Identity Documents Classification as an Image Classification Problem

Ronan Sicre[1](✉), Ahmad Montaser Awal[2], and Teddy Furon[1]

[1] Irisa/Inria, Rennes, France
{ronan.sicre,teddy.furon}@inria.fr
[2] AriadNext, Cesson-Sévigné, France
montaser.awal@ariadnext.com

Abstract. This paper studies the classification of identification documents, which is a critical issue in various security contexts. We address this challenge as an application of image classification, a problematic that received a large attention from the scientific community. Several methods are evaluated and we report results allowing a better understanding of the specificity of identification documents. We are especially interested in deep learning approaches, showing good transfer capabilities and high performances.

Keywords: Image forensic · Image classification · Document recognition

1 Introduction

Identity fraud is a major issue in today's societies with serious consequences. The threats vary from small frauds up to organized crimes and terrorist actions. The work presented in this paper is part of a research project IDFRAud[1] proposing a platform for identity documents verification. The first step classifies the query document according to its type and country of origin to prepare the verification of specific security checks, fake detection, document archiving etc. These later processes are out of scope of this paper.

For any supervised classification problem, the first task is to collect annotated data. In our application, these are specimens from various types of documents as well as emitting countries. Obtaining such data in large quantities is not always possible and we should therefore take into account this limitation. Some classes have more samples than others, giving unbalanced datasets. Moreover, query images vary from high quality scans to poor quality mobile phone photos with complex background, various orientations, occlusions, or flares, see Fig. 1.

There are two main approaches in the document classification literature. Methods based on the *layout* are mainly used when documents are composed of

[1] This work is achieved in the context of the IDFRAud project ANR-14-CE28-0012, co-financed by the french DGA: http://idfraud.fr/.

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 602–613, 2017.
https://doi.org/10.1007/978-3-319-68548-9_55

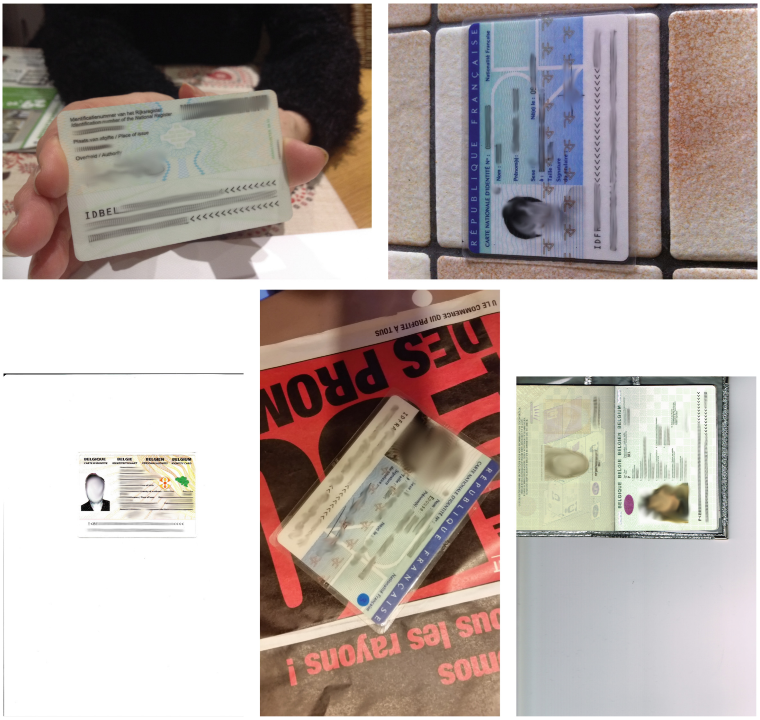
IDBEL
RÉPUBLIQUE FRANÇAISE
CARTE NATIONALE D'IDENTITÉ N°
BELGIQUE BELGIE BELGIEN BELGIUM

Fig. 1. Sample images from the databases

text blocks, figures, tables, etc. This is the case for journal articles, publications, books, or invoices. Documents are described by their spatial layout, *i.e.* the structure of text blocks, figures, and tables. Such descriptions are finally used to perform classification [2,16] or to compute similarities [8,24]. The second type of approaches is based on *text*. These methods build a description of the text content (extracted with an OCR in the case of scanned documents), such as bag of words or Word2Vec, which is given as input to classifiers [34]. More recently Recurrent Neural Networks (RNN) have been applied to classify documents [17].

