# OCR Based Slide Retrieval

Nabil Daddaoua [*] Jean-Marc Odobez [*]
Alessandro Vinciarelli [*]

IDIAP–RR 05-11

March, 2005

SUBMITTED FOR PUBLICATION

[*] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute www.idiap.ch
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-11

# OCR Based Slide Retrieval

Nabil Daddaoua Jean-Marc Odobez Alessandro Vinciarelli

March, 2005

SUBMITTED FOR PUBLICATION

**Abstract.** This work addresses the problem of acquiring, indexing and retrieving slides in the context of automatic oral presentation processing. Since the most suitable acquisition technique, in such a context, is the use of a framegrabber (a device capturing as images the slides displayed on a screen), the slides must be transcribed with an Optical Character Recognition system. Retrieval experiments performed on a corpus of 570 slides (26 presentations) gathered at a workshop show that performance obtained with the OCR transcriptions are close to those obtained by extracting the text from the electronic version (pdf or ppt) of the slides (through apposite API's).

# 1 Introduction

Presentations are a common event in many working environments, but their content is difficult to process and store. Often, after the presentation is given, no record is left and the only possibility to access again the information delivered is to contact the speaker. This results into a loss of potentially valuable information that can be avoided by developing effective indexing and retrieval techniques.

The most simple approach is the so-called *record and playback* approach [MS99]: each presentation is recorded and the resulting videos are available to interested users. This approach has evident limitations: it is difficult to identify a presentation of interest in a large collection and, once the presentation of interest is found, it is not evident how to identify the few minutes concerning a specific topic.

In this work we address the above limitations by applying Information Retrieval technologies to the slides used in presentations. This slide-based approach does not, in our opinion, impose a restrictive constraint. Slides are in fact very frequently, if not always, used as a support of oral presentations. Moreover, slides contain the main messages the speaker wants to convey as well as basic facts and figures. For such reasons, slides can be considered, in our opinion, as a good approximation of the whole presentation content.

The slides can be obtained essentially in two ways: the first is to obtain a copy of their electronic version (i.e. a ppt or pdf file) from the speaker, the second is to capture the images projected onto the screen during the talk through a framegrabber. In the first case, it is possible to convert the slides into text by using commercial software based on apposite API's. In the second case it is necessary to perform a transcription through an Optical Character Recognition (OCR) system [COB04, LW02, LDK00]. The first solution leads to transcriptions affected by few errors, but has several problems: the slides still need to be automatically synchronized with the presentation video, the slide formats are proprietary and the API's necessary to access them are expensive. Moreover, proprietary formats change frequently, thus the API's become obsolete after a relatively short lifespan. The second solution leads to transcriptions affected by some errors but it solves all of the problems left open in the previous case: the framegrabber output can be synchronized automatically with videocameras and and each slide can be linked to a video recording segment. In this way, the retrieval of video segments can be done through the retrieval of slides. The slides are captured as images (e.g. in jpeg format) that are independent of the original slide format. The transcription system becomes thus independent of proprietary formats and does not need to be changed each time the proprietary format is modified. Moreover, the acquisition of slides through a framegrabber is not affected by environmental effects (e.g. light changes or background noise), which facilitates the OCR processing.

For the above reasons, this paper investigates the possibility of applying Information Retrieval technologies to slide transcriptions obtained by using an OCR system over slide images. The same retrieval tasks have been performed over transcriptions of the same slide corpus (26 presentations for a total of 570 slides collected at a workshop [BB05]) obtained manually, using commercial API based software, or different versions of our OCR system. The experiments show that the retrieval performance obtained using the OCR transcripts are

close to the performance obtained with the API. Additionally, experiments shows that the OCR-based system has the ability to extract, index and retrieve text embedded in figures that are generally not accessible to the API software.
The rest of this paper is organized as follows: Section 2 describes the OCR system used in this work, Section 3 presents our retrieval approach, Section 4 shows experiments and results and Section 5 draws some conclusions.

# 2 Text recognition system

The method we employ in this article has been described in [COB04]. We present here an overview of this method, and the reader can refer to [COB04] for more details.
The method follows a top down approach, where text line regions are first localized in the image, and a text recognition system is then applied on the extracted regions.

