Document Analysis

Jürgen Schürmann
Daimler-Benz Research Ulm, Germany

Abstract

The written or printed text document has over centuries evolved to become the most important carrier of information for communication among humans and for human comprehension. With the rapid development of modern information technology computerized systems are beginning to play their role as intelligent assistants capable of reading and understanding documents prepared for human use. Being intrinsically an image understanding task, but driven by its specific goals and requirements document analysis has taken a path of its own in the computer vision field. The paper describes what the specialities are compared to general image understanding, how document analysis is related to natural language processing and what the reasons are for document analysis to take recourse to higher levels of understanding. On these foundations the general architecture of document analysis systems is discussed with some special emphasis on pattern classification as the basic technique for character recognition. The main fields of application are shortly presented.

1. Introduction

The visual sense is certainly the most important sense of higher living systems. Visual perception gives us the necessary information about the world we are in and is the basis for almost any intelligent behavior and action. The human visual sense also became the cornerstone of cultural development. The invention of script and writing, the dissemination and conservation of knowledge, the whole system of science and civilization would never have evolved to its present form without the human capabilities of viewing and understanding written signs and symbols on materials such as clay-tablets, parchment and paper, and -- as the preliminary endpoint of this development -- on computer screens.

Since the emergence of light sensing devices and suitable electronics the problem was tackled to design electronic reading devices. In the historic development one of the driving forces has been the intention to construct reading aids for those who weren't able to read themselves, reading aids for the blind [1]. Since then, what started as character recognition, has developed into document analysis and document understanding and moved considerably from mere image analysis and computer vision into the realm of natural language processing.

Certainly, document analysis is deeply rooted in image analysis since documents are primarily images. But these images convey messages which are expressed by language.

Therefore, knowledge about language is enormously helpful if we try to extract the messages from the image, especially since it turned out that unambiguous recognition is not in every case possible on just the character level.

Document analysis is a subfield of computer vision and has taken its own path with specific goals and approaches and with separate conferences, but always in viewing distance to that of the computer vision community. I understand my invitation to this talk as the invitation to transmit an impression about what the specialities of document analysis are compared to computer vision in general, about the state of art in this field and about its present trends.

2. Document Analysis, a Subfield of Image Understanding

Compared to the challenges of image understanding in natural three-dimensional scenes, document analysis, at first glance, seems to be a rather simple task. Surely, the general approach is very similar: we have to identify objects in the scene, determine their describing features, give then names to these objects based on their feature vectors, determine the geometrical relations between these objects, and build up a hierarchy of higher level objects according to a model representing the application dependent knowledge.

Document analysis has merely to deal with flat two-dimensional images, predominantly of the simple black and white type -- since all what cares ultimately is the difference between ink and background. Document analysis need not answer such complicated questions as for position and orientation in six degrees of freedom with ample possibilities of mutual occlusion among objects from perspective views. Document analysis need not cope with the troublesome impact of lighting conditions on the images gained from natural scenes -- since the illumination normally is part of the scanner which provides the document raster image to the document analysis system.

Whereas in general computer vision the camera system usually is an integral part of the whole image understanding system, in document analysis scanning has become a separate technology with a well-defined interface to document analysis. Document analysis works with high-resolution raster-images which may come from arbitrary sources. Interestingly, in many application contexts handling of digitized images is so much more convenient than handling of paper documents that it gave rise to an ever growing amount of already scanned and electronically stored documents -- mostly on CD-Roms -- which are simply waiting for being brought to a document understanding system.

The huge volume of information kept in raster image format presently is accessible only to humans. Certainly, compared to sending paper copies around by conventional mail services, distribution of information via electronic media is much easier if the document is stored in raster image format. But, as soon as actions shall be derived from the document content we first have to bring it on the computer screen -- or print it -- and let a human read and understand the document. Thus, from the viewpoint of computerized information processing the growing document image databases are buried treasures -- treasures to be raised by document understanding systems. Besides these promises of future importance already today there are many economically important applications of document analysis technology. I will return later to this point.

One of the specific challenges of document analysis comes from the high requirements on image resolution. Characters are tiny objects and usually we bring thousands of them on one page. The standard A4 document scanned with standard 300 dpi resolution results in an 8 Mpixel document raster image. With RGB scanning the image needs 24 MBytes of storage which shrink to 1 MByte after converting the document image to ink and background representation. Common image data compression techniques don't really help since the image would have to be decompressed for analysis. The only exception is linewise run-length coding in binary images which itself represents a well-suited preprocessing operation for connected component analysis since connected component analysis is in most cases the very first image processing operation to be applied.

In certain applications high resolution requirements meet with the demand for high throughput. In postal address reading e.g. the typical address reading zone is about 3×8 square inches in size. The mailpieces are passing the scanning station with about 40000 letters/hour. 300 dpi grayvalue scanning results in about 24 MByte/second input data rate.

Document analysis has to deal with a number of concepts arranged in interwoven hierarchies. On the geometrical side we have image-oriented concepts such as connected components, writing lines, text lines, blocks, and geometrical neigborhood relations. On the language-oriented side we have the character -- alphabetical, numeral, interpunctation mark -- on the lowest level of a conceptual hierarchy, followed by words, well-formed sentences, sections, paragraphs and finally the whole story of the document. These language-oriented concepts again mix with geometrical layout-oriented concepts such as bold-ness and italic-ness of fonts and the concept of writing order.