Section 2 explains why we have discarded these two classical approaches to propose an alternative based on the visual content of the identity documents. Image recognition has a large spectrum of tasks with applications in search engines, interest object detection, or image categorization/classification, which has been extensively studied over the last decades. The availability of large and/or complex datasets as well as regular international challenges has spurred a large variety of image classification methods. We propose to apply these approaches to deal with identity document classification.

This choice is not obvious as there are few graphical elements in identity documents. Moreover, the portrait photo of the owner is uninformative for classification. However, the recent work on Convolutional Neural Networks (CNN) showed great generic visual descriptions, which are transferable to a large variety of image recognition tasks, such as fine-grain image classification. Thus, our paper studies a wide range of image classification method as well as the transfer capabilities of CNN to the specific task of identity documents classification.

2 Previous Works

The introduction presented document recognition through three main trends: layout-based, text-based, and visual-based methods. We now explain why we choose this last trend.

Identity documents contain textual and graphical information with a given layout. From such well structured documents, one could expect to base their classification on the layout. However, the layout is not always discriminant. Some classes share very similar structure: this is especially the case of different versions of passport or ID card emitted by the same country. Other methods are based on text transcription. Unfortunately, such methods are not adapted to our application due to the following difficulties: The document is not localized a priori in the query image and background information might disturb the OCR tasks, see Fig. 1. Indeed, text information is difficult to extract before knowing the type of the document and where it is localized in the image. Moreover, a large part of the text is specific to the owner of the document and not to the class. Therefore, we prefer to rely on the graphical content of the identity document and we turn towards image classification techniques in search for robustness and diversity.

Image classification has received a large attention from the scientific community, *e.g.* see the abundant literature related to the Pascal VOC [9] and ImageNet [7] challenges. A large part of the modern approaches follow the bag-of-word (BOW) approach [6], represented by a 3 step pipeline: (1) extraction of local image features, (2) encoding of local image descriptors and pooling of these encoded descriptors into a global image representation, (3) training and classification of global image descriptors for the purpose of object recognition. Local feature points, such as SIFT [21], are widely used as local features due to their description capabilities. Regarding the second step, image encoding, BOW were originally used to encode the feature point's distribution in a global image representation [12,16]. Fisher vectors and VLAD later showed improvement over the BOW [14,23]. Pooling has also witnessed many improvement: for instance, spatial and feature space pooling techniques have been widely investigated [18,32]. Finally, regarding the last step of the pipeline, discriminative classifiers such as linear Support Vector Machines (SVM) are widely accepted as the reference in terms of classification performance [4].

Recently, the deep CNN approaches have been successfully applied to large-scale image classification datasets, such as ImageNet [7,15], obtaining state-of-the-art results significantly above Fisher vectors or bag-of-words schemes. These networks have a much deeper structure than standard representations, including several convolutional layers followed by fully connected layers, resulting in a very large number of parameters that have to be learned from training data.

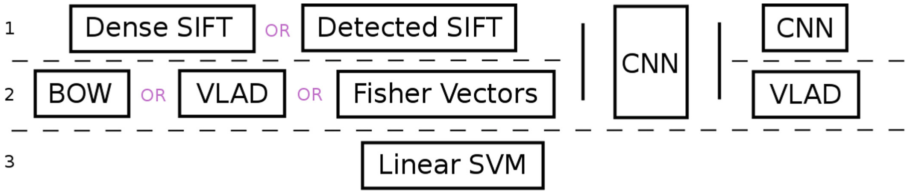

Fig. 2. Classification pipelines composed of (1) feature extraction on the first row, (2) feature encoding on the second row, and (3) classification on the final row.

