Complex Document Classification and Localization Application on Identity Document Images

Ahmad-Montaser Awal, Nabil Ghanmi, Ronan Sicre, Teddy Furon

Complex Document Classification and Localization Application on Identity Document Images

Ahmad-Montaser Awal, Nabil Ghanmi
AriadNEXT - Pôle R&D document
Rennes, France
Email: {montaser.awal, nabil.ghanmi}@ariadnext.com

Ronan Sicre
IRISA
Rennes, France

Teddy Furon
INRIA - LinkMedia
Rennes, France

***Abstract*—This paper studies the problem of document image classification. More specifically, we address the classification of documents composed of few textual information and complex background (such as identity documents). Unlike most existing systems, the proposed approach simultaneously locates the document and recognizes its class. The latter is defined by the document nature (passport, ID, etc.), emission country, version, and the visible side (main or back). This task is very challenging due to unconstrained capturing conditions, sparse textual information, and varying components that are irrelevant to the classification, *e.g.* photo, names, address, etc.**

First, a base of document models is created from reference images. We show that training images are not necessary and only one reference image is enough to create a document model. Then, the query image is matched against all models in the base. Unknown documents are rejected using an estimated quality based on the extracted document. The matching process is optimized to guarantee an execution time independent from the number of document models. Once the document model is found, a more accurate matching is performed to locate the document and facilitate information extraction. Our system is evaluated on several datasets with up to 3042 real documents (representing 64 classes) achieving an accuracy of 96.6%.

I. Introduction

Identity fraud has always been a cat and mouse play between counterfeiters and authorities multiplying the security checks in identity documents (*e.g.* holograms, watermarks, paper patterns). This raises the issue that only experts from border police departments are knowledgeable for a complete authentication of documents. The threats of identity fraud vary from small frauds up to organized crimes and terrorist actions. Most small forgeries are indeed not so elaborated aiming at deluding people with little expertise (hotels, casinos, telecom companies,...). In this case, an automatic fraud detection system is more feasible than the second case.

The work presented in this paper is part of a research project IDFRAud[1] proposing a platform for identity documents verification and analysis. First, the input document class is recognized (type, country, series, ...). Then, the security checks of this particular model are verified.

This paper presents an automatic classification method addressing the following challenges:

[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/

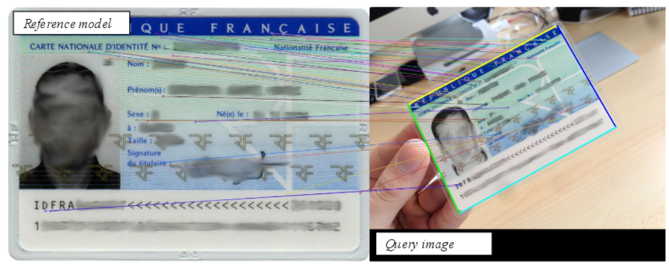

Fig. 1. Example of matched and cropped document

- Robustness to the capturing conditions: average image quality of smartphones in the wild (*e.g.* complex background, occlusions, flares)
- Scalability: the number of classes for an international coverage easily reaches a thousand.
- Scarcity of training documents: identity documents are not available in large quantities.
- Localization: the document must be geometrically localized with high precision in order to ease information extraction and security checks verification.

The next section briefly summarizes the state of the art. Section III presents our document classifier, evaluated on a set of real documents in section IV. The paper is finally concluded with a discussion of the system limitations and some perspectives.

II. State of the Art

Document classifiers can be divided into three main families analyzing the document layout, textual information, or the visual content. The *layout* based approach is adapted to well structured and text-rich documents such as journals, articles, or invoices. The document is described by the spatial layout of text blocks, figures, tables, *etc*. [15]. This layout is the keystone to calculate document similarities [11, 26] or to construct models of document classes [3, 10, 17]. This approach is not discriminant enough as identity documents of different classes may share similar layouts. The alternative *text* based approach typically constructs a global descriptor of the textual content (*e.g.* Bag of Words –BoWs or Word2Vec) which are then analyzed by classical classifiers (SVM, ANN, ...) [34].

