Quality, quantity and generality in the evaluation of object detection algorithms. *

Christian Wolf Jean-Michel Jolion

LIRIS
Laboratoire d'informatique en images et systemes d'information
UMR CNRS 5205
INSA-Lyon, F-69621, France
{christian.wolf,jean-michel.jolion}@insa-lyon.com

ABSTRACT

Evaluation of object detection algorithms is a non-trivial task: a detection result is usually evaluated by comparing the bounding box of the detected object with the bounding box of the ground truth object. The commonly used precision and recall measures are computed from the overlap area of these two rectangles. However, these measures have several drawbacks: they don't give intuitive information about the proportion of the correctly detected objects and the number of false alarms, and they cannot be accumulated across multiple images without creating ambiguity in their interpretation. Furthermore, quantitative and qualitative evaluation is often mixed resulting in ambiguous measures.

In this paper we propose an approach to evaluation which tackles these problems. The performance of a detection algorithm is illustrated intuitively by performance graphs which present object level precision and recall depending on constraints on detection quality. In order to compare different detection algorithms, a representative single performance value is computed from the graphs. The evaluation method can be applied to different types of object detection algorithms. It has been tested on different text detection algorithms, among which are the participants of the Image Eval text detection competition.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

*The work presented in this article has been conceived in the framework of two industrial contracts with France Telecom in the framework of the projects ECAV I and ECAV II with respective numbers 001B575 and 0011BA66.

1. INTRODUCTION

In the past, computer vision (CV) as a research domain has frequently been criticized for a lack of experimental culture, which has been explained by the young age of the discipline. However, experimental evaluation of the theoretical advances is indispensable in all scientific work. We are currently trying very hard to establish a real experimental culture, and the need of strict experimental procedures in applying and evaluating algorithms is widely recognized.

In text and video retrieval, the very successful TREC competition series[1] help to considerately advance the state of the art in this domain. In the field of document image analysis, the Image Eval text detection competition[2], the ICDAR page segmentation competitions [2], the ICDAR text detection competitions [8] and the GREC competition for line and arc detection [12] should be mentioned.

The introduction of the evaluation problem coincides with the emergence of the field of visual information retrieval. As a consequence, the first techniques have been naturally inspired by tools from this domain, as for instance precision/recall graphs which are frequently used in information retrieval. However, visual information has its own specificities, which need to be taken into account. This is the goal of this work.

In computer vision, a successful evaluation algorithm is rarely simple to design. Often it is necessary to conceive non-trivial algorithms in order to ensure an evaluation satisfying scientific requirements:

- A simple and intuitive interpretation of the obtained measures.
- An objective comparison between the different algorithms to evaluate.
- A good correspondence between the obtained measures and the objective performance of the algorithm to evaluate, taking into account its goal.

A particular problem in computer vision, which has already given birth to a multitude of solutions is the problem of detecting objects in images. In this context, by detection we also mean localization, thus tackling a two-part problem.

[1] `http://trec.nist.gov`
[2] `http://www.imageval.org`

We keep the general evaluation framework independent of the object type, defining an object as a visual entity with a spatial reality, and illustrate the concepts with experiments and examples from the field of text detection.

The main contribution of this paper concerns the following issues:

- The separation of detection quality and detection quantity. New performance graphs allow us to easily perceive the detection quantity ("how many objects have been detected?" and "how many false alarms have been detected?") as well as detection quality ("how accurate is the detection of the objects?").
- The influence of the data base is evaluated, i.e. the relationship between the performance of the detection algorithms and the structure of the image test database is put forward.
- The derivation of a single performance value which does not depend on quality related thresholds. Although this performance value, by definition, does not allow us to fully comprehend the behavior of a detection algorithm, it makes it easier to create a ranking of the algorithms to evaluate.

The reminder of this document is organized as follows: Section 2 gives an introduction to the problem and presents different evaluation modes on a hierarchy of different levels, which is formed by the different possible result representations. Section 3 presents a survey on the previous work on the evaluation of object detection algorithms. Section 4 introduces performance graphs for an easy and intuitive interpretation of the detection performance as well as a performance measure. Section 5 demonstrates the dependence of evaluation algorithms on the structure of the test database. Section 6 applies the evaluation measure to two different text detection algorithms and illustrates its intuitive usage. Finally, section 7 concludes.

2. EVALUATION LEVELS

Traditionally, object detection algorithms are evaluated using techniques developed for information retrieval systems. More specifically, the measures of precision and recall are widely used, since they intuitively convey the quality of the results:

$$\begin{aligned} \mathrm{R}_{IR} &= \frac{\text{N.o. correctly retrieved items}}{\text{N.o. relevant items in the database}} \\ \mathrm{P}_{IR} &= \frac{\text{N.o. correctly retrieved items}}{\text{Total n.o. retrieved items}} \end{aligned} \tag{1}$$

In order to have a single performance value for the ranking of methods, the two measures are often linearly combined. The harmonic mean of precision and recall has been introduced by the information retrieval community [10]. Its advantage is that the minimum of the two performance values is emphasized:

$$\mathrm{Perf}_{IR} = 2\,\frac{\mathrm{P}_{IR}\ \cdot\ \mathrm{R}_{IR}}{\mathrm{P}_{IR}\ +\ \mathrm{R}_{IR}} \tag{2}$$

For the object detection problem, the measures of recall and precision are not directly applicable, since the decision whether an object has been detected or not is not a binary one. Object detection algorithms may be evaluated at different levels w.r.t. the representation of the detection results, corresponding to different phases of the detection algorithms. The evaluation measures of the different levels differ in their relevance to the goal of the application and in their coverage, *i.e.* in the detection phases which are evaluated by the measure:

(a) Feature discriminance at pixel level At this level, the quality of the chosen features is evaluated without taking into account the classification decision taken in a later phase. Therefore, the result evaluated for each pixel p is not a binary decision but a feature vector $\boldsymbol{x}_p$. Splitting the pixels into two populations, where the first population consists of the pixels labeled as "object" according to the ground truth, and the second population consists of the "non-object" pixels, the goal of the evaluation measure at this level is to assess whether the features are well separated between the two populations.

