EMPIRICAL LEARNING METHODS FOR DIGITIZED DOCUMENT RECOGNITION: AN INTEGRATED APPROACH TO INDUCTIVE GENERALIZATION

F. Esposito, D. Malerba, G. Semeraro
Int. di Scienze dell'Informazione, Univ. of Bari, Italy

E. Annese, G. Scafuro
Olivetti Systems & Networks - D.O.R., TECNOPOLIS, Valenzano (BA), Italy

ABSTRACT

A hybrid method of empirical and supervised learning to acquire knowledge expressed in the form of classification rules is applied to optically scanned documents with the aim of automatic recognition and storage. An expert system devoted to classification recognizes a document as belonging to a class by its layout and the logical structure of a generic printed page. Not disposing of user defined classification rules, but only of classified examples of documents, decision rules for document classification are inferred by inductive generalization. The learning methodology combines a data analysis technique for linearly classifying with a conceptual method for generating disjunctive cover for each class of documents, taking advantage from the peculiarities of both the approaches.

INTRODUCTION

The problem of page layout recognition of a multimedia document, which is an important step in order to automatically store, retrieve and interchange multimedia office documents can be considered in the wider context of layout and logical structure detection. A document is intended as a related collection of printed objects (characters, columns, paragraphs, titles, figures, etc.), on a paper or microform, for example technical journals or reports. Here only single page documents will be considered. The transformation of such a document into a related collection of information to be stored, classified, retrieved, combined, updated is the objective of a document processing system.

The key idea of this paper is that, provided there is a set of documents with common page layout features, an optically scanned document can be classified in the early phase of its processing flow, by using a defined set of relevant and invariant layout characteristics, "the page layout signature", thus optimizing the system throughput.

The method we propose is derived from the observation that a human is generally able to classify office documents (invoices, letters, order forms, papers, indexes etc.) by a perceptive point of view, recognizing the structure of a form or reading only the content of particular parts of the document. A classification is correct enough for archival purposes, although an accurate retrieval must be realized on the whole content.

Actually, the problem of document recognition has received considerable attention, and several works have been presented relating to layout and logical structure detection. Some of them proceed top-down, [1], recursively subdividing a document into nested subregions according to certain grammar rules (XY-tree structure), while others assemble together primitive components in order to constitute larger regions until the document is completely described, [2]. The approaches proposed so far make use of a rich knowledge base and analyze successfully only those documents possessing a predefined structure. Moreover, the layout (physical level) and semantic structures (logical level) are recognized simultaneously.

In this paper, the characteristics, underlying methodology and experimental results of a system PLRS (Page Layout Recognition System), which recognizes generic documents by detection of layout structure, are presented: it may be considered as an expert system able to classify digitized documents without using optical character recognition or syntactic descriptions of the document given by the user. In fact, a printed page is treated by dealing only with automatically detected and constructed characteristics of the document, namely the geometrical characteristics of the blocks (height, width, spacing and alignment), and the document structure, whose description in a symbolic notation is created. The classification rules are automatically produced using a process of inductive generalization (learning from examples): some significant examples of document classes, relevant to the specific office,
are used as a training sample, in order to
discover the layout similarities and to
derive the discrimination rules
characterizing the classes of documents
and therefore useful in the recognition
step. One of the requirements of an office
automation system is the "in field"
customization and by applying these
learning capabilities we can satisfy such
a functional requirement. In particular,
the system knowledge base can be
automatically updated, taking into account
the variety of structure exhibited in the
office documents; structures ranging from
highly formatted forms and business
letters with standardized structures to
totally unstructured memos or magazine
pages. The learning methodology used to
derive knowledge automatically, in the
form of decision rules which are then
incorporated in PLRS, is based on a hybrid
approach integrating parametric and
conceptual learning methods. [3].