Recently, recurrent neural networks have been used to merge feature extraction and document classification [18]. However, our application presents specific difficulties, which prevents the use of the aforementioned methods. The document is not localized a priori in the image and drowned in background (see Fig. 1). Furthermore, textual information is not easy to extract before knowing the class and the layout of the document.

The *visual* based approaches are the best suited to address the problem of identity document recognition since they usually have a characteristic graphical structure. The authors in [25] proposed a method merging both visual and textual descriptors. A large part of other visual based techniques follow the BOW approach [9, 14]; where local features are extracted [20], encoded [16, 23], pooled [19, 32] and then used for classification [6].

Recently, deep learning networks have been successfully applied to image classification [13] and retrieval [33], giving results significantly above the BOW approach. These networks have a much deeper structure than standard representations, including several convolutionnal layers followed by fully connected layers. This comprises a very large number of parameters that have to be learned from big training datasets. Compact structured representation from intermediate to a high-level can be extracted from such networks [22]. Furthermore, deep learning representations can be encoded with VLAD [1] or Fisher vectors [8]. It is worth mentioning that part-based approaches, which learn a set of discriminative parts to model classes [12, 28, 29], are highly effective in similar fine grain classification but computationally expensive.
Some systems already apply the visual based approach to document classification. The authors of [31] use global image descriptor with the assumption that the document is already localized and extracted. Paper [2] proposes a classification of scanned identity document into two classes using local descriptors. A comparative study of local detectors and descriptors for document classification task is given in [24].

In our work, we make no assumptions on the capturing of the document images. They vary from high quality scans to low quality mobile phone photos. Our scheme is inspired from [2] with the following improvements:

- As in [2], class models are created from one reference image when available. Yet, a main difference of our model creation method is that it can also cope with several training images. This improves the quality of the model when variable zones cover a large part of the document and masking these areas removes too much information. Masking is thus switched off and the training of several images filters unnecessary key-points
- Masked zones are ignored during the key-points extraction instead of being replaced by white rectangles.
- The classification run-time is independent from the number of classes.
- The evaluation involves 64 document classes.
- The quality of the document localization is estimated to measure the classification confidence.

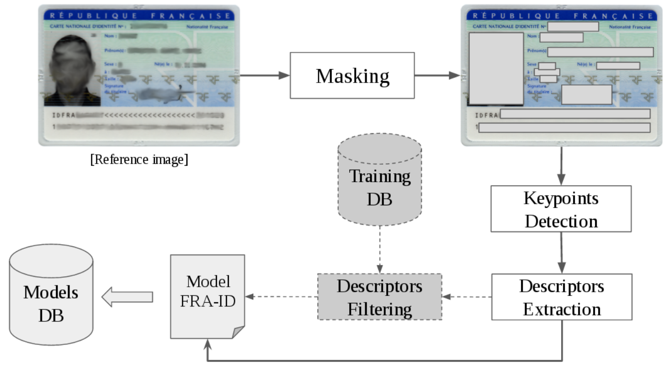

Fig. 2. Creation of the French ID reference model

- The classifier may abort if not confident enough. This allows the rejection of unknown document classes, which is of utmost importance in a fully automatic system.

III. Proposed approach

A. Creation of reference models

Identity documents contain invariable text zones (field labels: name, surname, etc.), variable text zones (personal data), and the same background. A reference model is created for each class in order to obtain the reference model base. Since the availability of document samples is very limited, a reference model may be created using either one document image or a set of documents when available.

As depicted in Fig. 2, the variable zones are first masked manually. Then, keypoints are extracted and characterized (ignoring those in the masked zones) by a local description method (SIFT, SURF, ORB, etc.). SURF [4] has been used for the experiments held for this paper. The advantage of local description is its invariance to affine transformations such as translation, scaling, rotation. When dealing with a training set, only keypoints that have a match in every training image are kept. The i-th model is denoted by a set M_i of n_i keypoints where $M_i = \{D_{i,1}, D_{i,2}, ..., D_{i,n_i}\}$ and D_{ij} is the set the extracted descriptors from the keypoint j. Each model is indexed with Random KD trees [30] from the FLANN library [21] in order to accelerate the direct matching process during the classification step.

B. Document classification

The classification of a query image is performed based on its keypoints. It consists in finding the winner class, which has the maximum keypoints matching with those of the query. The winner class is determined as follows (see Fig. 3).

