Shall deep learning be the mandatory future of document analysis problems?

Nicole Vincent, Jean-Marc Ogier

Shall deep learning be the mandatory future of document analysis problems?

Nicole Vincent[1], Jean-Marc Ogier[2]

[1] *Laboratoire LIPADE-SIP, ParisDescartes University, 45 rue des Saints-Pres, 75006, Paris, France*

[2] *Laboratoire L3i, La Rochelle University, 17000 La Rochelle, France*

Abstract

As the use of deep methods become widespread in the scientific community, causing major changes in systems architecture and position in terms of knowledge acquisition, we report here our insights about how document analysis systems are built. Where does the expertise really lie? In the features, in the decision making step, in the system design, in the data illustrating the problem to be solved? The examination of the practices of researchers in this field, and their evolution, allows us to conclude that the tools that are used, and related issues, have become more and more complex over time. Nevertheless, human skill is needed to activate these tools and to imagine new ones.

Keywords: Features, Machine Learning, Deep methods, Handcraft approach

1. Introduction

The widespread use of electronic documents including the digitally native documents is made easier today through effective means of electronic management. In the late 20th century, the main issue was to turn content on paper, or more precisely images of paper documents into digital and editable documents. It took many years for the struggling scientific community to implement reading systems that would be able to process a huge variability of documents going from forms [1], bank checks [2].... to the management of mail and financial documents [3].

Even though industrial systems are generally operational, the natural use of e-documents, which in fact are hybrid documents, is not actually obtained and may still be improved in many ways. If a document can be natively in a digital format, such as proprietary doc format or free pdf format, then it can be printed, scanned or snapped with a camera. The processing of hybrid documents, both paper and digital, represents a real scientific challenge for the document analysis community, despite the fact that most documents are now digitally produced. New processing must consider the hybrid nature of documents. It is also important to note that document analysis has evolved: it is not just about reading [4], managing or understanding a paper document from a screenshot [5] [6]. The aim is now to make the document "smart".

Indeed, through the print-and-scan processes, the initial structure as well as the content of the document may be lost. This explains the constant need for new document analysis algorithms, especially in the preprocessing phase. Document processing involves classifying documents according to their layout [7] and/or content [8], focusing on some parts of the document only, achieving reverse engineering [9] or information retrieval [10] [11] [12], authenticating documents [13] [14].

In this paper, we analyze the different types of characteristics or features that have been developed in recent years, with mixed results, to achieve a better understanding of current issues and prepare for a better future. Prior to our analysis of evolution of the complexity of features in section 3, we present various applications used for document analysis in section 2. We then show in section 4 how prior knowledge of data/documents, serving as a basis for analysis, can be incorporated into systems. The recent developments of deep learning-based approaches are examined in section 5. Provisional conclusions are presented in section 6.

2. Various applications

Most document analysis studies aim to simulate reading as a human task. However, in reality, when working on a document, the cognitive activity is

actually more complex than a "simple" reading task. For the reader, indeed, understanding the content of a document requires interpreting various elements such as printed or handwritten notes, text, tables, graphics or even images, as well as text contents, global document apparearance. These elements are sometimes mentioned as contextual information. Nevertheless, closer analysis shows that reading involves many different and difficult tasks, due tovariability of document types, content and layout :

- At first, character recognition in English or any Latin languages. The Asian and Arabic alphabets were considered later and more recently southern-eastern languages. Should the recognition methods remain the same for these different contexts? The universal aspect of a method is questionnable in various contexts, when the reading concerns forms, addresses, full text or graphical documents or even text inside images capturing natural scenes. Are language models really needed? Besides, the recognition process focused initially on segmented printed characters. Studies today aim to address the issues of constraint-free modern and historical handwritten documents, which appears to be a great scientific challenge. for the understanding of the document context.
- Layout, text or image extraction, graphical drawing interpretation.
- Alphabet separation as a preprocessing of text reading.
- Authentication, writer identification.
- Old documents analysis, classification or reading.
- Information retrieval, through multimodality queries.
- Security involving different levels, especially in a print-and-scan workflow process: are the content meanings of two documents equivalent or should the doucment be an occurrence of a hybrid document, as presented in the introduction?

3. Simple or Complex features

In order to achieve these various applications, the image content has to be analyzed. The systems rely on a common design structure, first comes the choice of a representation space and then the decision step chosen according to the application. In this section, we focus on the representation space that has long been a matter of debate when it comes to achieve best results. We try to give a coarse outline of the most classical features.

The first simplest representation is the representation of the image. At its core, the representation of an image is an array of pixels depicting the real world. Then, the simplest feature is the color (or grey level) of a single pixel, that can be extended to the whole image, forming a feature vector. Time going on, Through time, the single pixel approach has been replaced by a multiple pixels approach, a zone approach, such as the content of fixed or adaptive rectangular zones or some straight or curved lines in the image. From these geometrical fixed analyzed windows some algorithms have been developed to extract some significant zones in the binary images. For instance, connected components were studied for their size, geometric or density properties/features. In color images, superpixels have been defined. One of the difficulties of image analysis with respect to signal processing lies in the 2D dimension of the image signal. There have been attempts to transfer image content to chain representation. This is the case using the connected component contours and the Freeman code [15] in a binary image. Another way of doing it was to consider the image skeleton [16], with 1D contribution but fundamentaly a 2D object. The aim is to decrease the complexity of an image content. Nevertheless, the image is considered as a 2D object, either features can be extracted for describing its content or characteristic vectors can be considered as a way to get more information. Both are used to perform comparisons. In this modeling approach handcrafted features are obtained from the expert perception of the image content. Some features are derived from the general knowledge in pattern recognition, based on solid mathematical thorems, such as a decomposition of multivariate functions

(images are defined in a two-dimensional space). In this category of features, can be considered all the moments, (geometric, Zernike, Fourier-Mellin), the discrete Fourier transform. These features have been largely used. Since the works of Hu in 1961, invariant moments [17], which are based on combinations of regular geometric moments, have been very frequently used [18]. Among the various moments, one can thus cite Zernike moments [19], [20],[21], [22], [23]. They consitute a reference in the domain, giving pseudo-Zernike moments [19], Bamieh moments [24], and Legendre moments [25]. These invariant moments, which can be extracted from a binary or a grey-scale image, generally offer properties of reconstructibility, thus ensuring that extracted features contain all the information, at least enough information, about the shape under study. This completeness property shows the coarseness of description for such invariants. Good comparative studies about invariant moment can be found in [26] and [27], showing the superiority of Zernike moments in terms of recognition accuracy.