Fig. 1 - The office document handling system architecture

The main objective of a document
processing system is the transformation of
the page contents with the aim of
extracting information for automatic
document classification. A schematic
drawing of the document processing flow
is shown in fig.1. The processing phases
are:

- **preprocessing**: evaluation of a document
  skew through a study of the horizontal
  histogram, a possible rotation and the
  complexity degree of the document, [4];
- **segmentation**: the document is
decomposed into a series of
rectangular areas, using a Run Length
Smoothing Algorithm; quantitative
techniques, based on parameters such as
the black density or the number of
white-black transitions, allow to
discriminate text blocks, image
blocks, graphic blocks, horizontal and
vertical lines, [5];
- **layout analysis**: determines the physical
  layout structure of each page and the
  mutual relationships between different
  basic blocks;
- **page layout recognition**: the
discriminant rules that will be employed
later for classification purposes are
generated, using the physical page
layout;
- **OCR**: character recognition is performed
  by a segmentation process using a
  predefined library and certain Bayesian
decision rules.

Here the focus is on the page layout
recognition step: the recognition of a
document as belonging to a class is
feasible without the necessity of reading
texts or interpreting images. Documents
with significant layout features in
common, arising from the page segmentation
results, can be recognized by an expert
system into which the classification
knowledge has been incorporated. An expert
system, aimed at the classification of
documents and based on the layout
structure, has already been proposed, [6].
Nevertheless, the acquisition of this type
of expert knowledge is the major
"bottleneck" in system development, due to
the difficulty of human experts to explain
their mental process as a chain of
production rules. So, RES, [7], a learning
system for inductive inference of
classification rules, from real examples
of preclassified documents (training
examples), has been used. It has the task
of supplying the rules to the expert
system devoted to the recognition of
documents. This allows for the automatic
updating of the knowledge base when a set
of significantly new documents are
introduced and this is especially
important for the production of new
decision rules.
The layout analysis produces the data concerning document structure; these will be automatically translated into the symbolic and numerical descriptions, which constitute the input into the learning system. After a preliminary global analysis (studying the vertical projection), in order to detect eventual text columns, primitive blocks with homogeneous geometric features are assembled; then, the document is completely described in terms of pages, frames (related to paragraphs, columns) and basic blocks (single text rows, images), obtaining a hierarchical layout tree, as in fig. 2. The first frame level beneath the page level shows the macroscopic physical structure of a document: it allows for the identification of common layout features, that in a number of cases can be regarded as the generic layout structure.

The methodology of RES is definable as an empirical and supervised learning because it focuses on the classification of objects described by values of attributes, disposing of a training set of examples belonging to given disjoint classes; it combines a data analysis method for linear classification (Discriminant Analysis) with a conceptual method for generating disjunctive cover for each classes (INDUBI inspired to AQ algorithms), taking advantage of the respective characteristics of the parametric method which:
- efficiently manipulates quantitative knowledge;
- works on the global characteristics of a class of objects;
- is suitable with a poor a priori knowledge;
- is synthetic;
and of the conceptual method which:
- uses symbolic representation of knowledge;
- works on the structure of an object;
- generates context dependent hypotheses;
- is easily understandable.

Both the methods learn classification knowledge from training cases and are used sequentially in order to exploit the capability of the first in handling continuous descriptors and its robustness as to irrelevant and noisy variables, and the superior capability of the second in formalizing the inductive generalizations as decision rules involving human oriented descriptions that outperform the rules obtained by interviewing an expert.

The application of the Discriminant Analysis makes it possible to obtain a new metadescriptor by synthesizing the relevance, above all, of the continuous numerical descriptors in the formation of the classes; moreover the weight with which each initial descriptor enters to define the discriminant function may be used as a bias for the conceptual method in selecting universal laws by means of generalizations based on a defined number of evidences.

The methodology of RES is definable as an empirical and supervised learning because it focuses on the classification of objects described by values of attributes, disposing of a training set of examples belonging to given disjoint classes; it combines a data analysis method for linear classification (Discriminant Analysis) with a conceptual method for generating disjunctive cover for each classes (INDUBI inspired to AQ algorithms), taking advantage of the respective characteristics of the parametric method which:
- efficiently manipulates quantitative knowledge;
- works on the global characteristics of a class of objects;
- is suitable with a poor a priori knowledge;
- is synthetic;
and of the conceptual method which:
- uses symbolic representation of knowledge;
- works on the structure of an object;
- generates context dependent hypotheses;
- is easily understandable.