The query keypoints are first matched against all learned models altogether, since matching models one by one is slow and prevents scalability. Thus, all models compete with each others in this matching phase. The set of *direct matches* for model i is denoted m_i. The models are ranked by the ratio $r_i = \frac{|m_i|}{|M_i|}$, and the model with the highest ratio is the first candidate to be the winner class: $g = \arg\max_i(r_i)$.

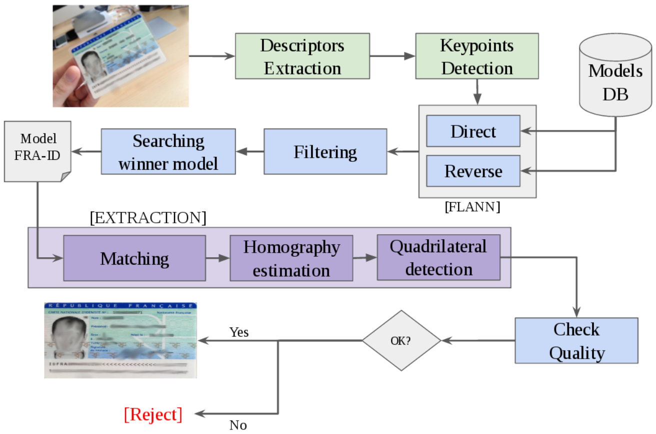

Fig. 3. Document classifier structure

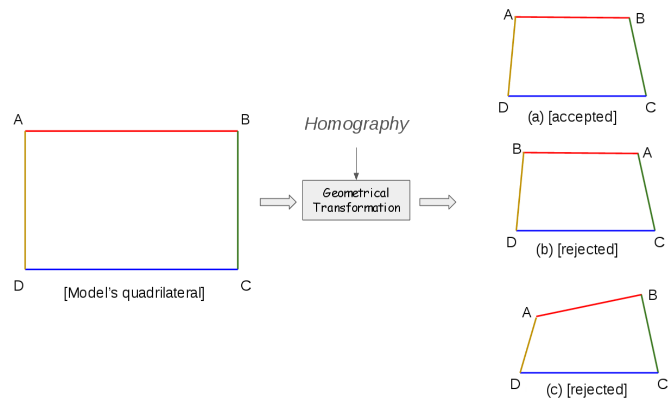

Fig. 4. Examples of valid and rejected quadrilateral

Then, all models are matched against the query producing a set m'_i of reverse matches. We define V_i as the set of valid matches of the i^{th} model as follows:

- Let S_i be the set of symmetric mapping, i.e. the set of couples of keypoints that match in the two directions (direct and reverse).
- The histogram of the orientation difference in each couple of points from S_i is computed to determine the dominant orientation difference $\Delta\theta$. The set O_i gathers couples sharing $\Delta\theta$.
- The RANSAC algorithm [7] finds the geometric transformation that maps the greatest number of point couples of O_i and eventually excludes outliers. This defines the set of valid matches V_i.

The other models are examined in the decreasing order of their ratios as follow:

for *each model* **do**
if $(|m_i| > |V_g|) \wedge (|S_i| > |V_g|) \wedge (|O_i| > |V_g|) \wedge (|V_i| > |V_g|)$ **then**
Update the winner class: $g = i$

This sequence of tests is checked in the same order (of their appearance in the condition) to avoid the computation of the set of valid matches V_i and the other subsets (S_i and O_i) when not necessary.

C. Extraction and quality estimation

The RANSAC method estimates a geometric transformation matrix H from the query image keypoints matching those of the winner model:

$$H = \begin{pmatrix} a_{11} & a_{12} & t_1 \\ a_{21} & a_{22} & t_2 \\ v_1 & v_2 & 1 \end{pmatrix} \tag{1}$$

Fig. 1 illustrates an example of the detection process (*i.e.* classification & localization). Once the document is localized and extracted, further analysis is required in order to read the document information (name, birth-date, ...) and to verify its authenticity (not detailed in this paper).