The text line localization algorithm has two components. The first one consists in classifying each pixel of the image into either text or non-text. This was achieved using simple feature computation and morphological processing. The second part of the algorithm aims at identifying individual text lines from the generated text-labeled binary map. This is achieved by searching in a systematic way for the top and bottom baselines of horizontally aligned text string regions with enough density.

The common approach to recognize text from individual text lines consist of applying a binarization algorithm on the text image followed by the use of a standard OCR software. Such an approach, however, generates many errors, as the distribution of gray-scale levels in the text region may not be bimodal (e.g. due to the use of several color, the presence of background...), and the OCR is often confused by similar-looking characters (*e.g.* l, I, 1, i,. . .). To address these issues, we proposed in [COB04] a scheme whose principle is illustrated in Figures 1, and which can be briefly described as follows: first, a segmentation algorithm that classifies the pixels into $K$ classes is applied to the text image. Then, the segmentation is exploited to produce binary text image hypotheses (*e.g.*, by assuming that a label, or a conjunction of labels, corresponds to the text layer). The resulting binary images are then passed through a post-processing step and forwarded to the OCR system, in this way producing different string hypotheses. The text result is selected from all the generated hypotheses based on a confidence value computed for each recognized string based on langage modeling and OCR recognition statistics. In the experiments, we have considered the three following variant of the method to produce the transcript:

1. **Trans2 :** this is the usual binarization process, where a segmentation process with K=2 classes is applied, resulting in the generation of two strings. The string with highest confidence is used as the transcript.

2. **TransBest :** as shown in Fig. 1, the recognition process is applied three times, by segmenting the image with a K value of 2, 3 and 4. >From all the generated text string hypotheses, the string with highest confidence is used as the transcript.

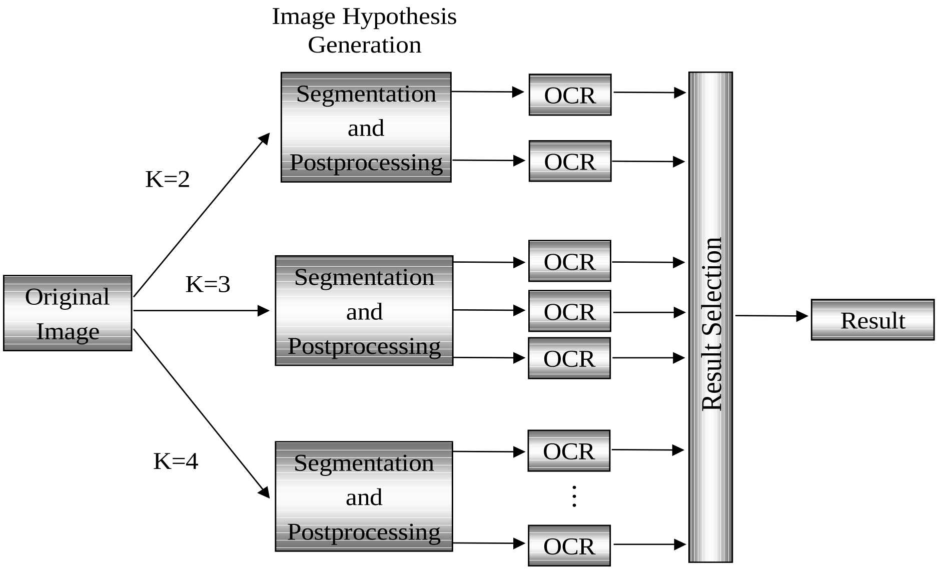
Image Hypothesis Generation
Original Image
K=2
K=3
K=4
Segmentation and Postprocessing
Segmentation and Postprocessing
Segmentation and Postprocessing
OCR
OCR
OCR
OCR
OCR
OCR
OCR
Result Selection
Result

Figure 1: Overall text recognition scheme.

3. **TransAll :** in the current application, the most important point is to obtain a transcript with as many slide words as possible correctly recognized. To optimize this criterion, we propose to use the following strategy. >From the set of text strings obtained for a single value of $K$ (see Fig.1), we keep the string with the highest confidence. The transcript is then generated by adding to the most confident of the obtained 3 strings all the words of the two other strings that are not present in the most confident one. With this strategy, we palliate the sensitivity of the OCR engine, which sometimes produces strings from different segmentations that only differ by one letter.