(b) Classification at pixel level Once the classification decision for each pixel is available, *i.e.* we know for each pixel whether it belongs to the object or not, the measures of recall and precision may be applied at pixel level. Alternatively, the classification error might be used for evaluation. Note, that if the performance is evaluated at pixel level, then the ground truth must be very precise in order to get robust measures.

(c) Detection at region level From the end user's point of view, a more natural way is to ask the question whether an object has been detected correctly or not. This assumes objects of compact shape, for which for instance a rectangle approach makes sense[3]. This is not appropriate for textures, or objects like snow, falling water, shadows, but does make sense for objects like humans, faces, text, tools etc. The reminder of this document deals with this evaluation level.

(d) Goal oriented evaluation In many applications, object detection is performed for a specific reason which is beyond the pure localization of the object. For instance, face detection might be a preliminary step for face recognition, text detection might be a preliminary step for text recognition, etc.

In this case, in order to take into account the specific goal, the evaluation algorithm should resort to the results of the application specific processing. In the context of text detection, a goal oriented evaluation scheme for a system which exploits the text content (as opposed to its position) should penalize lost text characters as well as additional characters which are not present in the ground truth. Possibilities are recall and precision on character level, or the string edit distance [11]. In the case of text detection for indexing video broadcasts, one might consider evaluation on an even higher level by weighting words according to their usefulness for the indexing process [6].

The evaluation level to choose depends on the application and the purpose of the evaluation. The pixel based evaluation measures are easy to calculate and easy to interpret.

[3]In the following, we will refer to this evaluation level as rectangle level.

However, they lack relevance to the goal of the process and are not very accurate.

The goal directed approaches are natural methods to employ for the final evaluation of the algorithm's performance. However, very often the localization of the object is the final goal of the application. For instance, in the case of face detection or text detection, recognition of the object might be impossible because of low data quality. In image and video indexing applications, the presence of a face or of text is valuable information which can be exploited.

Evaluation levels (a), (b) and (d) are easy to calculate, whereas region based evaluation (level (c)) is a non-trivial task: as the detection result is rarely exactly equivalent to the object as specified in the ground truth, we cannot easily say whether an object has been correctly detected or not. In the reminder of this work, we concentrate on the problem of evaluation on rectangle level.

3. PREVIOUS WORK

The goal of a rectangle based object detection evaluation scheme is to take a list G of ground truth object rectangles $G_i, i = 1..|G|$ and a list D of detected object rectangles $D_j, j = 1..|D|$ and to measure the quality of the match between the two lists. The quality measure should penalize information loss, which occurs if objects or parts of objects have not been detected, and it should penalize information clutter, *i.e.* false alarms or detections which are larger than necessary[4].

Most algorithms are based on an extension of the recall and precision measures which are calculated on the area of two rectangles G_i and D_i and on the area of the overlapping region:

$$\begin{aligned} \mathrm{R}_{AR}(G_i, D_i) &= \frac{Area(G_i \cap D_i)}{Area(G_i)} \\ \mathrm{P}_{AR}(G_i, D_i) &= \frac{Area(G_i \cap D_i)}{Area(D_i)} \end{aligned} \tag{3}$$

Recall illustrates the proportion of the ground truth rectangle which has been correctly detected, and precision decreases if the amount of additional incorrectly detected area increases. In the reminder of this work, we call these measures "area recall" and "area precision", respectively.

Whereas calculating these figures for a single pair of result and ground truth rectangles is straightforward, the extension to the realistic case of two lists of rectangles is not as easy. The existing evaluation methods differ in the way they treat the correspondence problem between the two rectangle lists, *i.e.* whether they consider single matches only or multiple matches, and in the way they combine the figures in order to generate a single measure for multiple rectangles and multiple images.

Doermann *et al.* present a configurable ground-truthing and evaluation system with a graphical java interface [3] for video segmentation. Their system also takes into account temporal matching of objects in videos and provides different temporal matching levels. However, the spatial matching algorithms supported by the tool are rather limited.

In [9], Mariano et al. propose a set of evaluation measures, among which are the area measures on rectangle bases given in equation (3) as well as measures on pixel level.

Antonacopoulos et al. propose an algorithm capable of comparing lists of rectangles [1] in the context of document page segmentation. "Partial misses", "misses" and "merges" are considered. The evaluation algorithm focuses on reporting the accuracy of the detection/classification of each rectangle, the authors do not provide performance measures for a whole document.

A simple evaluation scheme has been used to evaluate the systems participating at the text locating competition in the framework of the 7^{th} International Conference on Document Analysis and Recognition (ICDAR) 2003 [8]. Each rectangle in one list is matched with the best match in the opposing list:

$$\begin{aligned} \mathrm{R}_{ICD}(G, D) &= \frac{\sum_{i=1}^{|G|} BestMatch_G(G_i)}{|G|} \\ \mathrm{P}_{ICD}(G, D) &= \frac{\sum_{j=1}^{|D|} BestMatch_D(D_j)}{|D|} \end{aligned} \tag{4}$$

where $BestMatch_G$ and $BestMatch_D$ are functions which deliver the quality of the closest match of a rectangle in the opposing list:

$$\begin{aligned} BestMatch_G(G_i) &= \max_{j=1...|D|} \frac{2 \cdot Area(G_i \cap D_j)}{Area(G_i) + Area(D_j)} \\ BestMatch_D(D_j) &= \max_{i=1...|G|} \frac{2\ Area(D_j \cap G_i)}{Area(D_j) + Area(G_i)} \end{aligned} \tag{5}$$

A given rectangle may appear in only a single match. If a rectangle is matched perfectly by another rectangle in the opposing list, then the match functions evaluate to 1, else they evaluate to a value < 1. Therefore, the original measures taken from the information retrieval community, given by (1), are upper bounds for the measures given by (4). Both, precision and recall given by (4), are low if the overlap region of the corresponding rectangles is small.