The objects to be recognized may be subject to shift, scaling, skewing and rotational variations. This, however, is normally less variation than that caused by the affine transformations of perspective geometry. The character objects, however, may be distorted by the copying process, by sloppy handwriting and bad printing. One main source of variation is the use of countless different fonts not at all designed for easy discriminability but bearing lots of ambiguities which cannot be resolved on the lower levels of understanding.

It must be noted that the ambiguities among the different fonts are much more severe for automatic reading devices than for human use since in general the character recognizer has first a very limited view on the document and second is font-blind. The standard omnifont or polyfont recognition engine simply ignores the fact that there exists a multitude of different fonts and tries instead to read correctly any character found independent of the font it comes from.

Improvements in reading accuracy seem to be possible if font-dependent classifiers could be used, admittedly, at the cost of increased classifier complexity. But this requires reliable font recognition which is difficult to achieve from the narrow view the typical character recognizer has on its task. If, however, the recognizer can rely on the fact of large portions of text being printed in identical font, font recognition and font-specific classification is applicable.

Document analysis is not only recognition of documents written or printed in western (Latin) fonts. Large activities are also going on in recognition of numerous other alphabets like Greek, Arabian, Korean, Chinese, Japanese etc. where the number of classes to be discerned quickly can reach several thousands. Document analysis also has to deal with

the problem of multilingual text recognition where first the language is to be determined before the adequate language specific recognition engine system can be invoked.

Quite different recognition tasks arise from applications such as drawing and map interpretation or reading of musical notes. Besides the text parts, which also may appear in documents of this type, the primitive elements here are quite more complicated and obey a different language. In technical drawings a whole variety of geometrical objects can be constructed from strokes of different strokewidth and curvature. A specific difficulty in these applications derives from the fact that in contrast to character objects, which in most cases can be identified with connected components separated from each other, drawing objects are normally connected.

3. Relations to Natural Language Processing

Document images differ from general images in that they contain characters, words and text. Text is one of the two dominant media representing language, the other is speech. Language, however, is the object of linguistics, a science quite different from the engineering sciences. In view of the limited intelligence of present day reading devices it is not really necessary to become first an expert in computer linguistics before one can try to design a document analysis system. But earlier or later, everybody, working on document analysis will become involved in natural language processing to a certain degree.

The situation is rather similar to speech recognition where reliable recognition systems cannot be constructed without moving the recognition task from the phoneme level to higher levels of linguistic understanding such as the level of words or the level of dialog messages. In the same way in document analysis the ambiguities of recognition on the character level are generally overcome by recognizing words instead of characters. Techniques of this kind are commonly subsumed under the notion of contextual postprocessing.

There are many commonalities between speech and text recognition. Recognition of connected handwritten words e.g. has profited much from the technique of Hidden Markow Models developed for recognition of connected spoken words [2]. Indeed, much benefit and mutual fertilization can be earned, if research groups in speech understanding and text understanding are working closely together, since on a certain level of understanding speech and script understanding merge to language understanding.

4. Need for Recourse to Higher Level Understanding

The simple idea of translating the document image character by character into the corresponding sequence of ASCII-symbols does work only in exceptionally friendly cases. The common situation is that we have even difficulties to cleanly separate the single character images. Due to broken or merging character images the single character image is difficult to be determined. Things are even more complicated in connected handwriting recognition where begin and end of the character image can only be indicated after the whole word has been recognized.

The situation is made even worse by the fact that in certain cases identical graphemes are used for different character meanings, which is frequent among different fonts but happens also in the same font. If character recognition shall be made position and size independent we get problems with those character images which look the same or almost the same in lower and upper case. Thus, reliable character recognition requires recognition of writing lines as a precondition. From the image analysis point of view these writing lines are imaginary concepts. In general, writing lines can only be found if we have at hand a large enough number of characters with ascenders and descenders arranged in text lines.

The fact that recognition on the character level is often ambiguous is also valid for human recognition. The only way out of unreliability and ambiguity is recourse to higher level knowledge. That is the same kind of knowledge -- but up to much higher levels of understanding than is accessible to technical systems -- which is unconsciously used by humans in reading texts and understanding documents.

The level above the level of characters is the level of words. In principle, by exchanging the notion of a character class for the notion of a word class and designing class-membership functions for any word in the lexicon, we could switch over to the word as the elementary object of texts and recognize complete words instead of characters. The price we would have to pay for that is a tremendous increase in the number of classes since in common language the number of words is much larger than the number of characters.

Therefore, the word recognition approach can only work if the application allows to limit the size of the lexicon to a reasonable number. Indeed, there are numerous applications where this is the case.

Due to utilizing application specific constraints recognition on the word level in principle is more reliable than on the lower. There are different approaches possible to design recognition systems following this scheme. The direct way would be to determine the classes of the recognition task on the word level. The indirect way is to work with character recognizers and to equip them with the capability of forwarding multiple choices.