By learning these networks parameters on large image datasets, a structured representation can be extracted at an intermediate to a high-level [22,35]. Furthermore, Deep CNN representation have been recently combined with VLAD [1,11] or Fisher vectors [5,19] encodings.

It is worth mentioning that other approaches have been proposed in Computer Vision with the aim to build mid-level description [29] or to learn a set of discriminative parts to model classes [10,25,28]. They are highly effective in similar fine-grain classification scenarios but are extremely costly.

3 A Plurality of Methods

To perform image classification, we first follow the BOW-based pipeline. SIFT keypoints are extracted in either a dense fashion or by interest point detection. Dense extraction tends to offer better performance in classification, while interest points are rotation invariant [26]. Then, these features are encoded with BOW, VLAD or Fisher vectors and are used to classify images with SVM.

Secondly, we study CNN based features, where intermediate transferable representations are computed from pre-trained networks. Descriptors are computed using various networks, layers, orientations, and scales. Finally, a VLAD aggregation of activation maps across orientations and scales is proposed. These image descriptors are similarly given as input to SVM to perform classification, see Fig. 2.

3.1 Bag-of-Words

Assume that the local description output vectors in $\mathbb{R}^d$. The Bag of visual Words aims at encoding local image descriptors based on a partition of the feature space $\mathbb{R}^d$ into regions. This partition is usually achieved by using the k-means algorithm on a training set of feature points. It yields a set $\mathcal{V}$, so called *visual vocabulary*, of k centroids $\{\mathbf{v}_i\}_{i=1}^k$, named *visual words*. The regions are the Voronoi cells of the centroids. This process is achieved offline and once for all.

The local descriptors of an image $\{\mathbf{x}_t\}_{t=1}^T$ are quantized onto the visual vocabulary $\mathcal{V}$:

$$\mathsf{NN}(\mathbf{x}_t) = \arg\min_{1\leq i\leq k} \|\mathbf{x}_t - \mathbf{v}_i\|. \tag{1}$$

The histogram of frequencies of these mappings becomes the global image description whose size is k.

3.2 Fisher Vectors

Fisher vectors also start from a visual vocabulary $\mathcal{V}$ but used as a Gaussian Mixture Model (GMM). The distribution of the local descriptors is assumed to be a mixture of k Gaussian $\mathcal{N}(\mathbf{v}_i, \mathsf{diag}(\boldsymbol{\sigma}_i^2))$ with weights $\{\omega_i\}$. Covariance matrices are assumed to be diagonal, variances vectors $\{\boldsymbol{\sigma}_i^2\}$ and weights $\{\omega_i\}$ are learned from the training set as well.

Fisher vectors considers the log-likelihood of the local descriptors of the image $\{\mathbf{x}_t\}_{t=1}^T$ w.r.t. to this GMM. They are composed of two gradient calculations of this quantity per Gaussian distribution: The gradient G_μ^X w.r.t. $\mathbf{v}_i$ and the gradient G_σ^X w.r.t. to the variance vector $\boldsymbol{\sigma}_i^2$:

$$G_{\mu,i}^X = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^T \gamma_t(i) \mathsf{diag}(\boldsymbol{\sigma}_i)^{-1}(\mathbf{x}_t - \mathbf{v}_i), \quad (2)$$

$$G_{\sigma,i}^X = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^T \gamma_t(i) [\mathsf{diag}(\boldsymbol{\sigma}_i^2)^{-1}(\mathbf{x}_t - \mathbf{v}_i)^2 - \mathbf{1}_d], \quad (3)$$

where $\gamma_t(i)$ represents the soft assignment term, *i.e.* the probability that descriptor $\mathbf{x}_t$ derives from the i-th Gaussian distribution [23], and $\mathbf{a}^2$ denotes the vector whose components are the square of the components of $\mathbf{a}$. The concatenation of these gradients results in a global descriptor of $2kd$ components.