A disadvantage of the ICDAR evaluation scheme is that only one-to-one matches are considered. However, in reality sometimes one ground truth rectangle is "split" into several object rectangles or several ground truth rectangles are "merged" into a single detected object rectangle. This is a problem the authors themselves report in [8]. The problem is generally encountered in detection evaluation frameworks, where an over- or under segmented solution may very well be a correct detection.

Liang *et al.* present a method for the evaluation of document structure extraction algorithms [7]. From the two lists G and D of ground truth rectangles and detected rectangles, they create two overlap matrices σ and τ. The lines $i = 1..|G|$ of the matrices correspond to the ground truth rectangles and the columns $j = 1..|D|$ correspond to the detected rectangles. The values of these matrices correspond, respectively, to area recall and area precision between the row rectangle G_i and the column rectangle D_j:

[4]We should emphasize, that a comparison of the rectangles representing objects is not the same as comparing the objects themselves, since the rectangle based algorithm assumes that the object is identical to its bounding rectangle. In reality, a missed part of G_i may not contain object pixels, or a part of a false alarm in D_i may not contain detected pixels.

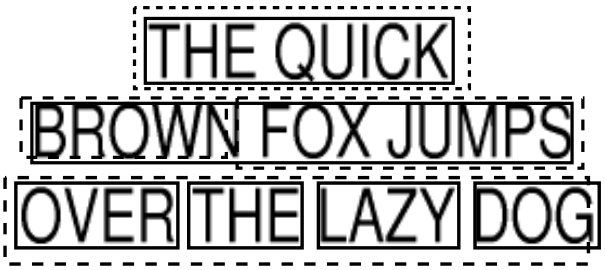


Figure 1: Different match types between ground truth rectangles and detected rectangles. Top: one-to-one match; Middle: a split; Bottom: a merge.

$$\begin{aligned} \sigma_{ij} &= \mathrm{R}_{AR}(G_i, D_j) \\ \tau_{ij} &= \mathrm{P}_{AR}(G_i, D_j) \end{aligned} \tag{6}$$

Matching rectangles is done by thresholding the values in the two matrices and clustering them into groups. Different match types are supported: one-to-one matches, one-to-many matches (splits) and many-to-one matches (merges). See figure 1 for an illustration of these concepts.

Hua *et al.* [4] also take into account splits and merges. They introduce two measures: "detection quality", which relates to recall, and "false alarm rate" which relates to (1 - precision). However, each measure is calculated as product of two factors: a factor which depends on the surface ratios and a factor which measures the rectangle fragmentation. The latter factor decreases in the case of splits and merges.

The evaluation protocol used for the ICDAR 2003 Page segmentation contest [2] is based on the same principles as Liang's method. The overlap matrices (they call them "MatchScore tables") are used to match ground truth entities to detected entities, where an entity (i.e., a region) may contain text, graphics, line-art, a separator or noise, which makes an adaptation of the overlap matrices necessary in order to evaluate the classification of each region. Splits and merges are supported. For each match, a performance value is calculated as the harmonic mean of a recall type measure and a precision type measure. The global performance value for all entities is computed as a weighted sum of the individual scores. The protocol suffers from the same drawbacks: the lack of intuitivity and the ambiguity of the response due to the mixture of detection quality and detection quantity.

Landais *et al.* propose an evaluation measure which is not based on the overlap information [6]: they consider a pair of detected/groundtruth rectangles as matching if and only if the centroid of one rectangle is contained in the other rectangle. This solution tends to accept matches with very low area recall and/or precision and it does not give an information on the quality of the detection.

4. OBJECT COUNT/AREA GRAPHS

Area recall and area precision are easy to interpret as long as there are only two rectangles involved. However, in the case of multiple images or a single image with multiple text rectangles, a combination of the measures is not straightforward.

This is the main drawback of the existing techniques described in the previous section: the way the overlap information is accumulated during the calculation of the evaluation measures leaves room for ambiguity. For instance, a recall of 50% could mean that 50% of the ground truth rectangles have been matched perfectly, or that all ground truth rectangles have been found but only with an overlap of 50%, or anything in between these two extremes. Quality and quantity of the detection are not apparent.

4.1 Requirements of an evaluation algorithm

We developed an evaluation scheme which addresses these problems. We propose a natural way to combine the contradictive measures on quality and quantity: two-dimensional plots which illustrate their dependence. More precisely, on the y-axis we plot the two measures which are the most interesting for us, object counts:

$$\begin{aligned} \mathrm{R}_{OB} &= \frac{\text{N.o. correctly detected rectangles}}{\text{N.o. rectangles in the database}} \\ \mathrm{P}_{OB} &= \frac{\text{N.o. correctly detected rectangles}}{\text{Total n.o. detected rectangles}} \end{aligned} \tag{7}$$

These two measures depend on the quality requirements, which are imposed using two measures: area recall and area precision. In other words, the detection performance is illustrated using two diagrams, where the first shows the dependence on area recall and the second shows the dependence on area precision. Each diagram, on the other hand, contains two graphs: one plots object recall, the other one object precision (see figure 4 in the results section for an example).

4.2 Rectangle matching

The computation of the measures given in (7) requires for each ground truth rectangle G_i the determination whether it has been detected or not, and for each rectangle D_i in the detection result the determination whether its detection is correct or not. These decisions are taken based on constraints imposed on the detection quality, *i.e.* the overlap between detection result and ground truth. In order to take into account one-to-one as well as one-to-many matches (splits) and many-to-one matches (merges), we calculate the overlap matrices σ and τ introduced in section 3.