Both the methods learn classification knowledge from training cases and are used sequentially in order to exploit the capability of the first in handling continuous descriptors and its robustness as to irrelevant and noisy variables, and the superior capability of the second in formalizing the inductive generalizations as decision rules involving human oriented descriptions that outperform the rules obtained by interviewing an expert.

The application of the Discriminant Analysis makes it possible to obtain a new metadescriptor by synthesizing the relevance, above all, of the continuous numerical descriptors in the formation of the classes; moreover the weight with which each initial descriptor enters to define the discriminant function may be used as a bias for the conceptual method in selecting, during the generalization process, those hypotheses involving the most discriminating descriptors.

The application of the conceptual method permits the production of a set of decision rules H, that are concept generalizations and must satisfy the completeness and the consistency conditions. The rules are structured as

Pattern :- Class

where Pattern is an expression in a formal language and Class represents a concept.

Both the inductive methods of learning, may be used to acquire knowledge automatically: their integration is a way of realizing the unification of statistical and syntactic approaches to pattern recognition as suggested by Fu, [9]. The first problem concerns the representation language, the second concerns the way of combining the results of two methods which emphasize different things.
The methods are both model driven but exhibit relevant differences as to:
- focus: symbolic methods describe concepts in the sense of what unifies a certain number of observations while the parametric methods describe boundary surfaces;
- strategy: symbolic methods work sequentially (one example at a time) while data analysis works in a global way;
- power of representation: conceptual methods allow the handling of structural representations of facts while the statistical methods use single-entity representations.

The possibility of applying an integrated methodology to obtain significant classification rules to recognize documents will be explored by examining and comparing the results arising from the application of the symbolic method alone, of the parametric method alone and, finally, of the combined methodology.

**SYMBOLIC LEARNING TECHNIQUES IN DOCUMENT PROCESSING**

A page layout is constructed from subobjects (or blocks) often closely related to each other and this structural nature of the problem makes the inductive process more difficult so that many well-known learning systems are unsuitable. Inductive generalization has been assigned to a conceptual learning technique, the STAR methodology \( \star \), according to which knowledge is represented using an extension of the first order predicate logic, the VL21 system, whose basic component is the selector or relational statement, written as:

\[
[L \# R ]
\]

where:
- \( L \), called referee, is a function symbol with its arguments;
- \( R \), called reference, is a set of values of the referee's domain;
- \( \# \) is a relational operator defining the relationship between referee and reference.

Selectors can be combined by applying different operators, such as AND, OR, decision operator \( (::>) \) and logic implication \( (\Rightarrow) \) in order to define:
- decision rules used for representing examples from a class (the action part of a rule specifies the class to which the observation belongs);
- inference rules used for representing background knowledge, i.e. the relationships between various descriptors (features);
- generalization rules applied to concept descriptions with the aim to include more examples.

The expressive power of the VL21 to describe the characteristics of objects is strengthened by the introduction of the annotations, a synthesis of the properties of a descriptor (or feature), such as the kind of domain (nominal, linear, tree-structured), the number of variables (or subparts of an object) it refers to, the domain itself, the cost associated and other useful information.

In order to test the approach, a set of 72 single page documents has been considered belonging to nine different classes. Four classes are letters, each class containing generic letters of the same company, with the same logo. Four classes are magazine indexes (PANI, Computer Communications, AT&T Journal, Puglia Scuola). The last class is a "reject" class, representing "the rest of the world". Fifty instances were selected as training examples, leaving the remaining 22 documents for the testing process. All the sample documents

<table>
<thead>
<tr>
<th>DESCRIBERS</th>
<th>DESCRIPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIDTH(Block)</td>
<td>linear domain: [very-very-small, very-small, small, medium-small, medium, medium-large, large, very-large]</td>
</tr>
<tr>
<td>HEIGHT(Block)</td>
<td>linear domain: [smallest, very-very-small, very-small, small, medium-small, medium, medium-large, large, very-large, very-very-large, greatest]</td>
</tr>
</tbody>
</table>

\( \text{ON-TOP(Block1,Block2)} \) denotes the relative position of the region Block within Doc:

- [east, north-east, north, north-west, west, south-west, south, south-east, centre]

\( \text{TO_RIGHT(Block1,Block2)} \) is true if Block1 is above Block2.