The character recognizer must observe its own behavior and determine when its decisions are unreliable and ambiguous. In ambiguous cases no longer unique decisions are forwarded but instead sets of alternatives and the final decision is thus left to the higher levels of the understanding system. The critical requirement is that the set of alternatives is large enough to contain the correct class and small enough not to overload the subsequent stages of the recognition system with the task of sieving out senseless alternatives.

Present-day document analysis systems make, as the key to reliable recognition, intensive use of application-specific knowledge. A good example is address reading. The address reader knowledge base represents the fact that certain expressions denote city names, others street names, that a certain city has certain streets, that along the streets are buildings which may have names of their own or simply numbers, that certain customers reside in certain buildings.

For more general applications document analysis systems are under development e.g. [3,4,5] having the architecture of knowledge-based systems. These systems contain a predefined body of knowledge about language, lexica and syntax -- and not to forget, knowledge about layout and typesetting -- together with the appropriate inferencing system.

This architecture constitutes a basic framework to be filled with application specific knowledge.

5. Architecture of Document Analysis Systems

The common document is represented by a high resolution raster image exhibiting a complex graphical structure of texts blocks, tables, inserted images and captions, graphical elements such as boxes, guiding lines etc. The first task is layout analysis and the basic tool applied for this purpose is connected components analysis [6,7] rendering a database of all graphical entities found in the image. This image content database stores the graphical entities together with their attributes, such as coordinates, bounding box, boundary polygon, inclusion hierarchy etc. The image content data base is complete. The original input image can be partly or entirely reconstructed from this database. Normally the input image can be disposed at this status of the procedure.

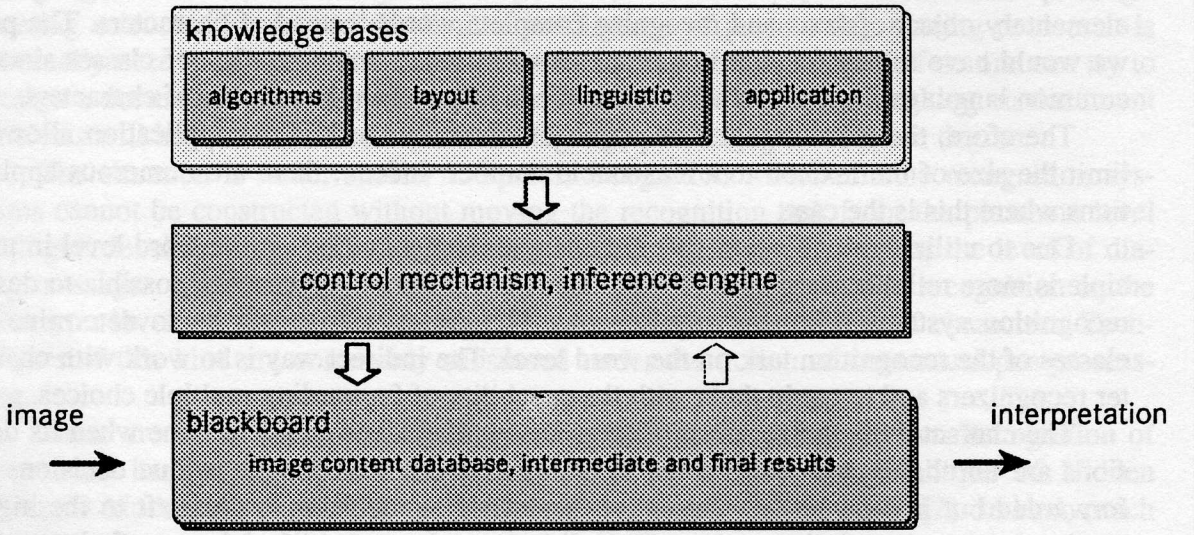

Fig.1: Architecture of the document analysis system

One fundamental advantage of changing the image representation from the raster image to the image content database in the context of document analysis is that in this form fast selections are possible. It is now very easy to select e.g. all objects with bounding box within a certain range, or all circular shaped objects within a certain range of size or all line-like objects not longer than a certain value and with a strokewidth in a given range. Another advantage is the scaling property of this representation with the pixel size. Whereas the number if pixels grows quadraticly with the number of pixel per inch, the complexity of the image content database grows much smaller.

Often, after finding the image objects certain image processing operations become necessary, e.g. size normalization, skew correction by shearing, rotations, morphological

filtering with different structuring elements, sceletonization. All of these operations can advantageously be implemented in the boundary polygon representation. In most cases remarkable gains in processing speed are gained at the cost of somewhat more complicated algorithms. By staying in the image content database representation the benefits of fast selections are kept.

Since the logical entities making up the document are not in every case separated by gaps filled with background pixels, the geometrical entities cannot in general directly be mapped into logical entities. Instead, in certain cases geometrical entities must be broken into pieces or merged. Examples are the determination of character images in case of characters consisting of more than one connected component or connected components consisting of more than one character.

Examples for logical entities principally consisting of larger numbers of geometrical entities are text lines, text blocks, columns etc. Since text is a linear sequence of characters the character images must be brought into reading order and specifically must be positioned into imaginary writing lines. There are cases where the character position in the writing line has a strong impact on the character meaning.