The most frequent description models deal with Fourier descriptors [28] or elliptic Fourier descriptors [29]. Taxt [30] proposed a comparative study between these descriptors. This study highlights the potentialities of these descriptors, in terms of simplicity and robustness. Taxt [30] in particular, proposed to use elliptic Kuhl moments [31] when characters orientation is known. This technique is also used by Trier [32] for character recognition on hydrographic maps. Authors claim to obtain a good recognition rate of 78 % for only 2,3% of misclassification on a test set of 1760 hand-written characters. The standard recognition rate has improved a lot since that time! The features were basic including template matching.

On the other hand, in the context of optical character recognition, structural invariant features can also be extracted from characters or thinned characters [33]. For instance, one can cite the number of occlusions, the number of T-joints or X-joints, the number of bend points. However, it has been shown that such features used alone do not lead to robust recognition systems [34]. Graph representations are then introduced [35]. This period corresponds to the beginning of the opposition between statistical methods and Artificial Intelligence that advo-

cated description which were natively invariant using elementary features and relations between them. Graphs were one of the mathematical tools enabling this kind of representation and the decision step was relying on graph similarity, among which graph edit distance, graph isomorphisms for instance. The theoretical aspect was attractive but the results at that time were not robust enough from an industrial point of view, especially because of complexity questions. In fact, the descriptions were too simple to cope with the shape variability, and the machine capacities could not allow to manage graph with too many vertices. Nowadays, larger graphs are used and comparison is achieved by some graph embedding methods [36] or based on graph kernel function [37] or on graph edit distance [38].

More recently, and in the same trend linked to the principle of the decomposition of a function according to a function basis. There have been various wavelet transforms leading to some sets of coefficients characterizing the image content. These transforms do not only consider the initial image but also different observation scales, leading to numerous possibilities of coefficients. A limited number of coefficents have to be selected among these coefficients in order to be considered as features. Many contributions have focused on the exploitation of wavelet theory for pattern recognition. A wavelet transform differs from a Fourier transform in that it is able to provide both a frequency and time representation of a 1D or a 2D signal. It allows to perform a local and global analysis of a shape [39], [40]. It is therefore particularly adapted to the extraction of a model of discriminant shape thanks to this multi-resolution aspect [41]. Among the works exploiting this wavelet theory, two can be cited, Shen [42] and Chen [43] approaches.

In Shen approach [42], the authors first present a general framework of invariant shape recognition with respect to rotation. By analyzing the different works presented in the literature, the authors infer that all moment-based approaches can be expressed through formula 1 .

$$F_{\mathrm{pq}} = \int \mathrm{S_q(r).g_p(r)rdr} \quad with \quad \mathrm{S_q}(r) = \int \mathrm{f(rcos\theta, rsin\theta)e^{iq\theta}d\theta} \tag{1}$$

In this expression, F_{pq} denotes an invariant moment of order pq (p and q are integers), and g_p denotes a function of the radial variable, let us say r in formula 1. By way of example, if $g_p(r) = r^p$, it is possible, by imposing constraints on p and q, to express the Hu moments. The author proposes also a formulation of $g_p(r)$ allowing the computation of the moments of Zernike, without however expressing all the compelling constraints on p and q. From this proposition of a unified moment extraction framework, the authors replace the function g_p with a basic function of wavelets, considering the family defined by formula 2.

$$\psi_{a,b}(r) = \frac{1}{\sqrt{a}}\psi(\frac{r-b}{a}) \tag{2}$$

The mother wavelet used in the article is the Cubic B-Spline wavelet, whose formulation, proposed in [Unser 1996], is given by formula 3

$$(r) = \frac{4a^{n+1}}{\sqrt{2\pi(n+1)}}\sigma_w cos(2\pi f_0(2r-1))exp(-\frac{(2r-1)^2}{2\sigma_w^2(n+1)}) = g_{\mathrm{n}}(r) \tag{3}$$

To model the shapes contained in the image, the authors then use the magnitude of F_{pq}, making it possible to obtain a shape description, which is invariant with respect to rotation. Invariance to a change of scale is obtained by means of a standard process of normalization comparable to regular moments.

Chen's approach [43] differs somewhat from the previous one. Indeed, starting from the qualities and defects of the Fourier transform and the wavelet transform, the author uses them both. The principle used can be described by the synoptic in figure 1. The first step in this model extraction is to normalize the shape and change the cartesian variables into the polar domain. To do this, the authors use a zoning of their shape, based on concentric circles and on radial probes. An averaging in each zone makes it possible to change Cartesian coordinates to polar coordinates. Then, a 1D Fourier transform is applied along the angular axis (θ), to obtain invariance regarding rotation, embedding the

famous transfer theorem in Fourier analysis. A wavelet transform is then applied, this time following the radial variable r in order to take advantage of the multi-resolution aspect of the wavelets.

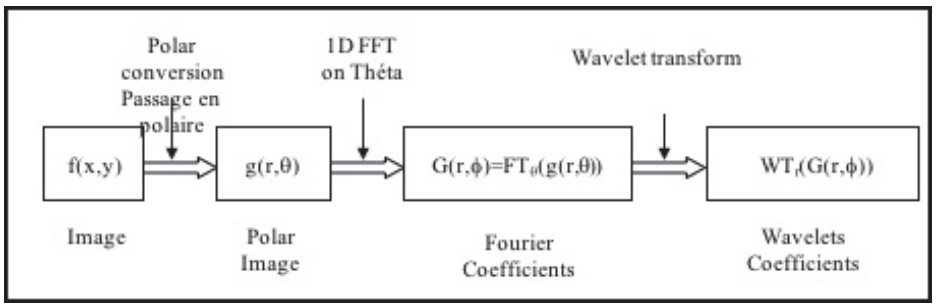

Figure 1: synoptic of Chen approach [43]

The use of this approach is still quite frequent in the community [44][45]. It is important to note that with these wavelet based approaches, all associated coefficients are linked to the projection of the image function on specific function families. Some properties of orthogonality among the couples of functions in the family render the properties of the coefficients more or less efficient. Most wavelet basis are chosen to optimize the number of coefficients to be retained. Indeed, what is important is the convergence speed of the truncated function model towards the initial image function. Mathematical theorems similar to Stone-Weierstrass theorem [46], ensure the theoretical convengence. Of course, the convergence speed is very important in a computer science application. Besides, theoretical properties most often hold for continuous function whereas computation is performed in a discrete space. Thus, some different sources of approximation as well as the number of efficient coefficients yield some variations in the use of these feature vectors according to the developped applications. The selection of coefficients is achieved by maximizing the information provided at the right level for any given application. This may be the reason why this more signal processing approach did not received wide acclaim in the document community more prone to pattern recognition approaches. Besides, the use of continous functions tends to introduce discretisation errors. Discontinous functions, peacewise constant functions, are more appropriate as in Harr transform

or Harr Wavelets. In [47], the optimum observation scale is looked for before application of an Harr transform.