Unknown documents are *rejected* based on the estimated quality of the homography H and the detected quadrilateral Q_r, given the model's quadrilateral Q_m:

$$Q_r = H \times Q_m. \tag{2}$$

The detected quadrilateral is defined by four vertices (A, B, C, D) and four edges. The homography must keep the vertices order as in the document model. The query image is rejected, *i.e.* considered as an unknown class, if the determinant d of H is negative (see Fig. 4.b), where:

$$d = a_{11} \times a_{22} - a_{12} \times a_{21} \tag{3}$$

Furthermore, an ideal detected quadrilateral approaches a rectangular form. However, document photo capture often generates a perspective deforming the rectangle into an isosceles trapezoid (one perspective) or a parallelogram (two perspectives). The query image is rejected if the detected quadrilateral does not meet the following criteria (e.g. Fig. 4.a):

1) At least one pair of the opposed edges is parallel (with a tolerance of 5°)

$$(\overrightarrow{[AB]} - \overrightarrow{[CD]}) < 5° \quad \Rightarrow \quad [AB] /\!/ [CD]$$
$$(\overrightarrow{[AD]} - \overrightarrow{[BC]}) < 5° \quad \Rightarrow \quad [AD] /\!/ [BC]$$

where $\overrightarrow{[AB]}$, $\overrightarrow{[CD]}$, $\overrightarrow{[AD]}$ and $\overrightarrow{[BC]}$ denote the edge angles with the horizontal axe.

2) The average difference of angles between each pair of opposed angles is less than 10°

$$[AB] /\!/ [CD] \quad \Rightarrow \quad \frac{(\hat{A} - \hat{B}) + (\hat{C} - \hat{D})}{2} < 10°$$
$$[AD] /\!/ [BC] \quad \Rightarrow \quad \frac{(\hat{A} - \hat{D}) + (\hat{B} - \hat{C})}{2} < 10°$$

3) Average perpendicularity of the four vertices is less than 25°

$$\left| \frac{\hat{A} + \hat{B} + \hat{C} + \hat{D}}{4} - 90 \right|° < 25°$$

TABLE I
DOCUMENT DATASETS

DB	#Classes	#Countries	#Images
BEL_DB	10	1 (Belgium)	446
FRA_DB	9	1 (France)	2494
INT_DB	64	12	3042

IV. EXPERIMENTATION

There is no publicly available dataset of identity documents as they hold sensitive and personal information. Three private datasets provided for this experimental work have been used (see Table I). Images are collected using a variety of sources (scan, smartphone, triple lightning scanners) without any imposed constraint. The document in a query image may have any dimension or orientation surrounded by a complex background.

The work of [27] performs an extensive evaluation on the image datasets using state of the art image classification methods. Our system is compared to CNN-based classification using the 'fast' network [5] on both FRA_DB and BEL_DB. Descriptors are extracted from the two first fully connected layers (*fc6* and *fc7*), as well as the last convolution layer (*c5*) and followed by pooling. Unlike the proposed method, these results were obtained using 527 training samples for the FRA_DB and a three fold cross validation for the BEL_DB. We observe from the Table II that the *fc6* descriptors outperforms those extracted from the other layers achieving 89.7% and 79.0% on FRA_DB and BEL_DB respectively. We observed that CNN-based approaches offer good classification performance. Nevertheless, they are not rotation invariant. Furthermore, performance degrades when training images are too few or when classes are unbalanced. The proposed approach overcomes these difficulties and reaches 95.8% and 94.7% accuracy on FRA_DB and BEL_DB respectively. Furthermore, it achieves 96.6% when tested on the INT_DB confirming the scalability of the proposed method. As illustrated in the classification confusion matrix (Figure 5), most of classes (38 out of 64) reaches an accuracy of 100%. In addition, only three classes have confusions greater than 10%. However, these cases correspond to documents sharing the same background with lightly different textual layout. In addition, other failures are due to poor image quality (noise, flash, ..) since the input images are very variable and we do not impose any constraints on document capturing process.

Finally, we evaluate the different filtering and the direct and reverse matching steps. The Table III illustrates the obtained accuracy using different configurations of matching and filtering. We note that the RANSAC filtering significantly improves the accuracy in every configuration. Similarly, using both direct and inverse matches improves the classification since uninformative matches are filtered out.