The transcripts obtained with any of these 3 methods will be used to index the slides as described in the next section.

## 3 Information Retrieval

Information Retrieval is the task of finding, among a large corpus of documents, those who are relevant to an information need expressed through a query. The literature proposes several approaches and in this work we use the so-called Vector Space Model (VSM), the most effective and widely applied one [BYRN99]. A system based on such an approach can be divided into two parts defined *offline* and *online* respectively.
The offline part takes as input the raw data (a corpus of documents) and gives as output the *term by document* matrix $A$, where each column corresponds to a document and each row corresponds to a term in the *dictionary* (the list of unique terms appearing in the corpus). The offline part is performed once for a given database and it is composed of four steps: *preprocessing*, *stopping*, *stemming* and *indexing*. Preprocessing simply removes all of the non-alphabetic characters (parentheses, punctuation marks, etc.). Stopping removes from

the documents all of the words (called *stopwords*) needed to make a sentence gramatically correct but not related to the document content (e.g. articles, prepositions, verbs of common use like *to be* or *to have*). Stemming replaces all of the morphological variants of the same word (e.g. *connection, connecting, connection*) with their stem (*connect*). The reason is that the meaning of the word is supposed to be carried by the stem and not by the morphological variations.

At the end of the stemming, the original documents have been converted into streams of terms, but this is not a suitable form for the retrieval process. It is thus necessary to index the documents, i.e. to convert them into vectors that can be processed by the online part. The document vectors can thus be arranged in the term by document matrix $A$ (see above). The element $a_{td}$ of $A$ accounts for the presence of term $t$ in document $d$. The literature proposes several alternatives to express the $a_{td}$ elements. In this work we use the Okapi formula which is the most effective and commonly applied [RWB00]:

$$a_{td} = \frac{tf(t,d) \cdot \log\left(\frac{N}{N_t}\right)}{k \cdot [1 - b + b \cdot NDL(d)] + tf(t,d)} \tag{1}$$

where $tf(t,d)$ is the number of times term $t$ appears in document $d$ (the *term frequency*), $N$ is the total number of slides in the database, $N_t$ is the number of documents containing term $t$, $k$ and $b$ are hyperparameters and $NDL(d)$ is the normalized document length (the length of $d$ divided by the average document length in the database). The logarithm is referred to as *inverse document frequency* (idf) and gives more weigth to the terms appearing in few documents because they are supposed to be more discriminative.

The online part takes as input a query $q$ expressed in natural language (e.g. *multimodal meeting analysis*) and gives as output a ranking of the documents based on their Retrieval Status Value $RSV(q,d)$, i.e. a score accounting for the relevance of $d$ to $q$. The documents relevant to $q$ are expected to appear at the top ranking positions. In the Okapi based systems, the $RSV$ is calculated as follows:

$$RSV(q,d) = \sum_{t \in Q} a_{td} \tag{2}$$

where $Q$ is the set of the terms contained in the query $q$. Since the value of $a_{td}$ is zero when term $t$ does not appear in document $d$, the above RSV expression tends to be higher when $d$ and $q$ share more terms. However, not all of the common terms contribute in the same way. The presence of $tf(t,d)$ at the numerator of $a_{td}$ (see Equation 1) gives more weight to the terms appearing more times in $d$ (they are supposed to be more representative of its content). The inverse document frequency makes the contribution of terms appearing in few documents higher (they are supposed to be more discriminative). The main limitation of such an approach is that long documents tend to have higher scores because the probability of sharing terms with a query is higher. The presence of the NDL in Equation 1 is aimed at smoothing such an effect by reducing the contribution of terms belonging to longer texts.