The matrices are analyzed in order to determine the correspondences between the two rectangle lists. In general, a non zero value in an element with indices (i, j) indicates that ground truth rectangle G_i overlaps with result rectangle D_j. However, the two rectangles are matched only if the overlap satisfies the quality constraints, *i.e.* if area recall and area precision are higher than the respective constraint:

$$\begin{aligned} (a) \quad \sigma_{ij} &> t_r \\ (b) \quad \tau_{ij} &> t_p \end{aligned} \tag{8}$$

where $t_r \in [0, 1]$ is the constraint on area recall and $t_p \in [0, 1]$ is the constraint on area precision. In detail, the different matches are determined as follows:

one-to-one matches: one ground truth rectangle G_i matches with a result rectangle D_j if row i of both matrices contains only one element satisfying (8) and column j of both matrices contains only one element satisfying (8). This situation is shown in figure 1a.

one-to-many matches (splits): one ground truth rectangle G_i matches against a set S_o of result rectangles $D_j, j \in S_o$ if

- a sufficiently large proportion of the ground truth rectangle has been detected (condition (8a) in a "scattered" version): $\sum_{j \in S_o} \sigma_{ij} \geq t_r$, and
- each contributing result rectangle overlaps enough with the ground truth rectangle to be considered a part of it (condition (8b) in a "scattered" version): $\forall j \in S_o : \tau_{ij} \geq t_p$.

many-to-one matches (merges): one result rectangle D_j matches against a set S_m of ground truth rectangles if

- A sufficiently large portion of each ground truth rectangle is detected (condition (8a) in a "scattered" version): $\forall i \in S_m : \sigma_{ij} \geq t_r$, and
- Each ground truth rectangle has been detected with enough area precision (condition (8b) in a "scattered" version): $\sum_{i \in S_m} \tau_{ij} \geq t_p$

many-to-many matches (splits and merges): this match type is currently not supported by our algorithm. Our experiments showed, that this situation does not occur very often in the case of text detection.

If a situation occurs which requires simultaneous splits and merges, then the algorithm translates this situation into several splits or a set of splits and one-to-one matches: each ground truth rectangle in the matching set is either part of a split if it is matched against several detected rectangles, or it is part of a one-to-one match if it is matched against a single detected rectangle. The drawback of this implementation is a slight unjustified punishment of combined splits and merges, since detected rectangles may be part of several sets of splits. In each set, the part of the detected rectangle which covers a ground truth rectangle of another set, is falsely reported as "missing" in the original set.

Based on this matching strategy, the recall and precision measures which we intuitively described in (7), can be finally defined as follows:

$$\begin{aligned} \mathrm{R}_{OB}(G, D, t_r, t_p) &= \frac{\sum_i Match_G(G_i, D, t_r, t_p)}{|G|} \\ \mathrm{P}_{OB}(G, D, t_r, t_p) &= \frac{\sum_j Match_D(D_j, G, t_r, t_p)}{|D|} \end{aligned} \tag{9}$$

where $Match_G$ and $Match_D$ are functions which take into account the different types of matches described above and which evaluate to the quality of the match:

$$Match_G \quad (G_i, D, t_r, t_p) = \begin{cases} 1 & \text{if } G_i \text{ matches against a single detected rectangle} \\ 0 & \text{if } G_i \text{ does not match against any detected rectangle} \\ f_{sc}(k) & \text{if } G_i \text{ matches against several } (\rightarrow k) \text{ detected rectangles} \end{cases}$$

$$Match_D \quad (D_j, G, t_r, t_p) = \begin{cases} 1 & \text{if } D_j \text{ matches against a single detected rectangle} \\ 0 & \text{if } D_j \text{ does not match against any detected rectangle} \\ f_{sc}(k) & \text{if } D_j \text{ matches against several } (\rightarrow k) \text{ detected rectangles} \end{cases}$$

where $f_{sc}(k)$ is a parameter function of the evaluation scheme which controls the amount of punishment which is inflicted in case of scattering, *i.e.* splits or merges. If it evaluates to 1, then no punishment is given, lower values punish more. In our experiments we set it to a constant value of 0.8.

Another possibility could be to use two different functions in the expressions $Match_G$ and $Match_D$ in order to punish over segmentation differently than under segmentation. This might be useful if text detection is followed by text recognition. Furthermore, more scattering might be punished more severely by adding a dependence to the number of rectangles k, for instance by setting $f_sc(k) = (1+\ln(k))^{-1}$, which corresponds to the fragmentation index suggested by Mariano et al. [9].

Note, that text which is only partly detected and therefore not matched against a ground truth rectangle, will correctly decrease the precision measure, in contrast to the ICDAR evaluation scheme described in section 3.

4.3 Multiple images

In the case of N images, we compare several lists $G^k \in \overline{\boldsymbol{G}}, k = 1..N$ of ground truth rectangles with several lists $D^k \in \overline{\boldsymbol{D}}, k = 1..N$ of result rectangles. As in information retrieval, the results on multiple images may not be accumulated by summing the recall or precision values. Instead, object recall and object precision are defined as follows:

$$\begin{aligned} \mathrm{R}_{OB}(\overline{\boldsymbol{G}}, \overline{\boldsymbol{D}}, t_r, t_p) &= \frac{\sum_k \sum_i Match_G(G_i^k, D^k, t_r, t_p)}{\sum_k |G^k|} \\ \mathrm{P}_{OB}(\overline{\boldsymbol{G}}, \overline{\boldsymbol{D}}, t_r, t_p) &= \frac{\sum_k \sum_j Match_D(D_j^k, G^k, t_r, t_p)}{\sum_k |D^k|} \end{aligned} \tag{10}$$

4.4 Constructing the graphs

As explained before, the object related measures introduced in equation (10) depend on two constraints t_r and t_p which impose constraints on the detection quality. The performance diagrams are produced by fixing one constraint to a set value, varying the second one (assigned to the x-axis) and plotting object recall and object precision on the y-axis of two graphs.