It worth mentioning that the classification time increases by 20 ms for each additional class (1.5 $seconds$ for 10 classes against 2.8 $seconds$ for 64 classes). However, a simple filtering of models with few matches before the reverse flann yields a constant classification time with a very light loss in accuracy (around 1%).

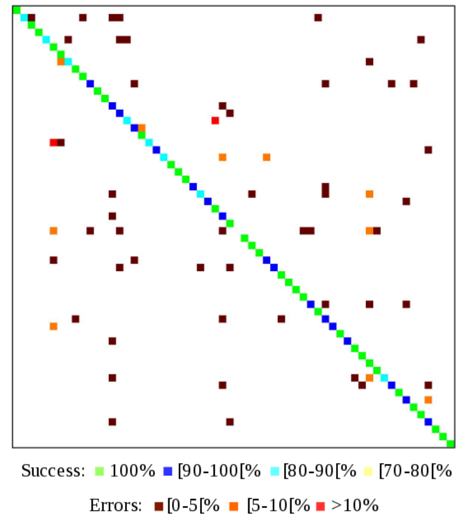

Fig. 5. INT_DB confusion matrix

TABLE II
CLASSIFICATION EVALUATION USING STATE OF THE ART METHODS AND OUR PROPOSED METHOD

DataSet	#Classes	#Samples	Classification	Accuracy%
BEL_DB	10	446	fastfc7 + SVM	78.6
			fast fc7 + SVM	79.0
			fast c5 + SVM	77.9
			Proposed method	**94.7**
FRA_DB	9	2494	fast fc7 + SVM	86.8
			fast fc6 + SVM	89.7
			fast c5 + SVM	88.5
			Implementation of [14]	87.0
			Proposed method	**95.8**
INT_DB	64	3042	Implementation of [2]	84.8
			Proposed method	**96.6**

V. CONCLUSION

In this work, the proposed approach successfully classifies, and extracts identity documents in the wild. First, a coarse

TABLE III
CLASSIFICATION EVALUATION ON THE INT_DB USING DIFFERENT CONFIGURATIONS OF OUR SYSTEM

Direct Matches	Reverse Matches	Symmetry Filter	Orientation filter	RANSAC	Accuracy%
✓					66.3
✓				✓	82.6
✓			✓		77.8
✓			✓	✓	85.9
✓	✓	✓			67.9
✓	✓	✓		✓	94.9
✓	✓	✓	✓		86.7
✓	✓	✓	✓	✓	***96.6***
	✓				59.6
	✓			✓	82.5
	✓		✓		77.7
	✓		✓	✓	85.8

keypoints matching associates the document image to one of the document models. Then, a fine-grained matching is employed in order to localize and extract the document. The system has been evaluated on real-world datasets and high performance is obtained. As future work, we would like to investigate denser keypoints extraction. This is particularly helpful when documents do not contain much visual elements. Furthermore, document extraction quality can be improved by adding geometrical and structural constraints to improve the valid matches set V_i. In addition, the quality of the extracted document can also be improved by preventing perspective transformations when images are captured by scans.

References

[1] Relja Arandjelović et al. "NetVLAD: CNN architecture for weakly supervised place recognition". In: *CVPR*. 2016, pp. 5297–5307.

[2] Olivier Augereau, Nicholas Journet, and Jean-Philippe Domenger. "Reconnaissance et Extraction de Pièces d'identité". In: *CIFED*. 2012, pp. 179–194.

[3] Andrew Bagdanov and Marcel Worring. "Fine-grained document genre classification using first order random graphs". In: *Sixth ICDAR*. 2001, pp. 79–83.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded Up Robust Features". In: *Computer Vision – ECCV 2006*. Vol. 3951. Lecture Notes in Computer Science. 2006, pp. 404–417.

[5] Ken Chatfield et al. "Return of the Devil in the Details: Delving Deep into Convolutional Nets". In: *British Machine Vision Conference*. 2014.

[6] Siyuan Chen et al. "Structured document classification by matching local salient features". In: *21st ICPR*. 2012, pp. 653–656.

[7] Ondrej Chum, Jiri Matas, and Josef Kittler. "Locally Optimized RANSAC". In: *DAGM-Symposium*. Vol. 2781. Lecture Notes in Computer Science. 2003, pp. 236–243.