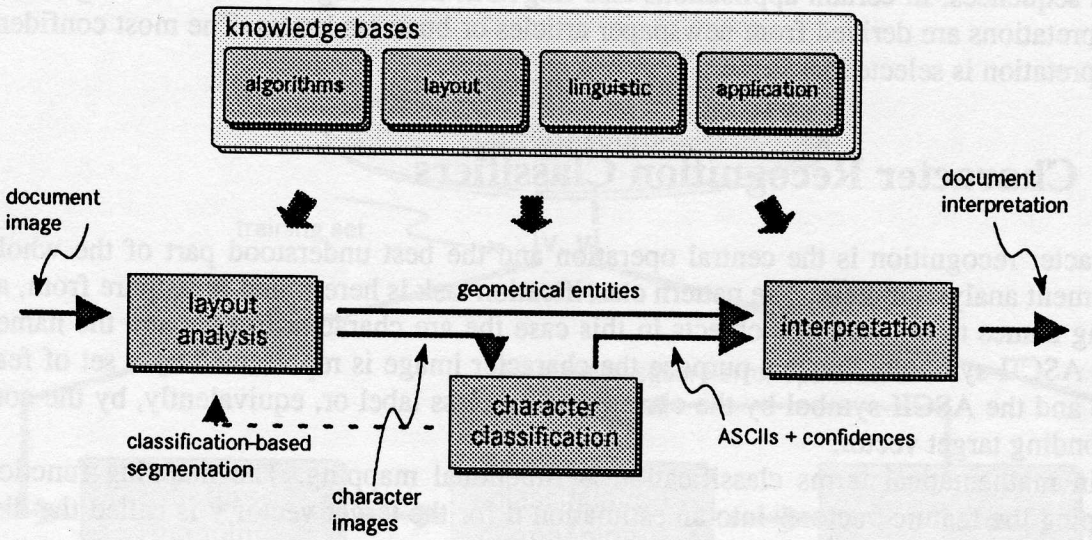

Fig.3: Processing Flow

The result of these operations is again stored in the document database which serves as the blackboard for a knowledge based reasoning process. Not in all cases these associations are final. In any case of ambiguities, alternatives are indicated and checked in later stages of the analysis process. A typical example is resolving character segmentation alternatives by invoking the character classifier. The finally accepted character segmentation is that leading to the overall interpretation with maximum confidence.

After the reading order has been established the character images are presented one after the other to the character classifier which maps them into the corresponding ASCII-

symbol. Text may be printed in almost countless different fonts, not to mention the multitude of variations in individual handwriting. The typical document analysis character recognizer ignores the font and tries to recognize the character meaning without knowing the font or even without knowing that different fonts exist at all. These systems are called polyfont or omnifont recognizers. Most of them also ignore size and slant variations by applying appropriate normalization operations before recognition.

Ambiguities on the character level are treated by forwarding sets of alternative character meanings together with their confidences. From these alternative words can be formed again weighted by confidences. In case of connected handwriting recognition the recognizer operates on whole word images instead of single character images. The usual technique is Hidden Markow Modelling again resulting in a sorted set of alternatives with their confidences.

The last step of the processing chain is document understanding based on a model world representing the application. On this level, reasoning operates with layout knowledge as well as with language knowledge. Layout reasoning takes positional information into account, language based reasoning cares for lexica describing which words are allowed in certain positions and for syntactical and semantical rules describing admissible word sequences. In certain applications also linguistic knowledge comes into the game, if interpretations are derived from newspaper articles or business letters. The most confident interpretation is selected.

6. Character Recognition Classifiers

Character recognition is the central operation and the best understood part of the whole document analysis system. The pattern classification task is here found in its pure from, as giving names to objects. The objects in this case the are character images and the names their ASCII-symbols. For this purpose the character image is represented by a set of features and the ASCII-symbol by the corresponding class label or, equivalently, by the corresponding target vector.

In mathematical terms classification is functional mapping. The mapping function mapping the feature vector **v** into an estimation **d** for the target vector **y** is called the discriminant function $\mathbf{d} = \mathbf{d}(\mathbf{v})$.

There are numerous ways possible of designing discriminant functions. All of them refer to collections of sample sets representing the relation between feature vector **v** and target vector **y** in the considered application. This is the paradigm of learning from examples and learning from examples is an intrinsically statistical concept.

From this point of view, the pattern to be recognized is a pair **[v,y]** of the two vector variables **v** and **y** describing how the pattern looks and what its correct meaning is. The application for which the character recognizer is to be developed is a stochastic source of pairs **{v,y}** -- the pattern source. The criterium for discriminant function design fundamentally is to make sure that the estimations **d** are valid in most of the cases -- minimum error rate optimization.

There is a universe of different constructions for discriminant functions imaginable. They differ in the determination of features **v**, in the basic construction of the discriminant

function **d**(.), in the optimization rule forcing the estimations **d** to come close to **y** as often or as closely as possible, and in the degree of human interaction needed in the learning process.

Pattern classifiers to be used in character recognition and document analysis systems are designed according to the specific requirements of the application in terms of set of classes, font mix, degradation of print quality. A consequence thereof is that classifier development based on learning and test sets here has become a routine job and is as far as possible automated. Therefore in this field, those feature determination and discrimination function design techniques dominate which require minimum human interaction and design ingenuity.