3.3 VLAD

VLAD is similar to Fisher vectors [14] aggregating only the difference between the local descriptors and hard-assigned cluster from the visual vocabulary:

$$\mathbf{d}_i = \sum_{\mathbf{x}_t : \mathsf{NN}(\mathbf{x}_t) = i} \mathbf{x}_t - \mathbf{v}_i. \quad (4)$$

The global descriptor $(\mathbf{d}_1^\top, \ldots, \mathbf{d}_k^\top)^\top$ has a size of dk. A power law, l_2 normalization, and/or PCA reduction are usually performed on Fisher and VLAD [23].

3.4 Convolutional Neural Networks

Deep Convolutional Neural Network [15] are composed of convolutional layers followed by fully connected ones with normalization and/or pooling performed in between layers. There is a large variety of network architectures [30,31], but a usual choice is 5 convolutional layers followed by 3 fully connected layers. The layers parameters are learned from training data.

The works [22,33] showed that extracting intermediate layer produces mid-level generic representations, which can be used for various recognition tasks and a wide range of data [27]. In our case, we use a fast network and a very deep network, both trained on ImageNet ILSVRC data. The fast network from [3] is similar to [15], while the deep network stacks more convolutional layers (19 layers in total) with smaller convolutional filters [30].

Following previous works [22,28,33], image representations are computed by either taking the output of the fully connected intermediate layers or by performing pooling on the output of the last convolutional layer [33].

Unfortunately rotation invariance can not be obtained with such networks. Thus, we enrich our datasets using flipped and rotated versions of each image to artificially enforce such invariance.

Recent works showed that fully connected layers can be kernelized to obtain a fully convolutional networks [20]. Such transformation allows input of various size, which is shown to be beneficial in [13] classification.

After showing the benefits of using several scales and orientations, we propose to aggregate multi-scale information using VLAD across a fixed set of scales and orientations. Specifically, each activation of the feature map is considered as a local descriptors, which are aggregated with equal weights. Unlike the similar NetVLAD [1], our method allows the aggregation over several scales and orientations. To our knowledge, such use of VLAD aggregation over scales and orientation of the activations of various layers has not yet been proposed.

4 Experiments

4.1 Datasets

There is no publicly available dataset of identity documents as they hold sensitive and personal information. Three private datasets are provided by our industrial partner. Images are collected using a variety of sources (scan, mobile photos) and no constraint is imposed. Thus, the documents have any dimension, any orientation, and might be surrounded by complex backgrounds. Figure 1 shows examples of such images.

Preliminary experiments is held on a dataset of 9 classes of French documents (FRA), namely *identity card (front), identity card (back), passport (old), passport (new), residence card (old front), residence card (old back), residence card (new front), residence card (new back), driving licence.* A total of 527 samples are divided into train and test, ranging from 26 to 136 images per class. Then, a larger dataset (Extended-FRA or E-FRA) of the same types of documents with a total of 2399 images (86 to 586 per class) is used. The last dataset consists of 446 samples (8 to 110 per class) of 10 Belgian identity documents (BEL), namely *identity card 1 (front), identity card 1 (back), identity card 2 (front), identity card 2 (back), residence card (old front), residence card (old back), residence card (new front), residence card (new back), passport (new), passport (old).*

4.2 Results

An extensive evaluation is carried out on the image datasets. Three measures are calculated: mean average precision (mAP), overall mean accuracy, and averaged accuracy per class.

Table 1. Evaluation of BOW, VLAD, and Fisher in terms of mAP for detected and dense features, on the FRA dataset.