\( \text{ALIGN(Block1,Block2)} \) is true if Block2 is on the right of Block1.

\( \text{CONTAIN_IN_POS(Doc,Block)} \) denotes the mutual alignment between Block1 and Block2:

- [first-row, last-row, mid-row, both_rows, first-column, last-column, mid-column, both_columns, no_alignment]

\( \text{CLASS_DISC_ANALYSIS(Doc)} \) denotes the results of the discriminant analysis. Its nominal domain is:

- [one, two, three, four, five, six, seven, eight, Reject]
are real letters received or sent by some firms or copies of the indexes of magazines, so that several forms of noise actually affected them. A page description example is in fig. 3.

The inferential inductive apparatus allowing the system to produce symbolic generalizations of a class of documents conducts a beam search through the space of possible generalizations since an exhaustive search is not possible due to the high demand of computational resources. The results are symbolic discriminant descriptions, satisfying both the completeness and the consistency conditions, expressed as disjunctive VL21 well-formed formulas.

INDUBI was used for hypotheses generation and testing, and the learning process took about 80 minutes on an OLIVETTI LXS 3020. A characteristic of this conceptual method is that of generating classification rules completely covering the training sample (100%). The results, eight maximally specific rules satisfying both the completeness and consistency conditions, may constitute the initial knowledge base of the expert system devoted to classification. Most of them show up the invariant parts of the physical layouts, such as logos and titles of fixed sizes in fixed positions, and their relations with other parts of the documents, as, for example, alignment with other blocks.

For instance, the discriminant rule produced for the class of OLIVETTI letters which the document of fig. 3 belongs to is:

- [ALIGN(S1, S2) = last_row][TO_RIGHT(S1, S2)]
- [WIDTH(S1) = medium_small]
- [HEIGHT(S1) = very_small]
- [WIDTH(S2) = medium_small, small]

Matching this rule with the symbolic description of fig. 3 a possible unification is: (S1/X6, S2/X7).

However, the presence of a disjunctive rule for one class and a trivial one for another must be emphasized: both the classes are magazine indexes. For example, the generalization concerning the class 3, containing Computer Communications indexes, is:

- [TO_RIGHT(S1, S2)] [TO_RIGHT(S1, S3)]
- [TO_RIGHT(S4, S2)] [TO_RIGHT(S3, S2)]
- [HEIGHT(S2) = very very large]

OR

- [ON_TOP(S1, S2)] [TO_RIGHT(S2, S3)]
- [ON_TOP(S1, S3)]
- [ALIGN(S1, S2) = beginning_column]
- [HEIGHT(S1) = medium]
- [HEIGHT(S2) = medium_small]
- [HEIGHT(S3) = medium, medium large, large, very large, very very large]

Moreover, when handling noisy documents, a complete matching procedure for classifying the testing documents cannot be used, so a syntactic distance defined for VL21 descriptions was introduced for a flexible matching. In spite of this, only 50% of the testing documents were recognized by the conceptual method alone.
A PARAMETRIC METHOD TO CLASSIFY DOCUMENTS

The statistical method to learn by examples, namely, Discriminant Analysis, allows, once the classes to which the examples belong are defined, for the calculation of a general function. The form of this function is known while the parameters, capable of identifying the "belonging regions" in the descriptors space, are unknown. Discriminant function \( g_i \) is a mapping from the set of the feature vectors to the real numbers, each value of \( i \) being associated with a single decision region: the discriminant analysis provides some decision rules partitioning the whole feature space \( \Omega \) into a number of regions \( \Omega_i \), \( i = 1, 2, \ldots, N \) where \( N \) is the number of classes.

If \( g_j(X) > g_i(X) \) for all \( j \) \( \iff X \in \Omega_i \)

The linear discriminating functions have the form

\[ g_i(X) = w_{0i} + w_{1i}x_1 + \ldots + w_{pi}x_p \]

where \( w_{pi} \) represents the weight with which the variable \( x_i \) enters to define \( g_i \). The functions are formed in such a way as to maximize the separation of the classes: their values are similar for objects belonging to the same class, while presenting the maximum differences between objects belonging to different classes.