At the very beginning of all processing there are raw measurements -- the pixel data, be they binary, grayvalue or color measurements. These go through several stages of processing in order to make them better suited for the intended purpose of analysis and recognition. We come out with the pixel array of the character's bounding box, usually after size normalization, rotational, shear and strokewidth normalization. Many character recognizers use the pixel data **v** directly as input data for the pattern classifier. Others first transform them into a new set of variables **w** which then constitute the classifier input data.

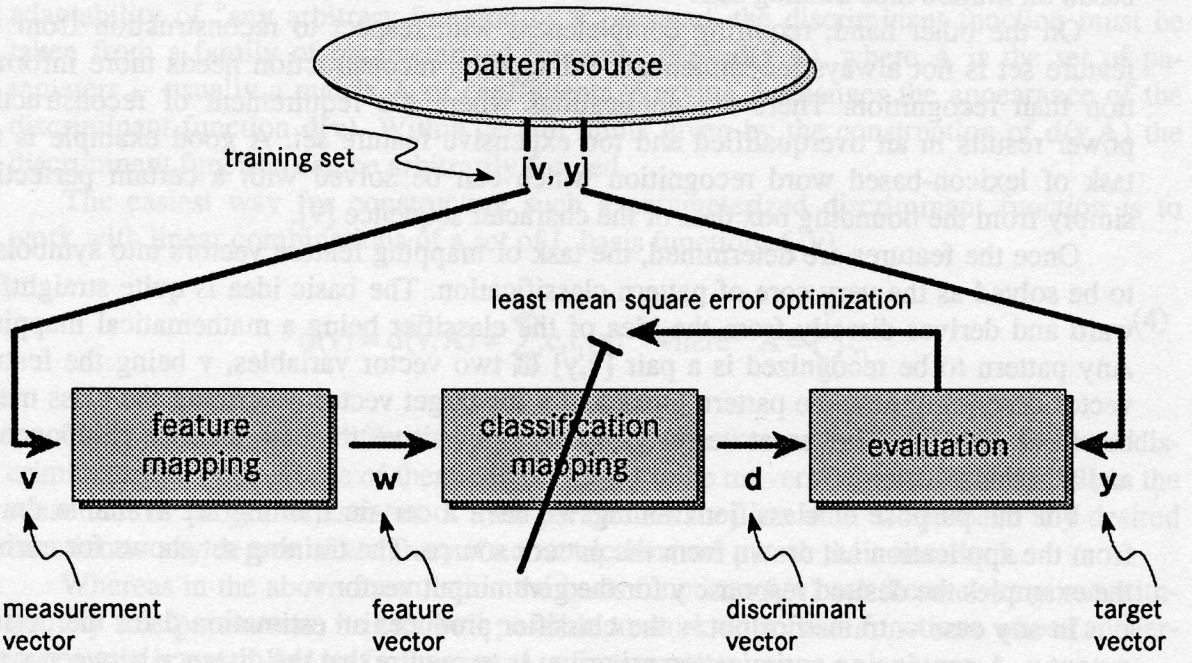

Fig.3: Basic structure of pattern classifiers

In mathematical terms, computing an improved feature vector **w** from the given feature vector **v** is simply a mapping from **V**-space to **W**-space just as the whole classification process is nothing else than a mapping from **V**-space to **D**-space. Thereby, **D** is the decision space with the embedded K target points **y** for the K classes to be discriminated

and **V** the input feature space. If the classifier works with the improved feature vector **w**, then its mapping goes from **W**-space to **D**-space.

From these considerations it is clear that arbitrary intermediate mappings can be inserted into the processing chain from **V**-space to **D**-space. A good criterion often, is to ask whether the secondary feature vector **w** has the capability of reconstructing from **w** the primary feature vector **v**. If that is the case, we are sure not to have lost relevant information while transforming the primary feature vector **v** into the improved feature vector **w**. Since the primary features may be correlated and therefore contain redundancy, an extremely interesting question is how many features are really needed.

These considerations match the design rationale of one of the most prominent feature mapping approaches, principal component analysis [8]. Applying this technique, the given coordinate system in feature space is replaced by a task dependent one consisting of the dominant eigenvectors computed from the covariance matrix of the pattern generating stochastic process {**v**}. Computationally, principal component analysis consists of solving an eigenvector/eigenvalue problem. Dominant eigenvectors are those corresponding to the largest eigenvalues. Using this technique the several-hundred-dimensional feature vector **v** normally representing the character raster image can be replaced by an improved feature vector **w** with about some ten dimensions. Reduction of the feature vector dimensionality without sacrifying discriminative power is one of the keys to reliable classifier training based on limited-size training sets.

On the other hand, requiring completeness with respect to reconstruction from the feature set is not always recommended. Principally, reconstruction needs more information than recognition. There are applications where the requirement of reconstructive power results in an overqualified and too expensive feature set. A good example is the task of lexicon-based word recognition which can be solved with a certain perfection simply from the bounding box data of the character sequence [9].

