Fine-grained Classification of Identity Document Types with Only One Example

Marcel Simon, Erik Rodner, and Joachim Denzler

Friedrich Schiller University Jena, Germany
{marcel.simon, erik.rodner, joachim.denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de/

Abstract

In this paper, we tackle the task of recognizing types *of partly very similar identity documents using state-of-the-art visual recognition approaches. Given a scanned document, the goal is to identify the country of issue, the type of document, and its version. Whereas recognizing the individual parts of a document with known standardized layout can be done reliably, identifying the type of a document and therefore also its layout is a challenging problem due to the large variety of documents. In our paper, we develop and evaluate different techniques for this application including feature representations based on recent achievements with convolutional neural networks. On a dataset with 74 different classes and using only one training image per class, our best approach achieves a mean class-wise accuracy of* 97.7%.

1 Introduction

Optical character recognition (OCR) is one of the core ingredients of many industrial computer vision applications and there has been plenty of research in this area. Classifying single or small groups of letters can be considered as being solved [1, 2] and even though some visual CAPTCHAs still pose a challenge for robust detection algorithms, they are also difficult to recognize for humans [3, 4].

Despite all of this success, OCR still remains a challenge for complex documents that contain a mixture of different fonts, images, background artifacts, and words from various languages without a restricted vocabulary. Among other applications, this is the case for identity cards and travel documents of unknown origin. Figure 1 shows six examples of such documents. Especially the vast amount of different layouts, languages, and fonts make it hard to robustly read all information using plain OCR. In addition, every document contains a different set of information and it is difficult to determine whether all information is read.

Hence information about the particular layout and language is valuable. In this paper, we therefore consider the task of automated categorization of identity cards and travel documents. Given a scanned document, the goal is to identify the country of issue, the type of document, its version and whether it is the front or back side. The document types include identity cards, regular and child passports, visas, and driver's licenses. This information allows for retrieving the layout of the document from a database and use this information to efficiently localize and read the document's content.

However, due to the text being different on every document and the visual similarity of some documents,

Figure 1. Anonymized examples of documents that are distinguished by our system. Each image represents a different class with a unique text layout in the dataset. From top to bottom: Indian passport version 5 and 6, Belgian and Portuguese ID card, German ID card version 4 and German passport version 10. Notice how the pictures and texts vary greatly within the same class while the discriminative parts are subtle. Best viewed in color.

this task poses a great challenge. In addition, identity documents contain sensitive information which leads to only very few samples available for training. More specifically, our system should be able to work with only one training image per class. As we show in the experiments, simply applying a general-purpose OCR to the whole scanned document and estimating the class given the recognized text is not an option. The reason for this are the different versions of a document which often cannot be distinguished by the available text but only by the layout of it. The OCR approach even fails, if only the country of issue needs to be determined, because the font and language is unknown before classification.

The contribution of this work is the evaluation of a challenging classification task in which only one independent training example is available per class. Besides popular state-of-the-art classification techniques, we evaluate the performance of convolutional neural networks, a technique that achieved astonishing results in current recognition competitions [5]. In addition, we also evaluate task-specific features and discuss their possible contribution to an accurate classification.

In the following sections, we briefly revise state-of-the-art pipelines for document classification. This is followed by a description of our classification system

and experimental results in section 3 and 4, respectively. Finally, section 5 concludes the paper with a summary of important results.

2 Related Work

In terms of application, our task is related to document classification. In this section, we will briefly review state-of-the-art techniques of this field.

Given a set of scanned documents, the goal is to distinguish different types of page layouts. We focus on approaches that use visual features, as these are the most related ones. Shin *et al.* [6] use numerous hand-crafted and text-layout-specific features to classify different document types with decision trees and self-organizing maps. Their features are mostly limited to documents with a large text block and hence are not applicable to our task. Bagdanov *et al.* [7] represent the document using attributed relation graphs and first-order random graphs. They assume that the document can be easily divided into multiple text zones. There are many works with similar layout assumptions [8, 9, 10]. In our application, however, the robust detection of text zones is a difficult task in itself for the same reason the OCR fails. Kumar and Doermann [11] use a bag-of-words approach with SURF features combined in hierarchical histograms in order to visually compare documents. While they focus on black and white documents, we also incorporate color documents. In addition, they use bag-of-words which turns out to be a bad choice for the task we are interested in as shown in the experiments. Some works [12, 13] match SIFT or SURF key point detections between test and reference images in order to classify documents. While this approach achieves very good results, its explicit matching of descriptors is a computationally demanding step. Since there are thousands of potential identify and travel document types, this approach is not a feasible solution for our application. In contrast, we perform fast spatial pyramid matching which can be seen as an approximate and robust matching of descriptors. Usilin *et al.* [14] perform document detection given an uncropped image based on the Viola-Jones detection framework. Sarkar [15] also uses the Haar-like features of the Viola-Jones-Framework and performs a maximum-likelihood estimation during classification. In contrast to both works, our approach is simpler and faster. It allows to distinguish over twelve times more classes with only one training image each at the same level of accuracy.