The Discriminant Analysis, working on a set of numerical descriptors characterizing the document as a whole, is used to calculate a function, very useful for synthesizing the relevance of such features, and also for quantitatively specifying their relationships in the definition of the hyperplanes separating the regions.

Many of numerical features concerning a document can be measured as the number, the type and the dimensions of the blocks (such as height, length, area, eccentricity, as well as their means, standard deviations, maximum and minimum values), while other measurements related to the number of black pixels per block can be directly provided by the segmentation process.

Such numerical information, only partially exploited in the first experiment because of the enormous complexity that would result, can be easily managed statistically by the discriminant analysis.

Initially, 93 features were picked out to describe each document and only 4 of them were selected by the stepwise variable selection process minimizing Wilks's Lambda, [11]:

1) standard deviation of the length of the text blocks;
2) maximum height of image blocks in the document;
3) mean number of black pixels in the blocks;
4) maximum number of black pixels in the and/or of the image blocks.

It's possible to see that image blocks, such as logos and magazine headings, play a fundamental role in the document recognition of the selected sample, even if it is not suitable to rely only on them to discriminate documents with a more complex layout structure.

Using such a linear classifier a 100% of the training examples were correctly classified while only three documents belonging to the test set were misclassified and one rejected: they had some form of noise in the image blocks. The percentage of the recognized documents rises to 82%.

THE INTEGRATED APPROACH TO INDUCTIONAL GENERALIZATION

In order to realize a hybrid approach, the results of the discriminant analysis are synthesized in a metaselector, which takes into account the contribution of the continuous numerical variables in the classification of documents. The first step towards the integration concerns the representation language VL21. First of all, descriptors are distinguished as statistical or conceptual in their annotations: the former are only unary VL21 descriptors, that is the reference of a selector contains a single variable which is the symbol used to name the document as a whole. A statistical descriptor appears once in each example and the reference of selectors whose referees are statistical descriptors should contain a single value: these are not syntactic constraints on the representation language, but only semantic tiles imposed by the process of feature vector extraction. According to the principle that each feature should be processed by the proper method, only interval or ratio level measurements are specified as statistical descriptors, since the problem of numerical coding arises for nominal or ordinal level measurements. The uniqueness of the representation, both for conceptual and for statistical discrimination, is due to the fact that in VL21 all the variables used to describe subparts of a document are considered existentially quantified and distinct.

So, the results of the statistical analysis are coded into a selector of the form:

\[ \text{class_disc_analysis}(\text{variable}) = \text{value} \]
where:

- **class_disc_analysis** is a descriptor specifying the class assigned to the example by the discriminant analysis;
- **variable** is the name of the document as a whole;
- **value** is the class identifier.

These selectors are appended to the VL21 descriptions of the training documents; **class_disc_analysis** is a synthesis of the statistical information and from this point of view the discriminant analysis can be considered as a form of constructive generalization. The classification rules induced from the training sample may contain such a metaselector in conjunction with other symbolic descriptions. The set of classification rules may be tested on a suitable testing set before being incorporated into the expert system devoted to the recognition of new documents.

By storing the discriminant function coefficients, it is possible, when a new document has to be classified, to compute its a posteriori class membership probabilities based on statistical descriptors; its description is matched with the symbolic classification rules according to the descending probability order. This should speed up the classification process when the true error rate of the discriminant analysis is low enough.

A schema of the page layout recognition system is in Fig. 4. The input of the document descriptor module is a table containing information about the blocks, superblocks and frames detected in the page layout. Such information is processed in order to produce a number of statistics and numerical features and a symbolic VL21 description of the document. The whole is stored in a document base, and is available to the remaining three modules. The statistical features of the training examples are treated by the discriminant analysis module and the coefficients of the discriminant functions are calculated; these are used for computing the a posteriori probabilities of the class membership of each document, so that the maximum probability determines the class a document belongs to. Finally, the result of the parametric method is coded into **class_disc_analysis** descriptors and a new selector is appended to each symbolic description of the documents. The augmented descriptions of the training documents are then processed by the symbolic learning system (INDUBI) and a classification rule for each class is produced. The last module is a knowledge based system whose purpose is to recognize a new document matching its symbolic description with the symbolic classification rules previously built. At this stage discriminant functions play their role: the **class_disc_analysis** descriptors are