The evaluation of the retrieval performance can be made through several measures, but none of them provides an exhaustive description of the retrieval results [BYRN99]. Moreover, depending on the application, some measures can be more appropriate than others. Given a query $q$, the set of the documents relevant to it is $R(q)$ and the set of the documents identified as relevant by the system is $R^*(q)$. The two fundamental measures in IR are *Precision*:

$$\pi(q) = \frac{|R(q) \cap R^*(q)|}{|R^*(q)|} \tag{3}$$

and *Recall*:

$$\rho(q) = \frac{|R(q) \cap R^*(q)|}{|R(q)|}. \tag{4}$$

Precision can be considered as the probability that a document identified as relevant by the system is actually relevant, while Recall can be thought of as the probability of a relevant document being identified as such by the system. The value of $\pi(q)$ is often calculated in correspondence of a predefined set of $\rho$ values (typically 10, 20,. . ., 100 percent) resulting in the so called *Precision vs Recall* curves. In order to obtain such a curve for a query set rather than for a single query, it is possible to perform a *macroaverage*, i.e. for each predefined value of $\rho$ the plotted Precision is the average of the $\pi$ values obtained for different queries:

$$\pi^M = \frac{1}{|T|} \sum_{q \in T} \pi(q) \tag{5}$$

where $T$ is the query set.

## 4 Experiments and Results

**Experimental setup :** The experiments performed in this work are based on a corpus of slides collected at a workshop held in June 2004 (Machine Learning in Multimodal Interfaces [BB05]). The slide authors were not aware of our experiments and they prepared their slides without any constraint. This data set is thus realistic with respect to similar situations. The corpus is composed of 26 presentations for a total of 570 slides. All of the slides have been acquired with a framegrabber (i.e. a device capturing the images displayed through a projector) resulting into 570 jpg images of dimension 1036×776 pixels (91.2 dpi resolution). The text contained in the presentations has been transcribed in three different ways. The first is by manually typing the content of the slides. This version, used as reference, will be referred to as *manual*). The second is by applying the different versions of the OCR system described in section 2 to the slide images (the transcriptions will be referred to as *Trans2*, *TransAll* and *TransBest*). The third is by using softwares converting the electronic versions of the slides (i.e. the Powerpoint or PDF files) into text (this version will be referred to as *API*).

**OCR performance evaluation:** the quality of the OCR transcripts can be evaluated using the *term recall* $TR$ and *term precision* $TP$ performance measures, defined by:

$$TR(d) = \frac{\sum_t \min\left(tf^\star(t,d), tf(t,d)\right)}{\sum_t tf(t,d)} \tag{6}$$

| OCR method | slide | | presentation | | database | |
|---|---|---|---|---|---|---|
| | $TR$ | $TP$ | $TR$ | $TP$ | $TR$ | $TP$ |
| Trans2 | 72.4 | 77.3 | 67.5 | 77.0 | 71.4 | 77.4 |
| TransBest | 77.0 | 78.4 | 72.6 | 77.3 | 76.7 | 79.0 |
| TransAll | 80.9 | 65.5 | 76.2 | 62.3 | 80.8 | 62.0 |
| API | 81.1 | 89.5 | 75.3 | 87.1 | 80.8 | 88.9 |

Table 1: Term recall $TR$ and term precision $TP$ averaged other slides, presentations, or computed on the whole database.

and

$$TP(d) = \frac{\sum_t \min\left(tf^\star(t,d), tf(t,d)\right)}{\sum_t tf^\star(t,d)} \quad (7)$$

where $tf(t,d)$ denotes the number of times the term $t$ really appears in the document $d$ ($d$ will be either a slide, a presentation, or the whole database), and $tf^\star(t,d)$ denotes the number of times the term $t$ appears in the transcript of the document $d$. The term recall can be interpreted as the percentage of terms in the document that have been correctly recognized by the OCR, while the term precision indicates the proportion of recognized terms that are actually true.

Table 1 provides the average term recall and precision computed over either slides, presentations or on the whole database. The overall values are good, showing that around 3 out of 4 terms are correctly recognized by the OCR systems, which means an average of 25 correct terms per slide document. These numbers, however, hide a large recognition variance depending on the slide type. While slides containing plain text only usually have term recall above 85%, slides containing images, plots or screen shots have lower and more diverse $TR$ values. The comparison of the OCR performance with the API results shows that the difference between the two approaches are not so large overall. While the OCR transcriptions are noisier, as indicated by the lower term precision, the term recall of the best performing OCR is equivalent to the API one (cf Table 1). Still, as expected, the API and OCR systems have different behaviours. While the API is almost error-less on text slides, it misses most of the text on slides with images, diagrams or plots.