Figure 4 in the experimental section shows an example of the two diagrams obtained this way. The diagram shown in figure 4a is generated by varying the constraint on area

Figure 2: An example rectangle detected with area recall = 100% and area precision = 50%.

recall, t_r, while constraint t_p is held to a fixed value. The diagram is composed of three graphs: object recall, object precision and the harmonic mean of the two measures. Similarly, figure 4b is created varying constraint t_p while constraint t_r is fixed.

The diagrams are easily interpreted by looking at the dynamics of the graphs: in this particular example, the fact that object recall never drops to zero when area recall approaches 1 means, that most of the text rectangles are detected with an area coverage of 100%, *i.e.* the detection rarely cuts parts of the ground truth rectangle. On the other hand, the fact that object recall does drop to zero when area precision approaches 1, means that all result rectangles exceed the ground truth boundaries. The particular amount of area which is detected additionally can be seen by the point/range where the object recall dramatically drops when area precision increases.

As stated above, during the creation of the graphs one of the two constraints is held fixed. The particular values assigned to the fixed constraints have been chosen empirically. However, we decided to pick different values for the two different constraints: while t_r is fixed to 0.8, we chose the lower value of 0.4 for constraint t_p. This decision is motivated by the fact that a detection result which cuts parts of the text rectangle is more disturbing than a detection which results in a too large rectangle. The value of 0.4 might seem very low, but keep in mind that the area of a square is a quadratic function of its side length. This fact is illustrated in figure 2, which shows a detection result with 50% area precision. The detected rectangle is twice as large as the ground truth rectangle, although the difference in the corner coordinates is quite small. Please refer to the discussion section for some remarks on the implications of this situation to text detection algorithms.

4.5 A single performance value

Very often it is useful and desirable to determine a single performance value for an algorithm, either for direct comparison of the performances of different algorithms, or to optimize the parameters of the detection algorithm.

For the reasons laid out in section 4.1, an objective comparison of the algorithms by a single scalar value is difficult, up to impossible. A single value is hardly able to characterize the complex behavior of a detection algorithm, which makes it necessary to resort to compromises. At first sight, a simple solution might be to hold the quality constraints t_p and t_r at fixed values, calculate object recall and object precision and combine them in a harmonic mean. However, this evaluation would depend heavily on the particular chosen values. One algorithm could outperform another one for given quality constraints, while it could show a weaker performance for other constraints. A good indicator should cover the performance of the evaluated algorithm across a whole range of quality constraints. We therefore propose the proportion of the graph area which is beneath the performance graphs as a reliable and objective measure, which is equivalent to the mean value of object measures over all possible constraint values.

More precisely, we first calculate the area proportion separately for object recall and object precision:

$$\begin{aligned} \mathrm{R}_{OV} &= \tfrac{1}{2T}\sum_{i=1}^{T} \mathrm{R}_{OB}(\overline{\boldsymbol{G}}, \overline{\boldsymbol{D}}, \frac{i}{T}, t_p) + \frac{1}{2T}\sum_{i=1}^{T} \mathrm{R}_{OB}(\overline{\boldsymbol{G}}, \overline{\boldsymbol{D}}, t_r, \frac{i}{T}) \\ \mathrm{P}_{OV} &= \tfrac{1}{2T}\sum_{i=1}^{T} \mathrm{P}_{OB}(\overline{\boldsymbol{G}}, \overline{\boldsymbol{D}}, \frac{i}{T}, t_p) + \frac{1}{2T}\sum_{i=1}^{T} \mathrm{P}_{OB}(\overline{\boldsymbol{G}}, \overline{\boldsymbol{D}}, t_r, \frac{i}{T}) \end{aligned} \tag{11}$$

The final performance value is the harmonic mean of the two measures.

The parameter T is a granularity parameter which controls the trade-off between the computational complexity of the evaluation algorithm and the precision of the integration approximation. In our experiments, we set the parameter to $T = 20$.

5. EVALUATING THE INFLUENCE OF THE TEST DATABASE

As for information retrieval (IR) tasks, the measured performance of an object detection algorithm highly depends on the test database. It is obvious, that the nature of the images determines the performance of the algorithm. An objective comparison between different algorithms will only be possible if the respective communities decide on shared common test databases. Alternatively, we recommend tackling this problem partly by performing different experiments for different test databases with different difficulties.

The structure of the data, *i.e.* the ratio between the relevant data and the irrelevant data, is another major factor which influences the results. In [5], Huijsmans *et al.* call attention to this fact and adapt the well known precision/recall graphs in order to link them to the notion of generality for an IR system, which is defined as the ratio between the number of relevant items and the number of all items in a database.

Very large databases with low generality, *i.e.* much irrelevant clutter compared to the relevant material, produce results with lower precision than databases with higher generality. A standard IR system presents the retrieved items to the user in a result set of predefined size. Since this size is fixed, with falling generality the amount of relevant material in the result set — thus the recall — will tend to be smaller. Thus, recall and precision depend on the generality of the database. In IR one is interested in the retrieval performance with respect to the generality as well as with respect to the size of the result set, which determines the search effort for the user. The dependence on two parameters makes three-dimensional performance graphs necessary. Alternatively, Huijsmans proposes two-dimensional graphs, which corresponds to a plane of the 3D space defined by Precision = Recall. Therefore, the graph plots Precision=Recall on the y-axis against generality on the x-axis.

However, unlike IR tasks, object detection algorithms do not work with items (images, videos or documents). Instead,

images (or videos) are used as input, and object rectangles are retrieved. Nevertheless, a notion of generality can be defined as the amount of objects which are present in the images of the database. We define it to be

$$\mathcal{G} = \frac{\text{N.o. object rectangles}}{\text{Average n.o. obj. per relevant image} \times \text{N.o. images}} \quad (12)$$

where the first factor in the denominator is the average number of objects per image containing objects, a constant which can be estimated for each database, or set to 1. Its actual value does not change the behavior of the mesure, only its scale.