![Diagram](image-url)
generated even for the new documents, and the a posteriori probabilities are exploited to define the order in which the classification rules (generalizations) must be matched with the new document. The result of classification may be a single class with probability equal to 1, that is the system is sure of its recognition of the new document. However it might occur that the document does not match any classification rule; in such a case the system shows a set of probabilities representing the degree of similarity between the description of the document and that of each class. If the system failed in the recognition process, the user could give his negative assessment together with the proper class and force the system to produce new classification rules.

Applying the integrated approach, all the documents belonging to the training set were covered by the classification rules. The generalizations changed significantly, in comparison with that obtained by the conceptual method alone, especially for those classes characterized by disjunctive or trivial rules. For example, the rule concerning class 3, becomes:

\begin{verbatim}
(CONTAIN_IN POS(S3,S2)=north) [TO_RIGHT(S2,S3)=beginning_column]
\end{verbatim}

The throughput time was positively and significantly affected. Moreover, the presence of the metaselector increased the capacity of recognizing the testing documents to 77%; although such a result may seem worse in comparison with the statistical one, it must be emphasized that the rule appears in a more comprehensible human-oriented form, due to the use of the symbolic language. This characteristic might seem less important for an automatic recognition system, however comprehensible classification rules are useful for document reconstruction aims, when it is necessary to connect the physical level with the logical one.

**COMPARISON OF THE EXPERIMENTAL RESULTS**

The results of the three experiments are summarized in Table II where a detailed description of the classification of the test examples is reported.

<table>
<thead>
<tr>
<th>Cor.</th>
<th>1st Exp.</th>
<th>2nd Exp.</th>
<th>3rd Exp.</th>
<th>COMBINED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc cl</td>
<td>CONCEPTUAL</td>
<td>STATISTICAL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1,1,2</td>
<td>1,1,2</td>
<td>1,1,2</td>
<td>1,1,2</td>
</tr>
<tr>
<td>4</td>
<td>2,2,2</td>
<td>2,2,2</td>
<td>2,2,2</td>
<td>2,2,2</td>
</tr>
<tr>
<td>5</td>
<td>3,3</td>
<td>3,3</td>
<td>3,3</td>
<td>3,3</td>
</tr>
<tr>
<td>6</td>
<td>3,5</td>
<td>8,9</td>
<td>8,9</td>
<td>8,9</td>
</tr>
<tr>
<td>7</td>
<td>4,5</td>
<td>4,5</td>
<td>4,5</td>
<td>4,5</td>
</tr>
<tr>
<td>8</td>
<td>4,5,6</td>
<td>4,5,6</td>
<td>4,5,6</td>
<td>4,5,6</td>
</tr>
<tr>
<td>9</td>
<td>5,6</td>
<td>5,6</td>
<td>5,6</td>
<td>5,6</td>
</tr>
<tr>
<td>10</td>
<td>5,6,7</td>
<td>4,5,6,7</td>
<td>4,5,6,7</td>
<td>4,5,6,7</td>
</tr>
<tr>
<td>11</td>
<td>5,6,7,8</td>
<td>5,6,7,8</td>
<td>5,6,7,8</td>
<td>5,6,7,8</td>
</tr>
<tr>
<td>12</td>
<td>6,6</td>
<td>6,6</td>
<td>6,6</td>
<td>6,6</td>
</tr>
<tr>
<td>13</td>
<td>6,6</td>
<td>6,6</td>
<td>6,6</td>
<td>6,6</td>
</tr>
<tr>
<td>14</td>
<td>6,7</td>
<td>6,7</td>
<td>6,7</td>
<td>6,7</td>
</tr>
<tr>
<td>15</td>
<td>7,5</td>
<td>7,5</td>
<td>7,5</td>
<td>7,5</td>
</tr>
<tr>
<td>16</td>
<td>7,5</td>
<td>7,5</td>
<td>7,5</td>
<td>7,5</td>
</tr>
<tr>
<td>17</td>
<td>7,7</td>
<td>7,7</td>
<td>7,7</td>
<td>7,7</td>
</tr>
<tr>
<td>18</td>
<td>8,7</td>
<td>8,7</td>
<td>8,7</td>
<td>8,7</td>
</tr>
<tr>
<td>19</td>
<td>8,7</td>
<td>8,7</td>
<td>8,7</td>
<td>8,7</td>
</tr>
<tr>
<td>20</td>
<td>8,7,8</td>
<td>8,7,8</td>
<td>8,7,8</td>
<td>8,7,8</td>
</tr>
<tr>
<td>21</td>
<td>9,5</td>
<td>9,5</td>
<td>9,5</td>
<td>9,5</td>
</tr>
<tr>
<td>22</td>
<td>9,5</td>
<td>9,5</td>
<td>9,5</td>
<td>9,5</td>
</tr>
</tbody>
</table>