Encoding	Dim.	Detected SIFT	Dense SIFT
BOW 1k	1k	80.7	79.2
BOW 10k	10k	87.0	85.9
VLAD 16	1k	78.5	81.7
VLAD 64	4k	86.6	90.7
VLAD 256	16k	90.1	91.0
Fisher 16	2k	88.9	88.3
Fisher 64	8k	92.8	93.1
Fisher 256	32k	92.7	92.8

Table 2. Performance of several CNN-based features, on the FRA dataset.

Net. layer	Dim.	mAP	Mean acc.	Acc./class
fast fc7	4k	91.1	85.4	85.6
fast fc6	4k	91.7	81.3	81.6
fast c5 - Avg	256	93.2	89.0	90.5
fast c5 - Max	256	92.9	85.7	88.0
vd19 fc7	4k	87.0	81.9	83.2
vd19 fc6	4k	89.4	85.4	86.0
vd19 c5 - Avg	512	89.6	85.4	86.0
vd19 c5 - Max	512	88.3	83.6	82.3

First, SIFT-based methods are evaluated on the FRA dataset, see Table 1. This comprises BOW, VLAD, and Fisher Vector encodings with several visual vocabulary sizes, and from detected or dense SIFT local descriptors. We note that SIFT descriptors are square-rooted and PCA is applied to obtain 64-dimensional vectors. We observe that Fisher Vector performs better than VLAD, which performs better than BOW. This is expected: the more refined the encoding, the longer the global descriptor, and the better the performances. Even when comparing similar global descriptor dimension, Fisher Vector offers the best performance. Note that Fisher Vector does not improve over 64 Gaussians. Secondly, dense local description overall outperforms detected feature except for the case of BOW encoding. These results agrees with general observations made in computer vision for classification tasks [26].

Then, we evaluate CNN-based descriptors on the same FRA dataset, see Table 2. Two architecture are compared: the 'fast' network [3] and the deep 'vd19' network [30]. Descriptors are obtained by extracting the output of the two first fully connected layers (*fc6* and *fc7*), as well as the last convolutional layer (*c5*). Average and max pooling of *c5* are evaluated as well. Surprisingly, the fast network outperforms vd19. Average pooling is also shown to outperform max pooling for convolutional layer and is preferred in the following experiments. Overall *c5* outperforms *fc6*, which outperforms *fc7*. In fact, lower layers (*c5*) encodes lower level and more generic information, which is less sensible to network training data.

Table 3. Orientation invariance of CNN features, on the FRA dataset.

Net. layer	mAP	Mean acc.	Acc./class
fast fc7	92.5	90.5	91.3
fast fc6	92.3	88.1	86.6
fast c5 - Avg	94.1	90.2	90.7
vd19 fc7	89.9	84.8	85.0
vd19 fc6	90.8	87.2	86.5
vd19 c5 - Avg	91.2	88.7	88.8

Since the CNN feature do not have any rotation invariance mechanism, we propose to enrich the training data collection by adding rotated and flipped images (ending up in 8 distinct descriptors per image), see Table 3. Such process offers a constant improvement for every descriptors.

Further experiments are achieved on the larger E-FRA dataset, see Table 4. Unlike for FRA dataset alone, we observe that *fc6* outperforms *c5*. Unsurprisingly, the more training data the better the performance reaching up to 99% mAP and more than 96% accuracy, when training on E-FRA. More experiment is performed on the BEL dataset, see Table 5. We divide the dataset into three folds, then learn on two third and test on the last one. Scores obtained on all permutations and finally averaged. As for E-FRA, the sixth fully connected layer offers the best performance. Also performances on the BEL dataset are much lower because some classes (residence card (old/front), residence card (old/back), residence card (new/back)) have very few (5 to 12) training samples.

