This is made possible through testing of the proposed application by analysts who operate in the medical negligence context.

References

[1] A. Torrisi, R. Bevan, K. Atkinson, D. Bollegala, and F. Coenen. Automated bundle pagination using machine learning. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, pages 244–248, 2019.

[2] R. Bevan, A. Torrisi, D. Bollegala, F. Coenen, and K. Atkinson. Extracting supporting evidence from medical negligence claim texts. In *Proceedings of the 4th International Workshop on Knowledge Discovery in Healthcare Data*, KDH '19, 2019.

[3] M. Hussain, J. Bird, and D. Faria. A study on cnn transfer learning for image classification. *CoRR*, 2018.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, 2012.

[5] S. Ahmed Khan Tareen and Z. Saleem. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–10, 2018.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. pages 2564–2571, 2011.

[8] S. leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. pages 2548–2555, 2011.

[9] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, 2015.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, 2015.

[12] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, 2016.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, 2017.

[14] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, 2016.

[15] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, 2017.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *the 3rd International Conference for Learning Representations, 2015*.