*Table II: Results of the experiments with the testing set*

C: correctly classified
M: misclassified
R: rejected (threshold = 0.85)

Looking at the totals reported in the lower part of the table, it turns out that the conceptual method present up to 4 misclassifications and 5 multiple attributions. This is fundamentally due to the trivial hypothesis generated for one class and its influence in the flexible matching when a measure of fitness is computed. The number of misclassifications decreases with the parametric method and in the combined one. The parallel increase in the number of rejections is less relevant because we prefer rejection to misclassification.

So the problem of defining an utility function in order to measure the performance of the recognizer in the three experiments arose, and the criterion adopted can be synthesized as follows:

\[
\text{N}_{\text{te}} = \text{N}_{\text{cc}} + \sum \text{p}(\omega_i|\text{x}_j) - \sum \text{p}(\text{Rej}|\text{Mis}^k_{\text{x}})\]

where \(\text{N}_{\text{te}}\) is the total number of test examples, \(\text{N}_{\text{cc}}\) is the number of correctly classified test examples, \(\text{p}(\omega_i|\text{x}_j)\) is the a posteriori probability of class \(\omega_i\) for the test example \(\text{x}_j\), and \(\text{p}(\text{Rej}|\text{Mis}^k_{\text{x}})\) is the probability of rejection given multiple attributions.
where:

\[ x_j \] is the j-th example belonging to the i-th class;

\[ N_{CC} \] is the number of test examples correctly classified;

\[ N_{t} \] is the number of test examples;

\[ R^e \] is the set of rejected examples;

\[ M_{is} \] is the set of examples that are erroneously recognized as belonging to the class \( \omega_i \);

\[ p(\omega_k | x_j) \] is the degree of belief that the example \( x_j \) can be classified in the \( k \)-th class.

In short, each rejection gives a positive contribution to the utility in proportion to the belief that an example \( x_j \) belongs to its right class. On the contrary, each misclassification weighs negatively with the utility in proportion to the belief in the misclassification.

Results printed in the last rows of Table II show that the combination of the symbolic and parametric methods produces more selective rules: these allow the misclassification rate to be reduced while keeping the rejection rate within acceptable limits. Moreover the complexity of the rules is not significantly increased and above all their intelligibility is not seriously affected.

**CONCLUSIONS**

A method to acquire knowledge directly from examples in form of classification rules has been applied to optically scanned documents. This classification knowledge may be incorporated in an expert system for the automatic recognition of documents based on the page layout. The rules, namely the hypotheses on the common properties of the layout structure of documents belonging to the same class, are obtained by inductive generalization on preclassified examples, working as a training set.

Two novelties in the problem of automatic document classification have been presented. The first one is the recognition of the layout structure of a generic printed page by using only spatial characteristics of its components (blocks) and relationships between the blocks. The second novelty consists in the proposal of a hybrid learning methodology, that is justified and evaluated by comparison with the conceptual and parametric methods respectively.

As for the first point, particularly in the third experiment, the rate of the correctly classified test documents display that 'understanding' a document type is possible, especially when printed documents with some standardized structure are considered.

With reference to the second point, the first experimental results show a better performance of the integrated approach. In such a comparison a utility function, complying with the heuristic rule that "no classification is better than misclassification", has been used.

**REFERENCES**