3 Identity Document Classification

In this section, we present the different techniques used for classification. This includes the different features, the feature quantization technique bag-of-words, the integration of spatial information using spatial pyramid matching, and the classification method.

Features We evaluate four different features in our experiments. The first feature is Pyramid Histogram of Oriented Gradients (PHOG) [16]. This approach represents each pixel in the input image by the discretized orientation of its gradient. All orientations are aggregated into histograms over increasingly smaller parts of the image. The global feature vector is then built by concatenating all histograms. The second feature type is Histogram of Oriented Gradients (HOG). HOG also uses gradient orientations, but, in contrast to PHOG, these gradients are aggregated for each fixed-sized block in the image. Neighboring blocks are then used to normalize the histograms. We also combine HOG with the popular feature quantization technique bag-of-visual-words (BOW). Bag-of-words is a technique to quantize local features into more abstract histograms over visual words. In training, the local features of all images are collected and unsupervised k-means-clustering is applied. This calculates k cluster means in the feature space, which represent the codebook. For encoding, a histogram over this codebook is created by assigning each local feature to the cluster center that is closest. The third feature mainly captures color information and is called Colorname descriptor [17]. Each RGB value in the input image is mapped to a histogram over ten semantic colors. The mapping is learned using a generic classification dataset and provided by the authors of [17].

The fourth and currently very popular feature type is the intermediate output of a pre-trained convolutional neural network (CNN) [18]. As part of current deep learning techniques, CNNs transform the input image into the desired output using one jointly trained model. The transformation is done using common operations like convolution, local normalization, pooling, matrix multiplication and element-wise non-linearities. Due to space constraints, we skip a detailed explanation and refer the reader to [5]. In our case, it is not possible to train a CNN from scratch since there is not enough training data available. Recent publications in the area of generic object classification use a pre-trained model in such a situation instead [18]. In most cases, the model is trained for the classification of ImageNet [19] pictures. These publications show that the intermediate feature encoding of such a pre-trained CNN is as generic as popular hand-crafted descriptors like HOG or SIFT, for example. This allows for using the same CNN to extract useful features in other tasks and datasets as well.

Spatial information Dense local features like HOG or Colorname are aggregated into one global feature vector by calculating a spatial pyramid (SP) of features. First, the statistics for the whole image are calculated. Second, the image is divided into four equally sized parts, the statistics for each of them are calculated and concatenated to obtain a global feature vector. This step is recursively repeated by subdividing each part again. Such a pyramid encodes spatial information while still keeping the robustness of histogram-based features.

Feature fusion and classification We combine multiple features using the early-fusion strategy, i.e. different features are concatenated into one large feature vector. The features are calculated for each image and used to train a Support-Vector-Machine (SVM). The one-vs-all strategy for multiclass classification and a linear kernel is used in all experiments.

4 Experiments

Datasets We evaluate our approach on a dataset containing 74 categories of identity and travel document from 35 different countries. The document types in our dataset include identity cards, regular and child passports, visas, driver's licenses with the backside of

some documents as additional class. In total, there are 375 images each showing a unique document. The images are already cropped, which means they only contain the document of interest and no background. Due to the difficulty of getting unique examples of such documents, 39 classes contain only one image. However, these classes are still included in training and hence can be the output of a prediction. Figure 1 shows six images of different classes.

Setup In each run of all experiments, one image per class is chosen randomly for training. All the remaining images are used for testing. The performance is measured as the mean value of all the class-wise accuracies. This is necessary since the number of samples per class varies greatly from one to 58. As mentioned above, the classes with only one image are still used for training and consequently can be predicted. However, they do not contribute to the calculation of the mean class-wise accuracies as there are no testing images for these classes available. Each experiment is repeated 100 times and the results are averaged in order to obtain a reliable performance measurement.