Another difference to IR systems is the lack of a result set window, because all detected items are returned to the user. Therefore, the generality of the database does influence precision, but *not* recall. Thus, the influence of the database structure on the system performance can be shown with simple two-dimensional precision/generality graphs. In reality, this graph should be very close to a straight line since it depends on a single property of the detection system: the average amount of rectangles detected in an image which does *not* contain any desired object. A very simple and straight forward way to evaluate the influence of generality on different detection algorithm is therefore to compare these values.

A decision needs to be made concerning the generality level of the database when result tables or graphs are displayed which contain a fixed level of generality. In other words, it is necessary to decide how many images with zero ground truth (no object present) should be included in the database. The exact amount depends on the particular application. The *a priori* probability of an image to contain exotic objects, as for instance water falls or fire might be very low. Another determining factor is the type of medium. In most cases, for applications working on single images the probability is higher than for applications working on video sequences. In this document, where experiments were performed on images containing text objects (see section 6), we chose a mixture of 50% images with relevant objects and 50% images without relevant objects.

6. EXPERIMENTAL RESULTS

We tested our metric on two different test detection algorithms applied to different test databases.

6.1 Evaluating text detection in video frames

The first test dataset contains two algorithms, which have been developed by the authors [14] [13]. For the sake of brevity, in the reminder of this paper we call them *algorithm 1* and *algorithm 2*. The two methods have been applied to a small set of video frames in the CIF format (384×288 pixels), which have been provided by INA[5] and France Telecom. This small database contains only 14 images, which makes it possible to visually show the detection results superimposed on the images (see figure 3). Thus, a direct comparison can be made between the detected object rectangles and the object/area performance graphs (figure 4).

The left column of figure 4 shows object recall and precision depending on the constraints imposed on area recall. Object recall and precision decrease only slowly when t_r approaches 1, which means that most of the object rectangles are detected with their entire area. Note, that the object recall graph drops faster for algorithm 2, illustrating a lack of the algorithm to detect the whole area of each rectangle. This can be confirmed looking at the superimposed results in figure 3a and figure 3b, respectively.

The right column of figure 4 shows object recall and precision depending on the constraints imposed on area precision. Object recall and precision drop to zero when t_p approaches 1, illustrating the fact that all object rectangles are larger than the corresponding ground truth rectangles. We can see that algorithm 1 is more precise, since object recall drops slower when the t_p is increased. Again, this is confirmed looking at the superimposed results in figure 3.

6.2 The Image Eval text detection competition results

The second dataset consists of the text detection algorithms participating at the ImageEval text detection competition 2006[6]. Pierre-Alain Moellic, the organizer of the competition, kindly provided results of the participants in XML format. The test image database consists of various images: postal cards as well as color and black and white photographs.

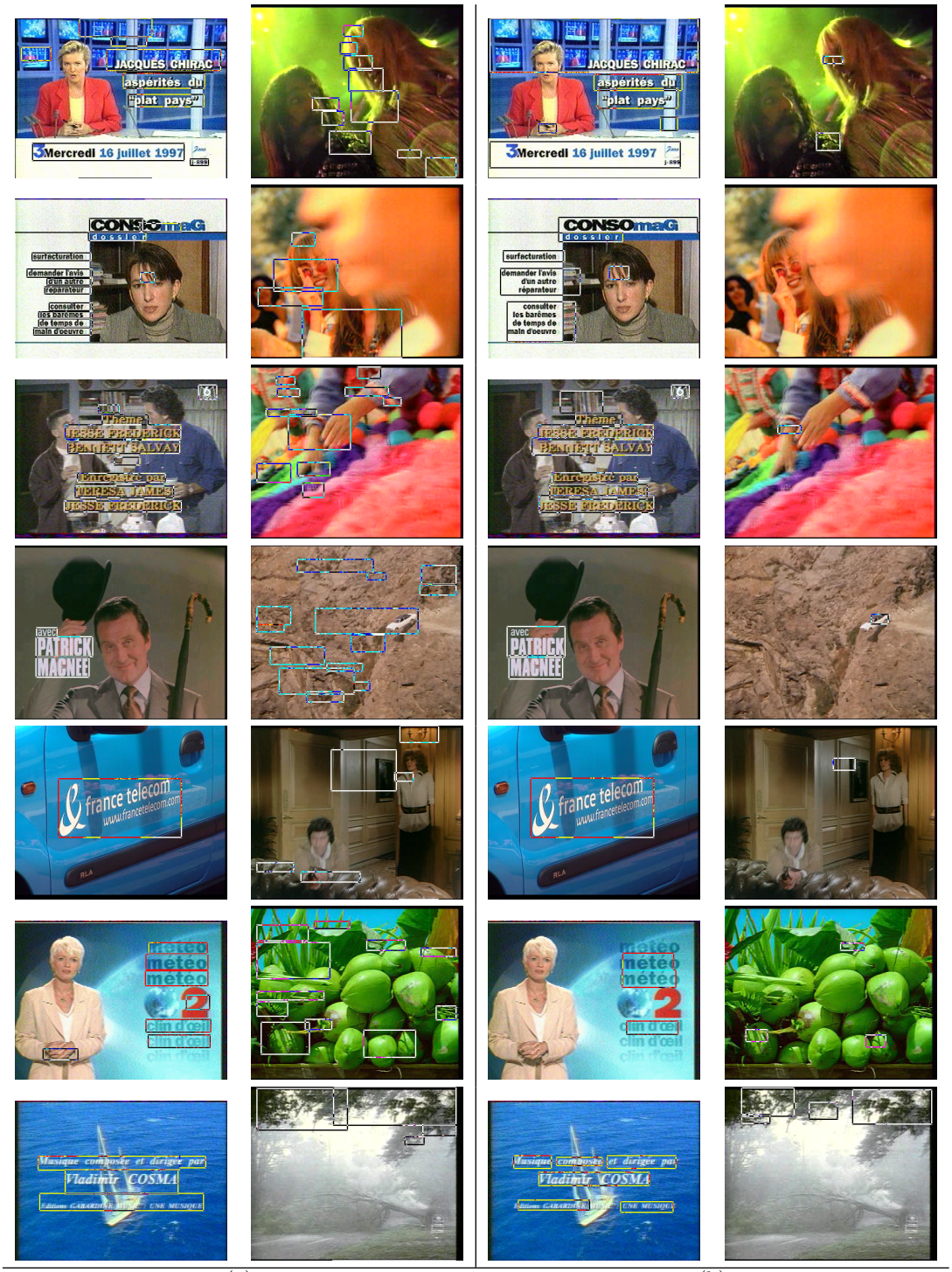
Five runs submitted by two different teams have been evaluated. The performance graphs of the best run for each team are shown in Figure 5 and the corresponding performance values are shown in table 1. The first run seems to have some performance troubles which can be noted immediately from the low overall performance of 18.3%. Looking at the graphs we get a clearer picture of the reasons: the left curve (varying t_r on the x-axis) shows that the object related performance drops drastically when the requirements on area recall are increased. An object level performance comparable to the other runs may be achieved when the area recall requirements are set to a small $\epsilon > 0$. In order words, the algorithm detects almost the same number of rectangles than the better algorithm, but at a very low precision, i.e. only detecting a very small part of the rectangles.