Some features bear similarities with the Human Visual System. Implementing those features led to some pioneering work by Marr [48]. Marr's work served as a basis for the use of Gabor transform [49] that is supposed to correspond to cabled operators present in the human nervous system [50]. Features can also be deduced from some perception theories such as Gestalt theory. In this case, alignment properties are looked for as well as distance from elements to their nearest neighbors. From this approach, one of the greatest lessons that has emerged is the concept of interest points, also known as keypoints. Two aspects are tightly linked to this notion, with respectively (i) key point detectors based for instance on Harris cover detection [51] or on SIFT keypoints (scale-invariant feature transform) [52] and (ii) key point descriptors that allow to describe image content around the keypoint [53][54]. This discriminates between methods relying only on the presence of characteristic points and dense methods that analyze all points in the image. For point description many solutions have been proposed that cannot all be mentioned here. Among them, one can cite shape context [55], SIFT and all their variations, HOG (Histogram of oriented gradients) [56], [57], LBP (Local binary patterns) [58], [59] [60] which all serve the same global purpose. Features can be used in themself or some histogram structures can be associated in spatial domains.

So far, we have presented the features that are studied in the literature. To improve results, researchers have tried to use those features in an optimal way. But as their primary goal is to obtain as much information as possible, there is a tendency to accumulate characteristics in large dimensional vectors, which may not be enough. In fact, this appears to be not so efficient in most occasions. Indeed, in a representation space with a dimension far too large, the famous curse of dimensionality holds and makes computation and comparison unrealistic. Besides, each characteristic may bring noisy information, some of which being more noisy than others, some others being redundant and making the process run longer than necessary. The aim of any system is to select the

most relevant features and the most diverse ones. A selection process reduces the number of efficient features and makes the system more robust: many approaches have been proposed for solving this question, such as filter, wrapper and hybrid procedures, which are among the most common methods. Wrapper methods are more efficient but are generally too time consuming; hybrid methods tend to optimize the efficiency and the computation work. For these approaches, the characteristics are not modified [61] [62]. Another approach is to try to define better features than the initial ones, based on a combination of the initially chosen features. This is the case when considering a Principal Component Analysis (PCA), trying to decrease the redundancy inherent to hancraft features or Independant Component Analysis (ICA). At the same time, it is possible to limit the number of features without damaging the result accuracy as the properties of the new representation axis are ordered. However, the new features are more difficult to understand since they correspond to a linear combination of the initial ones.

4. Knowledge or Learning

Features can be considered as an extraction of information from which decision making has to be achieved. We have considered the two main approaches that have been used in document analysis applications. Some are based on implicit or explict knowledge. The others applications are based on a learning phase that is supposed more objective.

4.1. Knowledge-based approach

Since the 1950s, document understanding has gone through multiple cycles alternating between knowledge-based approaches and learning-based approaches. Originally, systems were tightly linked to the expert, the knowledge of whom was modelled on the basis of more or less explicit formalisms: This knowledge could be introduced through parameters of the system (thresholds, rules, ...) or through theoretical formalisms such as semantic networks, graphs, grammars, and more recently ontologies. For all those systems, the document

analysis was often based on a sequential ordering of processes aiming at extracting the information from a document, so that the process could fit with the expert knowledge. Often, these approaches rely on an attempt to model the human visual system. Among the most famous systems, is ANON system proposed by Joseph in 1992 [63]. These context modellings suppose that the physical representations of a document are part of a visual language that has a reasonably formal grammar associated with it. The assumption is that the meaning of the message is fully embedded in the document and that the knowledge of the language in which it was expressed is sufficient to recover the meaning of the document. In the graphic recognition context, this is usually considered as the core domain of graphical document analysis. For these systems, the knowledge is implicit and used in an internal way often introduced as heuristics (in the algorithms) or externally (outside of algorithms), generating systems that are more or less generic. The external knowledge generally relies on knowledge representation methods. These approaches use many knowledge representation formalisms, such as rules, graphs, frames, semantic networks [64], graph based representation [65] and more recently ontologies [66] [67]. A number of other alternatives can be found in the literature (languages, databases, algebra, list, matrix, and so on). Also, other studies look into the neuroscience analysis, based on oculomotor studies or other models. Here Marr and Gestalt theory can be cited [68]. This was the foundation of original studies such as Ogier in 2000 [69]. Ogier proposed a complete cadastral map interpretation system based on a perception cycle principle. In the same way, one can also find pure rule based systems that use the advances in artificial intelligence , which are modelling expert knowledge through agents that are able to decide through argumentation and rounds of dialogues [70]. All these knowledge-based systems are involving an explicit learning. Whatever the systems may be, there are explicit and implict parameters involved, the most obvious ones being thresholds that are fixed and tuned according to experiments, so that the system accuracy is maximized.