The very recent work of [13] highlighted how input dimension of the image can have a large impact on performance. Therefore, we experiment various input sizes, ($s1 = 224 \times 224, s2 = 544 \times 544, s3 = 864 \times 864$), see Table 6. Concerning convolutional layers, the feature maps are averaged pooled as earlier. However, since fully connected layers require a fixed input feature maps dimension, we kernelize the layers so they are applied at every location of the larger feature map output by the last convolutional layer and finally perform max pooling. We observe a stable gain for every layer, *c5* and *s3* offering the best performance on the FRA dataset. We further note that higher dimensionality (1184×1184, 1504×1504) offers worst results in our experiments.

Since larger scales and multiple orientations encapsulate more precise information, we decide to aggregate the activations of several scales ($s1 = 224 \times 224, s2 = 544 \times 544, s3 = 864 \times 864$, and $s4 = 1184 \times 1184$) and 8 orientations

Table 4. Performance using various combination of the FRA and E-FRA datasets with orientation invariance. Tr Te and E represents Training set of FRA, Testing set of FRA, and E-FRA.

Train/test net. layer	mAP	Mean acc.	Acc./class
Tr/E fast fc7	83.5	81.2	76.7
Tr/E fast fc6	85.6	83.5	78.6
Tr/E fast c5 - Avg	87.8	85.4	83.8
Te/E fast fc7	83.3	83.7	77.8
Te/E fast fc6	84.7	85.8	81.3
Te/E fast c5 - Avg	83.6	86.1	82.1
TrTe/E fast fc7	89.5	86.8	83.9
TrTe/E fast fc6	91.4	89.7	87.6
TrTe/E fast c5 - Avg	90.0	88.5	86.8
TrE/Te fast fc7	97.9	93.6	95.4
TrE/Te fast fc6	99.0	96.3	96.7
TrE/Te fast c5 - Avg	96.6	94.8	95.4
TeE/Tr fast fc7	99.4	96.5	96.4
TeE/Tr fast fc6	99.6	98.0	98.2
TeE/Tr fast c5 - Avg	98.0	94.0	94.4

Table 5. Results obtained on the BEL dataset using 3 folds.

Net. layer	mAP	Mean acc.	Acc./class
fast fc7	71.7	78.6	64.9
fast fc6	73.8	79.0	66.3
fast c5 - Avg	70.9	77.9	60.0

Table 6. Varying scales CNN features with orientation invariance, on the FRA dataset.

Net. layer scale	mAP	Mean acc.	Acc./class
vd19 fc7 s2	94.9	93.9	90.8
vd19 fc7 s3	96.2	95.4	93.1
vd19 fc6 s2	95.1	94.8	91.9
vd19 fc6 s3	95.8	95.4	92.9
vd19 c5 - Avg s2	96.3	95.4	93.1
vd19 c5 - Avg s3	96.3	96.3	94.3

together in a VLAD descriptor. Each activation is centered, PCA reduced to 128 dimensions, and *l2*-normalized. Once concatenated in the VLAD, the final vector is power normalized. Table 7 shows the final performance on the FRA dataset and we observe a stable improvement for all layers reaching a very high performance, around 99% mAP and 98% mean accuracy.

Our application requires fast processing of the scanned documents. We report the computation time of SIFT and CNN features extraction in Table 8. Execution times hold for a single threaded i7 core 2.6 GHz. Note that image dimensions remained unchanged for SIFT features, while images are resized to 224 × 224 for CNN features using the fast network. CNN features are much faster than SIFT, and keypoints detection is quite slow especially for high-resolution images.

Table 7. VLAD aggregation over scales and orientations, on the FRA dataset.

Net. layer	mAP	Mean acc.	Acc./class
vd19 fc7	98.8	97.2	93.2
vd19 fc6	99.2	97.9	94.8
vd19 c5 - Avg	99.5	98.8	95.6

Table 8. Computation time for detected SIFT, dense SIFT, and CNN features extracted from (224 × 224) dimensional features on FRA train/test sets.