Finally, comparing the different OCR systems between each other, we can see that the standard approach consisting of binarizing the text image (Trans2) is not performing as well as the two other methods. For instance, the TransBest method improves significantly (by approx. 5%) the term recall with respect to the Trans2 OCR, without any degradation in the term precision measure. This demonstrates the validity of both the use of the multi-class strategy and the string selection scheme. Second, compared with the TransBest approach, the TransAll strategy further improves the term recall (by approx. 4%), but this is done at the expense of the term precision, which drops by around 16%, from 78% to 62%. This effect is understandable, as this method consists of adding complementary transcripts from different multi-class segmentations. A net effect is to produce longer transcripts (cf previous section) in which the additional terms (w.r.t. TransBest) are less reliable. However, as most of the

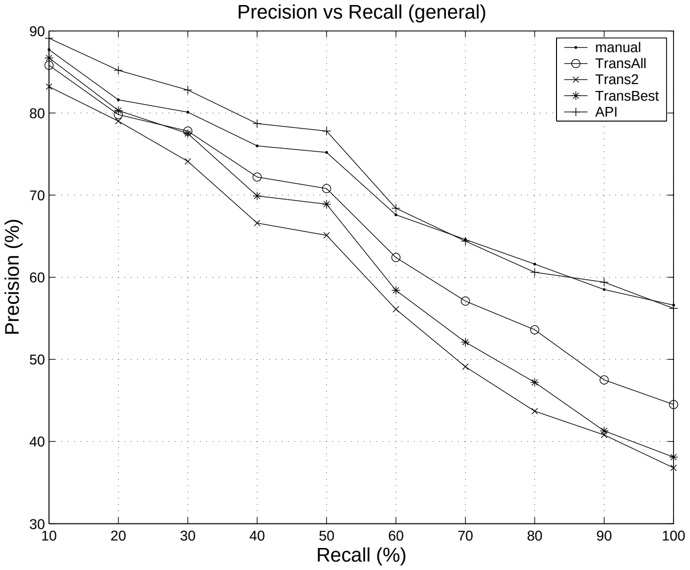

Figure 2: General. The plot shows the Precision vs Recall curves for the general task. The curves are reported for both manual and automatic transcriptions.

erroneous added terms do not correspond to true terms, and are not susceptible of being part of a query, their impact on the Retrieval Status Value of documents for a given query should be negligeable in principle. Hence, from a retrieval point-of-view, such a strategy should lead to better results.

**Retrieval tasks :** Two retrieval tasks have been performed that will be referred to as *general* and *image*. The general task is composed of 46 queries (containing 3.5 words on average) with 6.4 relevant documents on average. The general queries are expressed in natural language and they are submitted to find all of the documents about the topic they describe. The image task is composed of 85 queries expressed as a set of keywords (on average 3.6) and have one relevant document. The keywords have the particularity that they appear in a screen capture or a plot, but not in the actual text of the slide. The goal of such a task is to show that the OCR is capable of extracting the terms appearing in pictures, diagrams, plots and similar elements, while the API based systems can only access the text written as such on the slides. Image queries are submitted to find the slide containing the screen capture/plot/diagram the keywords are extracted from.

**Retrieval results :** Figure 2 reports the Precision vs Recall curves for the general task. The Precision achieved is higher on manual and API transcriptions (especially at high Recall) than on OCR based transcriptions. On the other hand, from a user point of view, such a difference does not require too much additional effort in order to find all of the relevant documents. At $\rho$=50 percent, the $\pi$ values range from 65.1 percent (Trans2) to 77.8 percent (API). Since the average number of relevant documents per query is 6.4, this means that the first 3 relevant documents can be found in the top 4 (API) to 5 (Trans2) positions. In other