4.2. Deep Learning

Totally different from knowledge-based approaches are learning-based approaches. Their principle consists in extracting the knowledge from some annotated datasets (such as annotated images for instance). These data are introduced as inputs for a training system. Here, the aim is not to recall the principle of deep learning nor to give an overview of the development studies defining the different architectures in terms of layers number. The objective is neither to present their family or the different objective functions that were proposed in the literature. In past times, artificial networks (ANN) were introduced [71] but became really popular but only after theoretical and practical methods to train networks, mostly based on the backpropagation algorithm designed by Y. Lecun, had been developed [72] or [73]. In the following years, Lecun introduced the convolutional networks (CNN) [74]. In 1990, he already used a five-layer network but at that time, it was not yet called a deep structure. The term deep network emerged in the noughties in the field of character recognition [75]. The complexity of the networks architecture systems is constantly increasing and with this complexity, the number of parameters and hyper parameters that have to be tuned during a learning phase. This implies to gather a large amount of annotated data according to the problem to be solved. The availability of the ImageNet [76] database released by Stanford Vision Lab from 2009 and its regular updating process has made developments in the field possible. ImageNet contains more than 14 million images with manual annotations consistent with WordNet hierarchy. Then it is a very large image database and it enables a real training of the networks. Since 2010, an ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is organized each year. The tasks are evolving but are always linked to classification of images and of objects in those images. In 2012, for the first time, deep neural networks were used by the participants. If the competition winner optimized 60 million of parameters [77], the old-style methods were still competitive [78]. From 2013 until today, most participants have used deep convolutional neural networks with architecture which are increasingly complex. In 2017 all participants used deep convolutional neural

networks. As we write these lines, several deep CNN architectures have become popular. Such architectures may either be modified and integrated into various applications or embedded in much larger architectures. Those deep CNN architectures are widely used currently, as shown in the studies presented at the ICDAR2017 conference. The design of the process has become the important thing. Some CNN or ANN are designed involving several layers for a specific application. When architectures are new, they usually increase the complexity of the wohle system by combining some basic classical architectures. The most famous deep current classical architectures have been named as ResNet [79], ResNet 101 [80], GoogLeNet [81], AlexNet [77], VGGNet [82], HCCR-CNN12Layer [83], PHOCNet [84], YOLOv2 [85], CaffeNet [86], LocNet [87], DeconvNet [88], DictNet[89], LeNet [90] or ConvNet [91], PixelCNN [92].

The term "deep learning" is associated with quite different studies. The global philosophy for using deep learning or deep approaches, is similar to that which prevailed with original ANN. Any multivariate function can be approximated by an ANN where the inputs are the variables and the outputs take the desired values. The number of parameters depends on the function to be approximated. The task of the architecture of neural networks is to provide the possibility to learn the function through available data. The input layer receives the image raw pixel values. Although the early development of ANN has been hindered by the lack of efficient learning algorithms, the number of hidden layers in ANN is in longer limited today because the back-propagation learning mode is still applicable. The explosion of computation time has been overcome by the availability of GPU that enables accelerating the learning process in a reasonable time. Within this context, the number of parameters to learn is tremendous and the number of labeled data involved in the learning process must be high, generally more important than what is available. This means researchers have to imagine strategies in order to artificially increase the number of annotated images in an automatic way. For this, they introduce transforms that they apply to the images available from the database. The choice of the transforms requires know-how in the field.

Because of these problems, the need for new strategies, the knowledge to be acquired were essentially problematic in the field of document analysis, alternative ways of using deep neural networks had to be explored. Today, mainly adaptations of already existing systems are performed or very basic architectures of deep methods are used. The approaches fall into two main categories: methods relying on transfer learning at different levels and proposal of methods involving the combination of one or several deep structures combined in complex architectures. Among the first category, the extraction of features that have proven to be efficient in some different problems dealing with image processing is the most frequent. The initial deep system, although it has been trained in an end to end fashion for solving another problem, has all the parameters already fixed. In fact, the system is seen as being made of two independent parts. The first part comprises of convolutional layers. It is considered as a feature extraction part whereas the second part corresponds to a classification part. The features that have been tuned for another problem are considered as "good" features, they are defined by the fixed weights in the left part. So this left part of the system is no longer modified. Besides, depending on the application, the classification part can take two aspects. Either the second part architecture is kept and parameters are learnt on the dataset available in the actual application or the second part is switched to another decision structure, for example a simple SVM. In this case, a learning step must be applied, as illustrated in Figure 2.

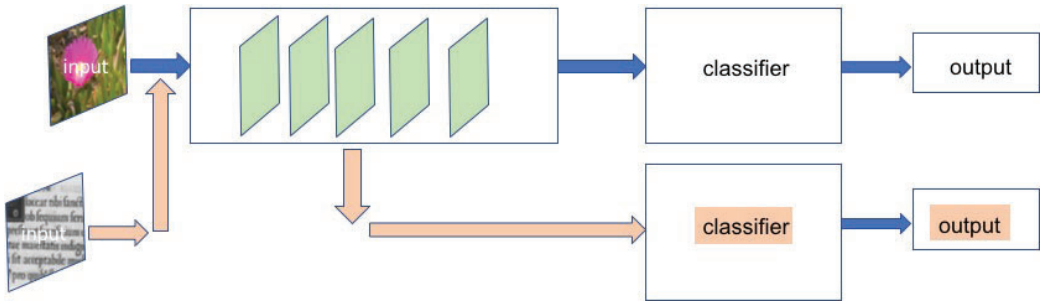

Figure 2: The deep learning method uses features from a deep CNN. The initial system is on the first line in green and in brown is the new architecture trained on a new problem.

Then come the architectures with a large number of layers used successfully

in previous applications. Further more, not only is the architecture used but the weights fixed in a previous application serve as a foundation to initialize the new system to be trained. The learning is further processed with the available data specific to the actual problem to be solved. Thus, the weights are modified through some rounds of optimization of the new system. Then, the problem of over fitting is emphasized as the quantity of data is usually limited. The principle is illustrated in Figure 3.

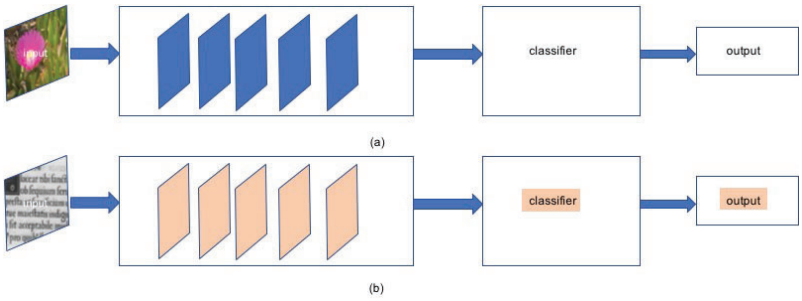

Figure 3: A transfer learning process: in (a) a pre-trained method and in (b) the use of the same architecture and same initial weights to train on new samples

The second category comprises of the proposal of a global system built from the combination of different subsystems, some elements of which are relying on deep methods. The main property of those systems and what makes them efficient is their ability to perform the learning of parameters or hyper parameters in an end to end process. The variety of global architectures is what differentiate these systems.