Once the features are determined, the task of mapping feature vectors into symbols is to be solved as the very core of pattern classification. The basic idea is quite straightforward and derives directly from the idea of the classifier being a mathematical mapping. Any pattern to be recognized is a pair [**v**,**y**] of two vector variables, **v** being the feature vector describing how the pattern looks and **y** the target vector describing the class membership $k \in \{1..K\}$. The target vector **y** usually is a unit vector with the k-th component 1 and all others 0.

For the purpose of classifier training we have a certain training set available drawn from the application i.e. drawn from the pattern source. The training set shows for each of the examples the desired response **y** for the given input vector **v**.

In any case -- trained or not -- the classifier produces an estimation **d** for the desired output **y**. A convincing optimization criterium is to require that the distance between **y** and **d** should be minimum on the average. Normally the squared Euclidean norm is applied to measure this distance. These considerations turns classifier adaptation into a problem of least mean square functional approximation

$$S^2 = E\left\{\left|\mathbf{d}(\mathbf{v}) - \mathbf{y}\right|^2\right\} = \min_{\mathbf{d}(\mathbf{v})} \tag{1}$$

It can be shown [10,11] that if no restrictions are put on the type of the discriminant function $\mathbf{d}(\mathbf{v})$ the result is the so-called regression function

$$\mathbf{d}(\mathbf{v}) = E\{\mathbf{y}|\mathbf{v}\} \tag{2}$$

Under the conditions considered here -- the target vector $\mathbf{y}$ being a unit vector with the k-th component 1 and all others 0 -- the discriminant function $\mathbf{d}(\mathbf{v})$ comes out to be the vector of a posteriori probabilities

$$\mathbf{d}(\mathbf{v}) = \mathbf{p}(\mathbf{v}) = \begin{pmatrix} \text{prob}(1|\mathbf{v}) \\ \text{prob}(2|\mathbf{v}) \\ \\ \text{prob}(K|\mathbf{v}) \end{pmatrix} \tag{3}$$

The a posteriori probabilities are the optimum class-membership indicator variables stating the probability of being right if the decision is made in favor of the respective class. In this sense they represent also the optimum confidence values.

For practical purposes we cannot work with functions having the flexibility and adaptability of "any arbitrary function of $\mathbf{v}$". Instead, the discriminant function must be taken from a family of parameterized functions $\mathbf{d}(\mathbf{x})=\mathbf{d}(\mathbf{v},\mathbf{A})$, where $\mathbf{A}$ is the set of parameters -- usually a matrix $\mathbf{A}$ of coefficients. Varying $\mathbf{A}$ changes the appearance of the discriminant function $\mathbf{d}(\mathbf{v})$. Within certain limits given by the construction of $\mathbf{d}(\mathbf{v},\mathbf{A})$ the discriminant function can be arbitrarily formed.

The easiest way for constructing such a parameterized discriminant function is to work with linear combinations of a set of L basis functions $f_l(\mathbf{v})$

$$\mathbf{d}(\mathbf{v}) = \mathbf{d}(\mathbf{v},\mathbf{A}) = \sum_{l=1}^{L} \mathbf{c}_l f_l(\mathbf{v}) \quad \text{where} \quad \mathbf{A} = \bigcup_{l=1}^{L} \mathbf{c}_l \tag{4}$$

There are numerous different approaches possible for constructing parameterized discriminant functions. Some of them can be shown to be universal approximators [12] in the sense that by proper adjustment of their parameters they are able to approach any desired function to any degree of accuracy if no restrictions are put on their number L.

Whereas in the above formulation the basis functions $f_l(\mathbf{v})$ are themselves given without any free parameters, it is as well possible to work with basis functions whose appearance again is controlled by further sets $\mathbf{A}_l$ of parameters

$$\mathbf{d}(\mathbf{v}) = \mathbf{d}(\mathbf{v},\mathbf{A}) = \sum_{l=1}^{L} \mathbf{c}_l f_l(\mathbf{v},\mathbf{A}_l) \quad \text{where} \quad \mathbf{A} = \bigcup_{l=1}^{L} \{\mathbf{c}_l, \mathbf{A}_l\} \tag{5}$$

A big advantage of the first mentioned type of discriminant function (4) is that the free parameters contained in $\mathbf{A}$ appear linear in the discriminant function, whereas the discriminant function itself $\mathbf{d}(\mathbf{v})$, however, may be arbitrarily nonlinear in the argument vari-

able **v**. The consequence is that carrying out the optimization leads to a linear system of equations for the coefficient matrix **A** having an economical direct solution.

In discriminant functions of the second type (5) also the adjustable parameters appear nonlinear. This fact requires the application of gradient descent or random search techniques for computing the optimum set **A** of parameters and bears the complications of multiple suboptimum solutions -- local minima.