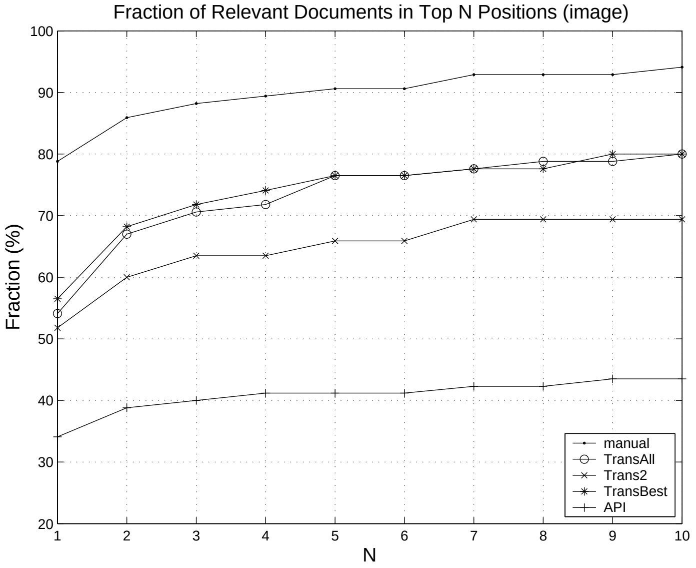

Figure 3: Fraction of relevant documents at position N. The plots reports the percentage of relevant documents appearing at the first N positions of the ranking.

words, in order to find half of the relevant documents, a user must browse, on average, four documents when using the API and manual transcriptions and five documents when using the OCR based transcriptions.The additional effort required to the user because of the recognition errors can thus be considered, in our opinion, acceptable.

The plots in Figure 3 show the results of the image task. Since the queries are supposed to retrieve only one document (see above), a suitable performance measure for this task is the percentage of relevant documents ranking in the first N positions. The curves can be also interpreted as the cumulative probability distributions of relevant documents ranking positions. The relevant document is at the top of the ranking around 80 percent of the times for the manual transcriptions, around 55 percent of the times for the OCR based transcriptions and around 35 percent of the times for the API. At the tenth position (i.e. at the end of the first results page in many IR system interfaces), the percentage of relevant documents rises to 94.1 percent, 80 percent and 43.5 percent for manual, TransAll and API transcriptions respectively. The OCR is thus almost two times more effective than the API based system in indexing the text contained in figures.

# 5 Conclusion

This work showed that it is possible to apply Information Retrieval technologies to slide transcriptions obtained with an OCR system. The use of an OCR rather than an API based commercial software has several motivations: the capture of the slide images through a framegrabber allows one to automatically link each slide to a presentation video segment (to retrieve slides is thus like to retrieve video segments); the use of slide images rather than electronic versions of the slides (ppt or pdf) allows the transcription system to be slide format independent. It is thus not necessary to access proprietary formats and to update the system each time the proprietary format is changed. Moreover, an OCR system can extract the texts embedded in figures (plots, diagrams, etc.) that are not accessible to API based commercial software.

The results show that the performance degradation due to the OCR errors is moderate. On average, half of the relevant documents can be found in the first four positions when using API based software and in the first 5 positions when using OCR. In other words, the additional effort required to the user because of the recognition errors is negligible. Since each slide is linked to a video segment, the approach we propose allows one to retrieve the parts of the oral presentations that are relevant to specific information needs.

The only information used so far for indexing is text, but slides contain many other important information sources: images, plots, diagrams, tables, etc.. The current system can thus be further improved by developing indexing approaches involving such elements. Our future work will focus in such a direction.

# References

[BB05] H. Bourlard and S. Bengio, editors. *Machine Learning for Multimodal Interaction: First International Workshop, MLMI'2004*, volume 3361 of *Lecture Notes in Computer Science*. Springer Verlag, 2005.

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[COB04] D. Chen, J-M. Odobez, and H. Bourlard. Text Detection and Recognition in Images and Videos. *Pattern Recognition*, 37(3):595–609, March 2004.

[LDK00] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital videos. *IEEE Trans. Image Processing*, 9(1):47–156, 2000.

[LW02] R. Lienhart and A. Wernicke. Localizing and segmenting text in images and videos. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(4):256–268, 2002.

[MS99] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. In *Proceedings of ACM International Conference on Multimedia*, pages 477–487, 1999.

[RWB00] S.E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management*, 36:95–108, 2000.