5. The use of Deep Learning is evolving

The first studies based on deep neural network architectures, and specially on convolutional networks appeared quite a long time ago, if we refer to the computer science time scale. The use of such models is more recent in the domain of document analysis and recognition. However, it is important to note the very rapidly-increasing use of such models. This is examplified by the examination of two international conferences (2016, 2017).

According to the proceedings of the 2016 International Conference on Pattern Recognition(ICPR2016), 61% of the total number of papers did not refer

to deep architecture for the document processing domain alone. At the ICDAR 2017 (International Conference on Document Analysis and Recognition), this number was down to 34%. This means that most papers referred to deep architecture. This evolution is visible in all the application domains but, as we mentioned in our previous section, the deep concept is used in different ways according to the studies. Furthermore, deep concepts are used but we must mention that they are not developed in the same way since they are presented in conferences such as Neural Information Processing Systems (NIPS) or Computer Vision and Pattern Recognition (CVPR). Only 19% of the papers referring to deep systems presented at the ICDAR 2017 mentionned the architecture of deep features, which are automatically fixed through the learning phase taking place to solve a problem different from the one raised at the onset.The most typical example for such a context is the use of the ImageNet database instead of documents. The trend is to see in these features some objectivity with respect to the traditional features that were chosen according to the system developer's expertise. To benefit from the different levels of computation performed in the network, deep-features are extracted at different levels of the convolutional layers. It seems that the selection of these levels is handcrafted but not engineered during the learning phase. It is obvious that the process increases the quality of the results in a significant way. In those studies, one trend to still increase the evaluation rates, is to multiply the inputs to the CNN computing deep-features. For example, a gray level image and the binary image are also fed to the network. Then, if the features are no longer handcrafted, some handcrafted transforms are applied to the material to be studied. There is no machine learning in the process because the transforms do not belong to a family of transforms, they are not obtained varying a hyper parameter but they are here to bring some complementary information difficult to compute in a finite CNN (however deep) in which first layers are linear convolutional layers. The limitation of the linear character of the convolutional operations is one of the problems that are considered in the community of the deep theoretical researchers.

Many studies attempt to achieve some transfer learning. They use some

architectures that happen to give some good results in "similar" applications and retrain them, either from scratch or using as initialization the weights obtained in a training phase of a previous application. This trend was confirmed in 19% of the papers presented at ICDAR 2017. It seems that in most cases several known architectures were chosen, but in the end the researchers' choice was essentially empirical. While in the early days, our community's choice went for either Zernike or Hu decomposition, now it is between Alexnet or VGG19 or any other architecture that we listed in previous section 4.2.

Finally, some other studies propose new architectures involving several deep structures or modifying the architecture of an already used network. Nowadays, these choices are only justified by the quality of their results on the specific applications. Based on these observations, we may safely state that some handcraft involvements remain present, but at a different level.

Comparing the content of ICPR2016 and ICDAR2017 theroretical papers provided an overview of evaluations on various applications.The overall impression was that the quality of the results is improving, even though it proves difficult to objectively appreciate those improvements. Then comparing results in the framework of competitions may make more sense. We considered three different tasks: i) classifying medieval writing ii) extracting layout and iii) binarizing document image. In these three cases, competitions were maintained for several years on equivalent testing sets and similar competitions, which allowed for a more objective assessment of the evolution of the community.

First, let us compare the competition on the Classification of Medieval Handwritings in Latin Script that took place at the ICFHR2016 (International Conference on Frontiers of Handwriting Recognition) [93] and ICDAR 2017 [94]. For each competition, seven and six systems were proposed respectively. In 2016, three of the seven systems relied on deep structures whereas the ratio was reversed in 2017 as four out of the six systems were relying on deep structures. Even though several deep systems were proposed, the winner in 2016 based the system on some local handcrafted features. As can be seen in Figure 4 (left handside), there is little difference between the winner and the second partici-

pant's results. The following year, the winner used a deep system based on the enhancement of the previous deep system, which improved its results. However, it is important to note that the second best competitor is still an "old" fashion system. We observe that the difference between the two best results has increased, as illustrated in Figure 4 (central part). For each deep system used in the competition, a reference deep system was used and empirically modified to adapt it to the data. In one case, the architecture of the network was chosen from some paper in arXiv, a recent but non-reviewed paper. In 2017, the competition comprised of a second task involving the classification of some images which were more difficult to process than those involved in the first task. The image contents were not as much homogeneous as those proposed in the first task but the training data was the same. Figure 4 (right handside) illustrates the accuracy loss between the two datasets. One possible conclusion is that the deep systems have learnt too much.They overfit the learning database and lack generalization capability. This is obvious when one compares the outcomes obtained from the use of features built by experts on handwriting examination. Such features, for instance, model the materials from medieval documents scrutinized by paleographers.

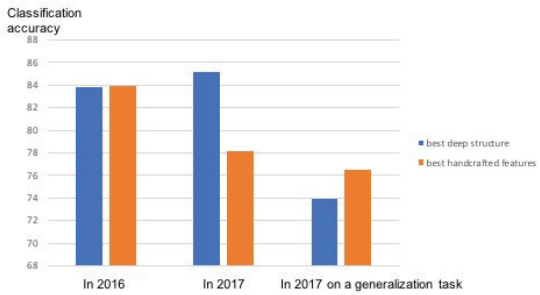

Figure 4: comparison of the results obtained in Clamm competitions in recent years

The layout competition on recognition of document layout is an older competition that has been open in all ICDAR conferences since 2001. But, in 2017 [95], no participant relied on a deep structure but used a multi-step procedure on rule-based decisions instead. Nevertheless, in the conference, one paper [96] segments document images to extract images and relies on a deep architecture.