In general, the performance characteristics of the detection algorithms are well illustrated by the graphs: the proportion of "recalled" objects and the proportion of false alarms is immediately visible for the quality a user might want to impose. Inflection points in the performance curves show the precision of the detection algorithm. Runs 2 to 5 show a rather flat dependence on area recall, with a drop of the performance around 80% - which is quite typical of object detection algorithms, e.g. compared to the graphs computed on the results of the ICDAR 2003 text detection competition[15].

Table 1 presents the performance values for each algorithm compared to the metric used during the ICDAR competition, introduced in section 3. The ranking of the algorithms stayed the same, although there are differences in the different performance values. More important, the interpretation of the values changes: recall according the ICDAR metric corresponds to the area recall, averaged across all images, which results in the ambiguity described in section 4. On the other hand, the new recall value corresponds to averaged object recall and may thus be interpreted as the

[5]The French national institute in charge of the archive of the public television broadcasts.

[6]`http://www.imageval.org`



(a) (b)

Figure 3: Some detection examples: (a) detection algorithm 1 (b) detection algorithm 2

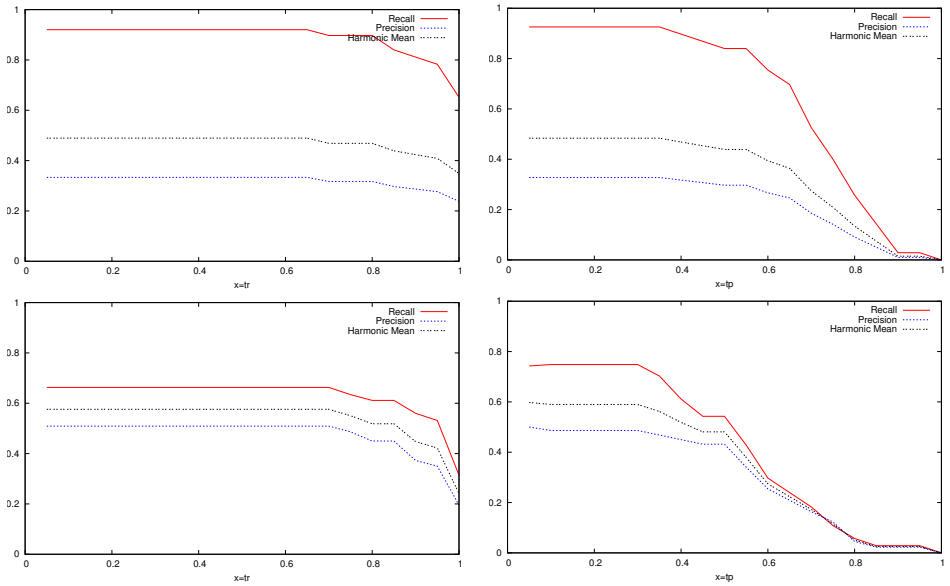


Figure 4: Results on the images shown in figure 3. Top: detection algorithm 1, bottom: detection algorithm 2; Left: varying constraint t_r (area recall) while t_p is constant and equal to 0.4, right: varying constraint t_p (area precision) while t_r is constant and equal to 0.8;

Method	Recall	Precision	H.Mean
	ICDAR Metric (eq. (4))		
Run 1	31.0	23.6	26.8
Run 2	65.1	48.3	55.4
Run 3	67.7	48.0	56.2
Run 4	67.9	47.6	56.0
Run 5	68.9	48.8	57.1
	New Metric (eq. (11))		
Run 1	18.3	15.9	17.0
Run 2	55.1	42.4	47.9
Run 3	57.6	42.3	48.8
Run 4	58.0	42.0	48.8
Run 5	59.7	43.8	50.5

Table 1: Single performance values on the Image Eval 2006 data set.

proportion of correctly detected objects, averaged across the whole range quality constraints a user might want to impose. Precision is interpreted in a similar manner.

7. DISCUSSION AND CONCLUSION

In this paper we have presented a novel method to evaluate object detection algorithms. The proposed method is applicable to any kind of object, as long as the detection result may be represented by a list of rectangles.

We introduced diagrams containing 2D graphs which depict measures on object level depending on quality constraints, making easy a clear and intuitive interpretation. A clear distinction is made between a quantitative evaluation of the detection algorithm and a qualitative evaluation. The dynamics of the graphs illustrate the behavior of the detection algorithm against different quality constraints which might be imposed by a user. The proposed evaluation method overcomes several shortcomings of the existing approaches, notably the ambiguity problem which follows from the direct accumulation of overlap proportions. Since the performance values are calculated on object level, a user can directly see the number of correctly detected objects and the amount of false alarms.