The parameterized discriminant functions $\mathbf{d}(\mathbf{v},\mathbf{A})$ underlie the same optimization criterion (1) as did the unconstrained function $\mathbf{d}(\mathbf{v})$.

$$S^2 = E\left\{\left|\mathbf{d}(\mathbf{v},\mathbf{A}) - \mathbf{y}\right|^2\right\} = \min_{\mathbf{A}} \tag{6}$$

The result of optimization -- in other words the result of classifier training -- is a least mean square approximation to the optimum solution (3) under the constraints of the chosen construction. Therefore, any pattern classifier trained according to (6) renders class-specific discriminant functions $d_k(\mathbf{v})$ to be interpreted as

$$d_k(\mathbf{v}) = \text{LMS approximation to prob}(k|\mathbf{v}) \tag{7}$$

Of the most prominent classifier types in use in document analysis systems, the multilayer perceptron belongs to the second category whereas the polynomial classifier belongs to the first one. The radial basis functions approach falls under the first category if centers and shapes of the radial basis functions are fixed and under the second category if they are not.

In historic development of pattern classification almost all ever invented pattern classification techniques have been tried out on the example of character recognition. Different from many other applications of pattern classification techniques, there has almost never been a shortage of sample sets in character recognition. Character recognizers normally are very rigorously tested with large test sets and performance diagrams measured as the error/reject operational characteristic or the trade-off between residual error rate and the mean number of alternatives.

In order to achieve comparable results rather early standardized and training and test sets have been collected and made publicly available. Performance contests are regularly carried out on the character as well as on the document level by specialized institutions as the National Institute of Standardization and Technology NIST or the Information Science Research Institute ISRI at the University of Nevada, Las Vegas. A remarkable amount of effort goes into the development and validation of statistical degradation models to be used for generation of synthetic sample sets from perfect character images that shall exhibit the typical perturbations as introduced by real scanning and printing [13].

The same technique of pattern classifier design from learning sample sets which is in the document analysis system so successfully applied for designing character recognizers is in other parts of the whole system employed for deriving decisions and case discriminations. An example is the task of finding potential cut positions in sequences of merged character images from inspecting a certain neighborhood of the potential cut -- cut segmentation classifier [14].

7. Main Fields of Application

One of the driving forces in the childhood of document analysis technology was the challenge to develop reading aids for the blinds [1,15]. With the commercial success in a wide variety of applications this specific goal lost importance. Now off-the-shelve scanners and low-cost OCR-software for PCs can be employed to individually configure tailored reading aids by including suitable speech and Braille output systems.

Presently, by far the biggest commercial potential lies in the postal sorting and mail distributing business. Its present market volume is estimated to be more than 1.5 billion US$ with a yearly increase of about 10%. With 45% thereof almost half of this market is in the US and Canada, 38% in Europe and the rest in East Asia and Australia.

In the industrialized countries the postal organizations rely very strongly on automated address reading. The standard reading and sorting machine has a throughput of more than 40000 letter/hour. The address is read only once and the resulting sort code printed in bar or fluorescent code on the envelope to allow repeated sort runs through large sorting machines based only on bar-code reading.

In most of the countries the whole address is automatically read consisting of zip code, city name, street name, customer name and affiliation if given. The zip code serves as a kind of oversized parity code introducing synthetic redundancy into the address information and making the recognition more reliable. The knowledge base contains all of the necessary postal knowledge which must be invoked in one tenth of a second. The result of the sorting process is a sequence of letters in the postman's walking order.

One of the reasons for the success of document analysis technology in this field has been that the imperfections of automatic reading and understanding are kept from having an influence on the recognition result by organizing an especially clever cooperation between the automatic system and human operators. Unreadable address images are stored in a reject image database and forwarded to computer screens where human operators manually key in the sorting code. The continuous progress in performance of information processing technology was even continuously recoined in reduction of the portion of mail to be manually coded. The most recent improvement in automatic address reading is the introduction of connected handwriting reading.

Another market segment with a long success story is form reading mainly for banking purposes. The developments in this area go far back to the roots of information technology. At this time the approach was guided by the idea of making the reading task simpler for the automated reading devices. For that purpose special fonts were invented for electronic reading, such as magnetic ink codes (MICR) and special fonts for optical reading (OCR-A and OCR-B), and specialized single-font readers designed. The improvements in computer technology and algorithms made these approaches obsolete. Form reading presently is mainly reading of handprinted forms, obviously with the capability of reading any printed material which may appear on the forms in any type of font. Form reading requires knowledge about the forms, their fields and the potential content to make the interpretation more reliable.

A common child of address reading and form reading is business reply mail processing. There are lots of documents in the mailstream carrying information in a form-like structure but not with the convenience of most forms to be of unique size and to be preprinted in so-called drop-out colors invisible to the scanner so that the form background can easily be separated from the content. Business reply mail-pieces are postcards, carrying on their reverse side mail order information and other data relevant for market surveys and similar purposes. These are read in business reply mail processing offices and stored on electronic media for further processing.

The next step in this direction will be automated handling of business mail correspondence. Large organizations with mail-based routine transactions such as banks and insurances are moving into a direction marked by the notion of document imaging. Incoming letters are opened, scanned and stored in CD-Rom archives for computer-based access from any desktop within the organization. Presently, at this occasion keyword information is manually added for indexing and retrieval purposes. This is a job to be taken over by suitable document analysis systems. Thus, document analysis will become part of workflow automation. The computerized assistant first reads the letter, categorizes it according to the message types relevant for the respective company and forwards the letter to the person responsible for this type of business. This approach can be developed further by designing more functionality into the computerized mail forwarding assistant and change it into a mail answering assistant.