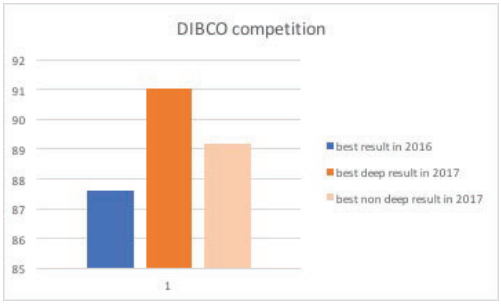

Figure 5: comparison of the results obtained in DIBCO competitions in recent years

The last competition we studied are the DIBCO competitions that have been organized in recent years. However we only considered the last two competitions on either printed or handwritten historical documents. In the 2016 competition [97] 12 systems were proposed and none of them relied on a deep method. In 2017, the organizers [98] received 26 participant systems. Seven of them only relied on a deep method. These belong to the machine learning methods whereas the other methods were more traditional. Results show that the seven methods were ranked in the top eight results. The first unsupervised method is ranked seven. This is illustrated in Figure 5.

In this section, we report the emerging of deep methods and the high quality of their associated results. All applications follow this trend, even though they can be used in various ways. These trends bring hope for the future.

6. Conclusion

After synthesing and examing the various steps of the community evolution, we may conclude that deep methods must not be opposed to other methods, and more specifically to heavily handcraft methods, for which everything is chosen by the system designer.

Hardware evolution, year after year, has led to increase both memory size and computation speed. The CPUs are more efficient and now GPU are necessary to handle the deep architecture working with very large databases. Then, from the analysis of the results, we can see a good synergy between hard and soft advances and this is good news for future developments. Besides, if human expertise was once turned towards the finding of features, it is obvious now that

it has to turn to higher levels in systems design. The level of abstraction has increased. From initial values, the objects of interest are now more like operators and sets of operators. The functions are replaced by families of functions and parameters give way to hyper-parameters. However human expertise is still needed. Researchers must not worry about losing their jobs as there is much to be discovered. Databases cannot yet handle all the natural or man-made variability and human will always be intelligent enough to propose cases that mislead the machine as it is the case in generative adversarial machine learning. Besides, tricks have to be found to adapt systems to low energy, to embed them in systems. Few works are tackling the problems of distributed resources. These points open up a large set of fascinating questions for future research.

References

[1] S. Ramdane, B. Taconet, A. Zahour, Classification of forms with handwritten fields by planar hidden markov models, Pattern Recognition 36 (4) (2003) 1045–1060.

[2] N. Gorski, V. Anisimov, E. Augustin, O. Baret, D. Price, J.-C. Simon, A2ia check reader: a family of bank check recognition systems, in: Fifth International Conference on Document Analysis and Recognition, 1999, pp. 523–526.

[3] M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, J. Lladós, Multimodal page classification in administrative document image streams, IJDAR 17 (4) (2014) 331–341.

[4] Y. Y. Tang, S.-W. Lee, C. Y. Suen, Automatic document processing: A survey, Pattern Recognition 29 (12) (1996) 1931 – 1952.

[5] M. Christy, A. Gupta, E. Grumbach, L. Mandell, R. Furuta, R. Gutierrez-Osuna, Mass digitization of early modern texts with optical character recognition, J. Comput. Cult. Herit. 11 (1) (2017) 6:1–6:25.

[6] S. He, L. Schomaker, Beyond ocr: Multi-faceted understanding of handwritten document characteristics, Pattern Recognition 63 (2017) 321 – 333.

[7] H. Alhéritière, F. Cloppet, C. Kurtz, J. Ogier, N. Vincent, A document straight line based segmentation for complex layout extraction, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017, 2017, pp. 1126–1131.

[8] A. Gordo, F. Perronnin, E. Valveny, Large-scale document image retrieval and classification with runlength histograms and binary embeddings, Pattern Recognition 46 (7) (2013) 1898–1905.

[9] D. B. Lysak, P. M. Devaux, R. Kasturi, View labeling for automated interpretation of engineering drawings, Pattern Recognition 28 (3) (1995) 393 – 407.

[10] K. Khurshid, C. Faure, N. Vincent, Word spotting in historical printed documents using shape and sequence comparisons, Pattern Recognition 45 (7) (2012) 2598–2609.

[11] A. P. Giotis, G. Sfikas, B. Gatos, C. Nikou, A survey of document image word spotting techniques, Pattern Recognition 68 (2017) 310 – 332.

[12] G. Kumar, V. Govindaraju, Bayesian background models for keyword spotting in handwritten documents, Pattern Recognition 64 (2017) 84 – 91.

[13] I. Siddiqi, N. Vincent, Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features, Pattern Recognition 43 (11) (2010) 3853–3865.

[14] S. He, L. Schomaker, Writer identification using curvature-free features, Pattern Recognition 63 (2017) 451 – 464.

[15] H. Freeman, Computer processing of line-drawing images, ACM Comput. Surv. 6 (1) (1974) 57–97.

[16] R. Stefanelli, A. Rosenfeld, Some parallel thinning algorithms for digital pictures, J. ACM 18 (2) (1971) 255–264.

[17] M. K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory 8 (1962) 179–187.

[18] I. Rothe, H. Susse, K. Voss, The method of normalization to determine invariants, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (4) (1996) 366–379.

[19] M. Teague, Image analysis via the general theory of moments, Journal of Optical Society of America (JOSA) 70 (1980) 920–930.

[20] A. Khotanzad, Y. H. Hong, Rotation invariant image recognition using features selected via a systematic method, Pattern Recognition (PR) 23 (1990) 1089–1101.

[21] A. Khotanzad, Y. H. Hong, Invariant image recognition by Zernike moment, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (5) (1990) 489–497.

[22] S. Liao, M. Pawlak, On the accuracy of Zernike moments for image analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (12) (1998) 1358–1364.

[23] H. tao Hu, Y. dong Zhang, C. Shao, Q. Ju, Orthogonal moments based on exponent functions: Exponent-fourier moments, Pattern Recognition 47 (8) (2014) 2596 – 2606.

[24] B. Bamieh, R. De Figueiredo, A general moment-invariant/attributed graph method for three dimensional object recognition for a single image, IEEE Journal of Robotics Automation 2 (1986) 31–41.

[25] Y. Chen, N. Langrana, A. Das, Perfecting vectorized mechanical drawings, Computer Vision and Image Understanding 63 (2) (1996) 273–286.

[26] C. Teh, R. Chin, On image analysis by method of moments, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (4) (1988) 493–513.