We use the following parameters for the algorithms. HOG features are calculated using the variant of [20] with the default cell size of 8 pixels and 9 bins. The PHOG features are calculated using 40 bins and a pyramid depth of three. The CNN features are calculated using the framework and the pre-trained ImageNet model of [21]. We used the output of layer *conv5*, which yielded the best results. For bag-of-words, k-means is used to find 500 cluster centers. Hard voting is used to encode the features into histograms. The spatial pyramid of feature histograms is calculated for a depth of three.

Results The results for different features are shown in Table 1. Comparing individual features, HOG achieves the best performance with a mean class-wise accuracy of 92.7%, closely followed by PHOG and Colorname with 91.6% each. The combination of the best intensity feature HOG and the color feature Colorname help to further boost the performance to 97.7%. Visualizations suggest, that HOG features are well suited to capture background patterns and the position of the photo. In contrast, Colorname features are well suited to distinguish the general coloring of the document.

If the CNN features are used, a slightly lower mean class-wise accuracy of 96.5% is achieved. Nevertheless, it is interesting to see that features learned on ImageNet achieve such a good performance on a completely different dataset. In contrast to many generic datasets, the feature quantization technique bag-of-words does not improve the performance. Using bag-of-words even decreases the performance by 15.2% in the case of HOG features. The combination of all features does not help to improve the accuracy. This is most likely due to the feature dimension being too high for the small amount of training data.

The major contribution to the performance comes from the spatial pyramid. We compared different depths of the spatial pyramid when using the best performing combination of HOG and Colorname features. As can be seen in Figure 2, the performance is relatively low with 78.9% if no spatial information is encoded. Adding only one level of the pyramid improves

Table 1. Performance of different features and the influence of feature quantization. The performance is measured as mean class-wise accuracy. Abbreviations: HOG - histogram of oriented gradients, BOW - bag-of-words, SPk - spatial pyramid with k levels, PHOG - pyramid of histogram of oriented gradients, CNN - convolutional neural network activations.

Features	Mean class-wise accuracy
HOG+BOW+SP3	77.5% ± 4.0%
Colorname+SP3	91.6% ± 2.5%
PHOG	91.6% ± 1.7%
HOG+SP3	92.7% ± 0.9%
CNN	96.5% ± 1.9%
HOG+Colorname+SP3 +CNN+PHOG	96.7% ± 2.1%
HOG+Colorname+SP3	**97.7%** ±1.6%

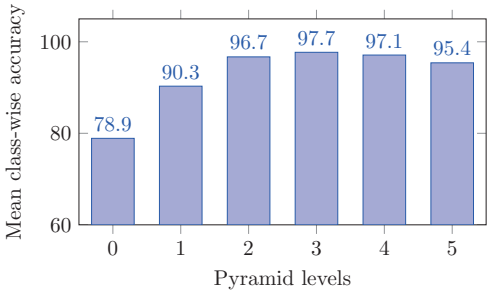

Figure 2. The influence of the spatial pyramid depth. Only one level already increases the performance by over 11% compared to a feature vector with no spatial information. The best performance is achieve with three levels.

the performance by over 11%. This result is intuitive, because all documents have a fixed alignment. Hence the top left part of the image corresponds to the top left part of the document and the same for the other parts. The best performance is achieved using three pyramid levels with a mean class-wise accuracy of 97.7%. On an Intel i7 processor with 3.4 GHz, this classification of a single image took less than 100 ms.

OCR baseline We also evaluated the reliability and usefulness of the task specific features. In order to get a baseline, we evaluated an OCR-based approach. We used tesseract [22], an open-source library developed by Google. In this experiment, we merged documents of the same country into one class. This is necessary because the OCR approach is not able to distinguish different document versions in most cases. We defined a set of discriminative words for each category. If one of these words appear on a test document, we assign it the corresponding class label. In contrast to the other experiments, all images are used for testing. Even in this simplified task, this approach achieved an average recognition rate of only 39.4%. The main reason for that is the unknown language and font on the document. We also evaluated other task specific features like whether there is a photo on the document

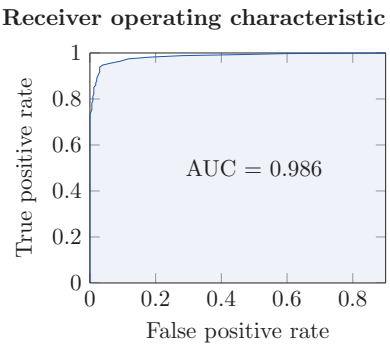

Figure 3. The ROC curve of our novelty detection system that is used to reject unknown documents.

and the aspect ratio of the document. However, none of these increased the recognition performance.