8. Concluding Remarks

Document analysis has since its beginnings been one of the topics in pattern recognition and image processing conferences. For a number of years the US Postal Service organized the Advanced Technology Conference series [16] which concentrated on questions relevant for postal automation and among those with document analysis for address reading. Since 1991 within IAPR (International Association for Pattern Recognition) a specialized series of Conferences developed dealing exclusively with all aspects of document analysis, the International Conference on Document Analysis and Recognition ICDAR, which took place St. Malo (France) in 1991 [17], in Tsukuba (Japan) in 1993 [18], in Montreal (Canada) in 1995 [19]. The 1997 ICDAR will happen in Ulm, Germany. Several workshop series are dealing with document analysis related topics such as the Document Analysis Systems workshops [20], the International Workshop on Frontiers of Handwriting Recognition [21], the Workshop on Structural and Syntactical Pattern Recognition [22,23], and the International Graphonomic Society workshop [24].

References

[1] R. Kurzweil, "The Age of Intelligent Machines", MIT Press, Cambridge Massachusetts, 1990

[2] A. Kaltenmeier, F. Class, P. Regel-Brietzmann, T. Caesar, J. Gloger, E. Mandler, "Hidden Markov Models -- A Unified Approach to Recognition of Spoken and Written Language", Proceedings 15. DAGM Symposium, Lübeck, September 1993, pp. 191-198

[3] T.A. Bayer, "Understanding Structured Text Documents by a Model Based Document Analysis System", Proceedings 2. International Conference on Document Analysis and Recognition, Tsukuba, Japan, October 1993, pp. 448-453

[4] A. Dengel, R. Bleisinger, F. Fein, R. Hoch, F. Hönes, M. Malburg, "OfficeMaid -- A System for Office Mail Analysis, Interpretation and Delivery", Proceedings IAPR Workshop on Document Analysis Systems DAS'94, Kaiserslautern, Germany, October 1994, pp. 52-75

[5] K. Loris, "From Document to Recognition -- For People, Not Machines", Advanced Imaging, February 1995, pp. 14-18

[6] N. Bartneck, "A General Data structure for Image Analysis Based on a Description of Connected Components", Computing 42, 1989, pp. 17-34

[7] E. Mandler, M.F. Oberländer, "One-Pass Encoding of Connected Components in Multi-Valued Images", Proceedings 12. International Conference on Pattern Recognition, Atlantic City, 1990, pp. 64-96

[8] K. Fukanaga, "Introduction to Statistical Pattern Recognition", Academic Press, San Diego, 1990

[9] A. L. Spitz, "An OCR Based on Character Shape Codes and Lexical Information", Proceedings 3. International Conference on Document Analysis and Recognition, Montreal, August 1995, pp. 723-728

[10] J. Schürmann, "Pattern Classification, A Unified View of Statistical and Neural Approaches", J. Wiley&Sons, to appear in 1996

[11] J. Schürmann, "Polynomklassifikatoren", Oldenbourg, München 1977

[12] K. Hornik, M. Stinchcombe, H. White, "Multilayer Feedforward Networks Are Universal Approximators", Neural Networks, 2, 1989, pp. 359-366

[13] T. Kanungo, R.M. Haralick, H.S. Baird, "Power Functions and their Use in Selecting Distance Functions for Document Degradation Model Validation", Proceedings 3.

International Conference on Document Analysis and Recognition, Montreal, August 1995, pp. 734-739

[14] T. Bayer, U. Kressel, "Segmentation of Merged Characters", Proceedings 11. International Conference on Pattern Recognition, The Hague, September 1992, pp. 346-349

[15] H. F. Schantz, "The History of OCR -- Optical Character Recognition", Recognition Technologies Users Association 1982

[16] Proceedings Fourth USPS Advanced Technology Conference, Washington DC, USA, November 1990, USPS 1990

[17] Proceedings First International Conference on Document Analysis and Recognition, Saint Malo, France, AFCET - IRISA / INRIA, 1991

[18] Proceedings Second International Conference on Document Analysis and Recognition, August 1993 , Tsukuba, Japan, EEEI Computer Society Press, 1993

[19] Proceedings Third International Conference on Document Analysis and Recognition, August 1995, Montreal Canada, EEEI Computer Society Press 1995

[20] A. Lawrence Spitz, A. Dengel (eds.), Proceedings IAPR Workshop on Document Analysis Systems DAS, October 1994, Kaiserslautern, Germany, World Scientific 1995

[21] Proceedings Fourth International Workshop on Frontiers in Handwriting Recognition IWFHR, Taipei, Taiwan, December 1994

[22] H.S. Baird, H. Bunke, K. Yamamoto (eds.), Extended Proceedings IAPR Workshop on Structural and Syntactic Pattern Recognition SSPR, June 1990, Murray Hill NJ, USA, Springer 1992

[23] D. Dori, A. Bruckstein (eds.), Proceedings IAPR Workshop on Structural and Syntactic Pattern Recognition SSPR, October 1994, Nahariya, Israel, World Scientific 1995

[24] Proceedings Seventh Biennial Conference of the International Graphonomics Society IGS'95, London ON, Canada, 1995