Features	Detected SIFT	Dense SIFT	CNN
Average per image	54 s/43 s	5.1 s/4.9 s	0.2 s/0.2 s
Total	180 m/240 m	17 m/27 m	40 s/53 s

To conclude, CNN generate highly effective compact description, largely outperforming earlier SIFT-based encoding schemes from the classification performance and run-time point of view. Secondly, our evaluation provides insight regarding the amount and balance of data required to reach very high performance. Finally, the proposed VLAD aggregation across scales an orientations shows superior performance.

5 Conclusion

This paper addressed the problem of identification documents classification as an image classification task. Several image classification methods are evaluated. We show that CNN features extracted from pre-trained networks can be successfully transferred to produce image descriptors which are fast to compute, compact, and highly performing.

References

1. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR (2016)
2. Bagdanov, A., Worring, M.: Fine-grained document genre classification using first order random graphs. In: ICDAR, pp. 79–83 (2001)
3. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
4. Chen, S., He, Y., Sun, J., Naoi, S.: Structured document classification by matching local salient features. In: ICPR, pp. 653–656. IEEE (2012)
5. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: Proceedings of the IEEE CVPR, pp. 3828–3836 (2015)
6. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: International Workshop on Statistical Learning in Computer Vision (2004)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR. IEEE (2009)
8. Eglin, V., Bres, S.: Document page similarity based on layout visual saliency: application to query by example and document classification. In: ICDAR (2003)
9. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010). Springer
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Trans. PAMI **32**(9), 1627–1645 (2010)
11. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 392–407. Springer, Cham (2014). doi:10.1007/978-3-319-10584-0_26
12. de las Heras, L.P., Terrades, O.R., Llados, J., Fernandez-Mota, D., Canero, C.: Use case visual bag-of-words techniques for camera based identity document classification. In: ICDAR, pp. 721–725, August 2015
13. Herranz, L., Jiang, S., Li, X.: Scene recognition with CNNs: objects, scales and dataset bias. In: CVPR, pp. 571–579 (2016)
14. Jégou, H., Perronnin, F., Douze, M., Schmid, C., et al.: Aggregating local image descriptors into compact codes. Trans. PAMI **34**, 1704–1716 (2012)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
16. Kumar, J., Doermann, D.: Unsupervised classification of structurally similar document images. In: ICDAR, pp. 1225–1229. IEEE (2013)
17. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI, vol. 333, pp. 2267–2273 (2015)
18. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
19. Liu, L., Shen, C., Wang, L., van den Hengel, A., Wang, C.: Encoding high dimensional local features by sparse coding based fisher vectors. In: NIPS (2014)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
21. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
22. Oquab, M., Bottou, L., Laptev, I., Sivic, J., et al.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR (2014)

23. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_11
24. Shin, C., Doermann, D.: Document image retrieval based on layout structural similarity. In: IPCV, pp. 606–612 (2006)
25. Sicre, R., Avrithis, Y., Kijak, E., Jurie, F.: Unsupervised part learning for visual recognition. In: CVPR (2017)
26. Sicre, R., Gevers, T.: Dense sampling of features for image retrieval. In: ICIP, pp. 3057–3061. IEEE (2014)
27. Sicre, R., Jégou, H.: Memory vectors for particular object retrieval with multiple queries. In: ICMR, pp. 479–482. ACM (2015)
28. Sicre, R., Jurie, F.: Discriminative part model for visual recognition. CVIU **141**, 28–37 (2015)
29. Sicre, R., Tasli, H.E., Gevers, T.: Superpixel based angular differences as a mid-level image descriptor. In: ICPR, pp. 3732–3737. IEEE (2014)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
31. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
32. Tasli, H.E., Sicre, R., Gevers, T., Alatan, A.A.: Geometry-constrained spatial pyramid adaptation for image classification. In: ICIP (2014)
33. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: ICLR (2016)
34. Xing, C., Wang, D., Zhang, X., Liu, C.: Document classification with distributions of word vectors. In: APSIPA
35. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS (2014)