[27] S. Belkasim, M. Shridar, A. Ahmadi, Pattern recognition with moment invariants : A comprative study and new results, Pattern Recognition 24 (1991) 1117–1138.

[28] S. Pei, C. Lin, Normalisation of rotationally symetric shapes for pattern recognition, Pattern Recognition 25 (1992) 913–920.

[29] C. Lin, New forms of shape invariants from elliptic fourier descriptors, Pattern Recognition 20 (1987) 535–545.

[30] T. Taxt, Recognition of handwritten symbols, Pattern Recognition 23 (11) (1990) 1155 – 1166.

[31] F. Kuhl, C. Giardina, Elliptic fourier features of closed contour, Computer Vision, Graphics and Image Processing 18 (1982) 236–258.

[32] O. Trier, T. Taxt, A. Jain, Extraction methods for character recognition a survey, Pattern Recognition 29 (1996) 641–662.

[33] S. Shimotsuji, O. Hori, M. Asano, Robust drawing recognition based on based-guided segmentation, in: in Proc. of IAPR Workshop on Document Analysis Systems, DAS 1994, Kaiserslautern, Germany, October 23-26, 1994, 1994, pp. 337–348.

[34] K. Y. S. Mori, C.Y.Suen, Historical review of ocr research and development, Proceedings of the IEEE 80 (7) (1992) 1029–1058.

[35] A. C. Shaw, Parsing of graph-representable pictures, J. ACM 17 (3) (1970) 453–481.

[36] K. Riesen, H. Bunke, Graph classification based on vector space embedding, International Journal on Pattern Recognition and Artificial Intelligence 23 (6) (2009) 1053–1081.

[37] Recent advances in graph-based pattern recognition with applications in document analysis, Pattern Recognition 44 (5) (2011) 1057 – 1067.

[38] A. Fischer, K. Riesen, H. Bunke, Improved quadratic time approximation of graph edit distance by combining Hausdorff matching and greedy assignment, Pattern Recognition Letters 87 (2017) 55–62.

[39] I. Daubechies, the wavelet transform, time-frequency localization and signal analysis, IEEE Transaction on Information Theory 36 (1990) 961–1005.

[40] I. Daubechies, Ten lectures on wavelets, in: CMBS-NSF, Regional conference series in applied mathematics, 1992.

[41] S. Deng, S. Latifi, E. Regentova, Document segmentation using polynomial spline wavelets, Pattern Recognition 34 (12) (2001) 2533 – 2545.

[42] D. Shen, H. H. Ip, Discriminative wavelet shape descriptors for recognition of 2-d patterns, Pattern Recognition 32 (2) (1999) 151 – 165.

[43] G. Chen, T. D. Bui, Invariant Fourier-Wavelet descriptors for pattern recognition, Pattern Recognition (PR) 32 (1999) 1083–1088.

[44] M. Suryani, E. Paulus, S. Hadi, U. A. Darsa, J.-C. Burie, The handwritten sundanese palm leaf manuscript dataset from 15th century, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Osaka, Japan, November 13-15, 2017, 2017, pp. 796–801.

[45] H. Turki, M. B. Halima, A. M. Alimi, Text detection based on mser and cnn features, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Osaka, Japan, November 13-15, 2017, 2017, pp. 949–954.

[46] W. Rudin, Real and Complex Analysis, McGraw-Hill, 1987.

[47] A. Ghorbel, J. Ogier, N. Vincent, A segmentation free word spotting for handwritten documents, in: 13th International Conference on Document

Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015, 2015, pp. 346–350.

[48] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, MIT Press, 1982.

[49] D. Gabor, Theory of communication, J. Inst. of Elect. Eng. Part III, Radio and Communication 93 (1946) 429–457.

[50] N. Sharma, S. Chanda, U. Pal, M. Blumenstein, Word-wise script identification from video frames, in: 13th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, USA, August 25-28, 2013, 2013, pp. 867–871.

[51] C. Harris, M. Stephens, A combined corner and edge detector, in: 4th Alvey Vision Conference, 1988, pp. 147–151.

[52] D. G. Lowe, Object recognition from local scale-invariant features, in: International Conference on Computer Vision, 1999.

[53] N. Nguyen, C. Rigaud, J.-C. Burie, Comic characters detection using deep learning, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Osaka, Japan, November 13-15, 2017, 2017, pp. 41–46.

[54] D. Sharma, N. Gupta, C. Chattopadhyay, S. Mehta, Daniel: A deep architecture for automatic analysis and retrieval of building floor plans, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Osaka, Japan, November 13-15, 2017, 2017, pp. 421–426.

[55] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–521.

[56] M. Hangarge, C. Veershetty, R. Pardeshi, G. Mukarambi, Word retrieval from kannada document images using HOG and morphological features,

in: Recent Trends in Image Processing and Pattern Recognition - First International Conference, RTIP2R 2016, Bidar, India, December 16-17, 2016, Revised Selected Papers, 2016, pp. 71–79.

[57] A. Berenguel, O. R. Terrades, J. Llados, C. Canero, e-counterfeit: a mobile-server platform for document counterfeit detection, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Osaka, Japan, November 13-15, 2017, 2017, pp. 15–20.

[58] S. Dey, A. Nicolaou, J. Lladós, U. Pal, Local binary pattern for word spotting in handwritten historical document, in: Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2016, Mérida, Mexico, November 29 - December 2, 2016, Proceedings, 2016, pp. 574–583.

[59] H. Mohammed, V. Margner, T. Konidaris, H. S. Stiehl, Normalised local nave bayes nearest-neighbour classifier for offline writer identification, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Osaka, Japan, November 13-15, 2017, 2017, pp. 1013–1018.

[60] S. Shan, H. Xu, F. Su, A new method for spatiotemporal textual saliency detection in video, in: 23rd International Conference on Pattern Recognition, ICPR 2016, Cancun, Mexico, December 4-8, 2016, 2016, pp. 3229–3234.

[61] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, Meta-des.oracle: Meta-learning and feature selection for dynamic ensemble selection, Information Fusion 38 (2017) 84–103.

[62] M. Reif, F. Shafait, Efficient feature size reduction via predictive forward selection, Pattern Recognition 47 (4) (2014) 1664 – 1673.