[8] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. "Deep filter banks for texture recognition and segmentation". In: *CVPR*. 2015, pp. 3828–3836.

[9] Gabriella Csurka et al. "Visual categorization with bags of keypoints". In: *Workshop on statistical learning in computer vision, ECCV*. 2004.

[10] Michelangelo Diligenti, Paolo Frasconi, and Marco Gori. "Hidden tree Markov models for document image classification". In: *IEEE Transactions on PAMI* 25.4 (2003), pp. 519–523.

[11] Véronique Eglin and Stephane Bres. "Document page similarity based on layout visual saliency: application to query by example and document classification". In: *Seventh ICDAR*. 2003, pp. 1208–1212.

[12] Pedro F Felzenszwalb et al. "Object detection with discriminatively trained part-based models". In: *IEEE transactions on PAMI* 32.9 (2010), pp. 1627–1645.

[13] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[14] Lluis-Pere de las Heras et al. "Use case visual Bag-of-Words techniques for camera based identity document classification". In: *13th ICDAR*. Tunis, Tunisia: IEEE, 2015, pp. 721–725.

[15] Jianying Hu, R. Kashi, and G. Wilfong. "Document image layout comparison and classification". In: *Fifth ICDAR*. 1999, pp. 285–288.

[16] Hervé Jégou et al. "Aggregating local image descriptors into compact codes". In: *IEEE Transactions on PAMI* (2012), pp. 1704–1716.

[17] Jayant Kumar and David Doermann. "Unsupervised Classification of Structurally Similar Document Images". In: *12th ICDAR*. 2013, pp. 1225–1229.

[18] Siwei Lai et al. "Recurrent Convolutional Neural Networks for Text Classification." In: *AAAI*. Vol. 333. 2015, pp. 2267–2273.

[19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *CVPR*. Vol. 2. 2006, pp. 2169–2178.

[20] David Lowe. "Object recognition from local scale-invariant features". In: *ICCV* (1999), pp. 1150–1157.

[21] Marius Muja and David G. Lowe. "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration". In: *VISSAPP*. INSTICC Press, 2009, pp. 331–340.

[22] Maxime Oquab et al. "Learning and transferring mid-level image representations using convolutional neural networks". In: *CVPR* (2014), pp. 1717–1724.

[23] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. "Improving the Fisher Kernel for Large-Scale Image Classification". In: *ECCV*. 2010, pp. 143–156.

[24] Marçal Rusiñol et al. "A comparative study of local detectors and descriptors for mobile document classification". In: *13th ICDAR*. IEEE. 2015, pp. 596–600.

[25] Marçal Rusiñol et al. "Multimodal page classification in administrative document image streams". In: *IJDAR* 17.4 (2014), pp. 331–341.

[26] Christian Shin and David Doermann. "Document Image Retrieval Based on Layout Structural Similarity". In: *IPCV*. 2006, pp. 606–612.

[27] Ronan Sicre, Ahmad Montaser Awal, and Teddy Furon. "Identity documents classification as an image classification problem". In: *ICIAP*. 2017.

[28] Ronan Sicre and Frédéric Jurie. "Discriminative part model for visual recognition". In: *Computer Vision and Image Understanding* 141 (2015), pp. 28–37.

[29] Ronan Sicre et al. "Unsupervised part learning for visual recognition". In: *CVPR*. 2017.

[30] C. Silpa-Anan and R. Hartley. "Optimised KD-trees for fast image descriptor matching". In: *IEEE CVPR*. 2008, pp. 1–8.

[31] Marcel Simon, Erik Rodner, and Joachim Denzler. "Fine-grained Classification of Identity Document Types with Only One Example". In: *14th IAPR ICMVA*. Tokyo, Japon: IEEE, 2015, pp. 126–129.

[32] H. Emrah Tasli et al. "Geometry-constrained spatial pyramid adaptation for image classification". In: *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 1051–1055.

[33] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. "Particular object retrieval with integral max-pooling of CNN activations". In: *ICLR*. 2016.

[34] Chao Xing et al. "Document classification with distributions of word vectors". In: *APSIPA*. IEEE. 2014, pp. 1–5.