For the comparison of different detection algorithms we have proposed a single performance measure which is directly derived from the performance graphs.

Our evaluation method is based on the amount of overlap between the ground truth rectangles and the result rectangles, not on the location of this overlap. In many applications, e.g. in the case of text detection, however, the amount of overlap between two rectangles is not a perceptively valid measure of quality: Error space around the rectangle might be less harmfull than a concentration of the error space at one side of the rectangle.

As specified in section 4.4, in order to prevent the rejection of detection results as the one in figure 2, the precision constraint t_p is set to a very low value. This is necessary because the error surface grows with the square of the additional rectangle length (or height). However, we still might want to reject detections of cases where the error is concentrated at one side of the rectangle.

A statistical test using all error pixels would be overkill given the fact that the functional form of the error distribution is known and that it depends on 4 parameters only: the absolute differences of the left (respectively right, upper and lower) coordinates of the rectangle pair. We chose therefore a simpler yet more effective method, which directly checks these parameters. In the more specific case of text detection, we are more interested in detecting a horizontal disequilibrium. Therefore, we concentrate on two of the differences measures: the absolute differences of the left (respectively right) coordinates of the rectangles to match need to be smaller than a constraint which depends on the width of the ground truth rectangle.

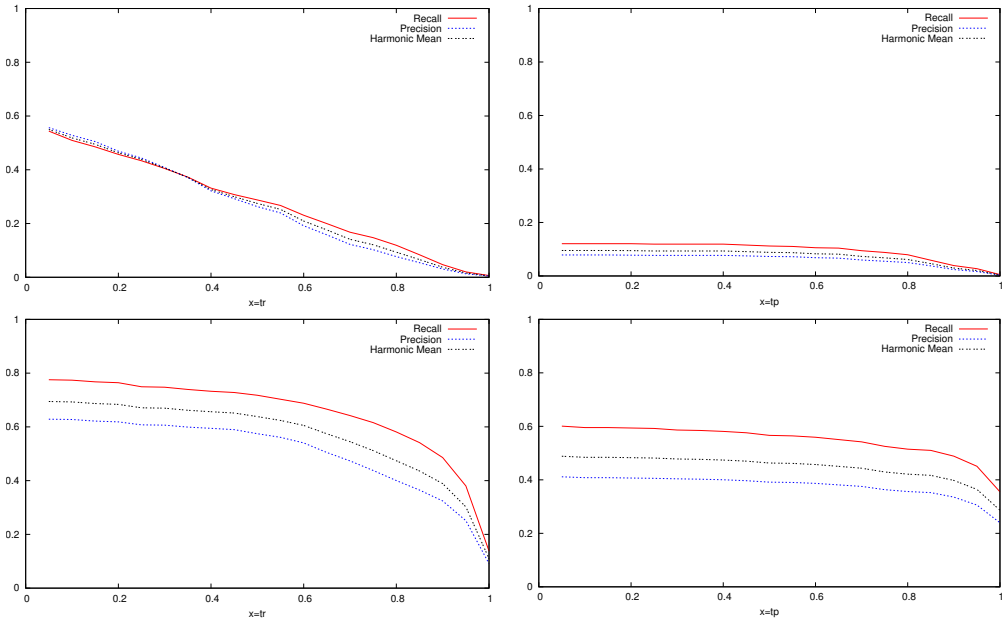


Figure 5: Results on the Image Eval 2006 data set. The two rows show the best run of each participant. Left column: varying constraint t_r (area recall) while t_p is constant and equal to 0.4; Right column: varying constraint t_p (area precision) while t_r is constant and equal to 0.8;

8. REFERENCES

[1] A. Antonacopoulos and A. Brough. Methodology for Flexible and Efficient Analysis of the Performance of Page Segmentation Algorithms. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 451–454, 1999.

[2] A. Antonacopoulos, B. Gatos, and D. Karatzas. ICDAR 2003 Page Segmentation Competition. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 688–692, 2003.

[3] D. Doermann and D. Mihalcik. Tools and Techniques for Video Performance Evaluation. In *Proceedings of the International Conference on Pattern Recognition*, volume 4, pages 4167–4170, 2000.

[4] X.-S. Hua, L. Wenyin, and H.-J. Zhang. An Automatic Performance Evaluation Protocol for Video Text Detection Algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):498–507, 2004.

[5] N. Huijsmans and N. Sebe. Extended Performance Graphs for Cluster Retrieval. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 26–31, 2001.

[6] R. Landais, L. Vinet, and J.-M. Jolion. Ciblage de l'optimisation d'un système de détection d'objets. *to appear in Traitement du signal*, 2007.

[7] J. Liang, I.T. Phillips, and R.M. Haralick. Performance evaluation of document layout analysis algorithms on the UW data set. In *Document Recognition IV, Proceedings of the SPIE*, pages 149–160, 1997.

[8] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. ICDAR 2003 robust reading competitions. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, volume 2, pages 682–687, 2003.

[9] V.Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer. Performance Evaluation of Object Detection Algorithms. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 965–969, 2002.

[10] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[11] R.A. Wagner and M.J.Fisher. The string to string correction problem. *Journal of Assoc. Comp. Mach.*, 21(1):168–173, 1974.

[12] L. Wenyin and D. Dori. A Protocol for Performance Evalution of Line Detection Algorithms. *Machine Vision and Applications: Special issue on performance evaluation*, 9(5-6):240–250, 1997.

[13] C. Wolf. *Text Detection in Images taken from Videos Sequences for Semantic Indexing*. PhD thesis, INSA de Lyon, 2003. `http://liris.cnrs.fr/christian.wolf`.

[14] C. Wolf and J.-M. Jolion. Extraction and Recognition of Artificial Text in Multimedia Documents. *Pattern Analysis and Applications*, 6(4):309–326, 2003.

[15] C. Wolf and J.-M. Jolion. Object count/Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms. *International Journal on Document Analysis and Recognition*, 8(4):280–296, 2006.