Novelty Detection In real world applications users might scan previously unseen documents. It is necessary to reject them in order to avoid unpredictable behavior of the document reading system. We build a simple yet effective novelty detection system by using L2-regularized logistic regression. During prediction, we calculate probability estimates and reject the input image if the probability is below a calculated threshold. We evaluated our system using a "leave-one-class-out"-strategy. From all but one class, we randomly select one image for training. All remaining images are used for testing. This is repeated for each class in the dataset. The performance is evaluated using ROC-AUC. As can be seen in Figure 3, our system is able to consistently distinguish between known and unknown documents with an AUC of 0.986.

5 Conclusions

This paper tackles the challenging task of automated identity and travel document classification. Since general-purpose OCR fails in most cases, document classification is an important pre-processing step for localizing the information in documents. For evaluation of different state-of-the-art methods, we used a dataset consisting of 375 unique documents categorized into 74 partly very similar classes. Using only one training image per class, the combination of HOG and Colorname features achieved a mean class-wise accuracy of 97.7%. Spatial information turns out to be a key ingredient in this dataset. Unknown documents are also detected and can be reliably recognized using logistic regression achieving a ROC-AUC of 0.986.

References

[1] Sermanet, P., Chintala, S., LeCun, Y.: Convolutional neural networks applied to house numbers digit classification. In: ICPR. (2012) 3288–3291

[2] Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint (2013) http://arxiv.org/abs/1312.6082.

[3] Bursztein, E., Bethard, S., Fabry, C., Mitchell, J.C., Jurafsky, D.: How good are humans at solving CAPTCHAs? A large scale evaluation. In: Symposium on Security and Privacy. (2010) 399–413

[4] Chellapilla, K., Larson, K., Simard, P.Y., Czerwinski, M.: Computers beat humans at single character recognition in reading based human interaction proofs. In: CEAS. (2005)

[5] Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS. Volume 25. (2012) 1097–1105

[6] Shin, C., Doermann, D., Rosenfeld, A.: Classification of document pages using structure-based features. IJDAR **3** (2001) 232–247

[7] Bagdanov, A.D., Worring, M.: Fine-grained document genre classification using first order random graphs. In: ICDAR. (2001) 79–83

[8] Heroux, P., Diana, S., Ribert, A., Trupin, P.: Classification method study for automatic form class identification. In: ICPR. Volume 1. (1998) 926–928

[9] Eglin, V., Bres, S.: Document page similarity based on layout visual saliency: application to query by example and document classification. In: ICDAR. (2003) 1208–1212

[10] Diligenti, M., Frasconi, P., Gori, M.: Hidden tree markov models for document image classification. PAMI **25** (2003) 519–523

[11] Kumar, J., Doermann, D.: Unsupervised classification of structurally similar document images. In: ICDAR. (2013) 1225–1229

[12] Chen, S., He, Y., Sun, J., Naoi, S.: Structured document classification by matching local salient features. In: ICPR. (2012) 653–656

[13] Infantino, I., Maniscalco, U., Stabile, D., Vella, F.: A fully visual based business document classification system. In: SAI. (2014) 339–344

[14] Usilin, S., Nikolaev, D., Postnikov, V., Schaefer, G.: Visual appearance based document image classification. In: ICIP. (2010) 2133–2136

[15] Sarkar, P.: Image classification: Classifying distributions of visual features. In: ICPR. Volume 2. (2006) 472–475

[16] Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: CIVR. (2007) 401–408

[17] Van De Weijer, J., Schmid, C.: Applying color names to image description. In: ICIP. Volume 3. (2007) III–493 – III–496

[18] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint (2013) http://arxiv.org/abs/1310.1531.

[19] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. CVPR **0** (2009) 248–255

[20] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2010) 1627–1645

[21] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint (2014) http://arxiv.org/abs/1408.5093.

[22] Smith, R.: An overview of the tesseract OCR engine. In: ICDAR. Volume 2. (2007) 629–633