[63] S. Joseph, P. Pridmore, Knowledge-directed interpretation of line drawing images, IEEE Trans. on PAMI 14 (1992) 928–940.

[64] S. D. Antoine, J. S. Collin, J. K. Tombre, Analysis of technical documents: The redraw system, Structured Document Image Analysis, Springer Verlag, Berlin/Heidelberg (2002) 385–402.

[65] J. Llados, E. Marti, J. Villanueva, Symbol recognition by error-tolerant subgraph matching between region adjacency graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (10).

[66] S. Belongie, J. Malik, J. Puzicha, Towards ontology-based retrieval of historical images. applied ontology, Applied Ontology 10 (2) (2015) 147–167.

[67] S. Ebert, M. Liwicki, A. Dengel, Ontology-based information extraction from handwritten documents, in: Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, 16-18 November 2010, Kolkata, India, 2010.

[68] A. Desolneux, L. Moisan, J. Morel, Gestalt theory and computer vision, in: Seeing, Thinking and Knowing, Kluwer, 2002, pp. 71–101.

[69] J. Ogier, R. Mullot, J. Labiche, Y. Lecourtier, Semantic coherency: the basis of an image interpretation device-application to the cadastral map interpretation, IEEE Trans. Systems, Man, and Cybernetics, Part B 30 (2) (2000) 322–338.

[70] F. Cloppet, P. Moraitis, N. Vincent, An agent-based system for printed/handwritten text discrimination, in: PRIMA 2017: Principles and Practice of Multi-Agent Systems - 20th International Conference, Nice, France, October 30 - November 3, 2017, Proceedings, 2017, pp. 180–197.

[71] W. McCulloch, W. Pitts, A logical calculus of ideas immanent in nervous activity, Bulletin of Mathematical Biophysics 5 (4) (1943) 115 – 133.

[72] Y. LeCun, Learning processes in an asymmetric threshold network, in: E. Bienenstock, F. Fogelman-Soulié, G. Weisbuch (Eds.), Disordered systems and biological organization, Springer-Verlag, Les Houches, France, 1986, pp. 233–240.

[73] S. Becker, Y. LeCun, Improving the convergence of back-propagation learning with second-order methods, in: D. Touretzky, G. Hinton, T. Sejnowski (Eds.), Proc. of the 1988 Connectionist Models Summer School, Morgan Kaufman, San Mateo, 1989, pp. 29–37.

[74] Y. LeCun, O. Matan, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, H. S. Baird, Handwritten zip code recognition with multilayer networks, in: IAPR (Ed.), Proc. of the International Conference on Pattern Recognition, Vol. II, IEEE, Atlantic City, 1990, pp. 35–40.

[75] M. Ranzato, F. Huang, Y. Boureau, Y. LeCun, Unsupervised learning of invariant feature hierarchies with applications to object recognition, in: Proc. Computer Vision and Pattern Recognition Conference (CVPR'07), IEEE Press, 2007.

[76] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proc. Computer Vision and Pattern Recognition Conference (CVPR2009), IEEE Press, 2009.

[77] A. Krizhevsky, L. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25 (4).

[78] T. Harada, Y. Kuniyoshi, Graphical gaussian vector for image categorization, in: Advances in Neural Information Processing Systems (NIPS 2012), 2012.

[79] S. Ren, K. He, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2017) 1137 – 1149.

[80] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. Computer Vision and Pattern Recognition Conference (CVPR2016), IEEE Press, 2016.

[81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. Computer Vision and Pattern Recognition Conference (CVPR2015), IEEE Press, 2015.

[82] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. International Conference on Learning Representation (ICLR2014), IEEE Press, 2014.

[83] X. Xiao, L. Jin, Y. Yang, W. Yang, J. Sun, T. Chang, Building fast and compact convolutional neural networks for offline handwritten chinese character recognition, Pattern Recognition 72 (2017) 72 – 81.

[84] S. Sudholt, G. Fink, PHOCNet : A deep convolutional neural network for word spotting in handwritten documents, in: Proc. International Conference on frontiers of Handwriting Recognition (ICFHR2016), IEEE Press, 2016.

[85] J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in: Proc. Computer Vision and Pattern Recognition Conference (CVPR'17), IEEE Press, 2017.

[86] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 675 – 678.

[87] S. Gidaris, N. Komodakis, Locnet: Improving localization accuracy for object detection, in: Proc. Computer Vision and Pattern Recognition Conference (CVPR2016), IEEE Press, 2016, pp. 789 – 798.

[88] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the International Conference on Computer Vision, 2015.

[89] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Synthetic data and artificial neural networks for natural scene text recognition, in: NIPS Deep Learning Workshop, 2014.

[90] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[91] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, CoRR abs/1311.2901.

[92] A. Bansal, X. Chen, B. Russell, A. Gupta, D. Ramanan, Pixelnet: Representation of the pixels, by the pixels, and for the pixels, arXiv:1702.06506.

[93] F. Cloppet, V. Eglin, V. Kieu, D. Stutzmann, N. Vincent, ICFHR2016 competition on the classification of medieval handwritings in latin script, in: Proc. International Conference on frontiers of Handwriting Recognition (ICFHR2016), IEEE Press, 2016.

[94] F. Cloppet, V. Eglin, M. Helias-Baron, V. Kieu, D. Stutzmann, N. Vincent, ICDAR 2017 competition on the classification of medieval handwritings in latin script, in: Proc. International Conference on Document Analysis and Recognition (ICDAR2017), IEEE Press, 2017, pp. 1371–1376.

[95] C. Clausner, A. Antonacopoulos, S. Pletschacher, ICDAR2017 competition on recognition of documents with complex layouts RDCL2017, in: Proc. International Conference on Document Analysis and Recognition (ICDAR2017), IEEE Press, 2017.

[96] S. Tsutsui, D. Crandall, A data driven approach for compound figure separation using convolutional neural networks, in: Proc. International Conference on Document Analysis and Recognition (ICDAR2017), IEEE Press, 2017.

[97] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, Handwritten document image binarization contest (H-DIBCO 2016), in: Proc. International Confer-

ence on frontiers of Handwriting Recognition (ICFHR2016), IEEE Press, 2016.

[98] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, ICDAR2017 competition on document image binarization (DIBCO 2017), in: Proc. International Conference on Document Analysis and Recognition (ICDAR2017), IEEE Press, 2017.