Project no. FP6-507752

MUSCLE

Network of Excellence
Multimedia Understanding through Semantics, Computation and Learning

DN4.1:
Report on Benchmark-Based Evaluations

Due date of deliverable: 30.11.2006
Actual submission date: 28.02.2007

Start date of project: 1 March 2004 Duration: 48 months

Deliverable Type: P
Number: D 4.1
Nature: Report
Task: WP 4

Name of responsible:
Andreas Rauber, TU Vienna-IFS, Austria

Revision 1.0

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	×
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contents

Introduction

T. Lidy[1], A. Rauber[1], J. Louradour[2], K. Daoudi[2], E. Benetos[3], M. Kotti[3], C. Kotropoulos[3], S. Aksoy[4], P. Duygulu[4], G. Akçay[4], E. Ataer[4], M. Baştan[4], T. Can[4], Ö. Çavuş[4], E. Doğrusöz[4], D. Gökalp[4], A. Akaydın[4], L. Akoğlu[4], P. Angın[4], G. Cinbiş[4], T. Gür[4], M. Ünlü[4], O. Ferret[5], R. Besancon[5], G. Grefenstette[5]

[1] Vienna University of Technology (TU Vienna-IFS)
[2] Institut de Recherche en Informatique de Toulouse (IRIT)
[3] Aristotle University of Thessaloniki (AUTH)
[4] Bilkent University (Bilkent)
[5] Commissariat à l'Energie Atomique (CEA LIST)

MUSCLE, the Network of Excellence on Multimedia Understanding through Semantics, Computation and Learning consolidates research groups that are exploring methods for knowledge extraction from multimedia data. Research within work-package 4 focusses on processing of text and audio data (including speech, sound and music). The aim of this work-package is to extract semantic concepts from these modalities which are then utilized for knowledge extraction, for instance in topic or genre recognition tasks. The extracted numerical data from the different modalities can be used individually, or in a combined manner, which is of particular value to the Cross-Modal Integration research work of work-package 5.

Comparability of results is a key issue in many disciplines and has proven to be extremely helpful for both collaboration and competition among research groups. Benchmarking corpora are clearly essential in order to devise well-defined tasks on which researchers can test their algorithms. In this context, we can define a corpus as an annotated collection of files, documents, or digital objects, where the annotations represent the criteria the algorithms will be evaluated against.

The evaluation of the effectiveness of MUSCLE techniques for knowledge extraction is an important factor for further progress. The plethora of tasks and approaches make clear the necessity for a suitable means of comparing results amongst researchers across institutional backgrounds and domains. While MUSCLE internal data sets are provided within the work-package 2 on Evaluation, Integration and Standards, the scientific research communities provide well-defined and well-known test corpora for mutual comparison of algorithms. It is very important for MUSCLE to participate in these international scientific benchmarking forums and to compare and collaborate with the research community outside MUSCLE. Besides evaluation of the devised methods on common standard benchmark corpora, these benchmark-based evaluations moreover allow a comparison of the novel methods to international state-of-the-art concepts.

This report on benchmark-based evaluations within MUSCLE describes the promising results that have been achieved by MUSCLE teams in scientific benchmarking campaigns. The report is organized in two parts: Part I includes evaluations of the methods for the different modalities of audio and Part II describes the participation in benchmarking campaigns that use textual information as input.

In Part I, Chapter 1 reports on the participation of a team from IRIT in the annual Speaker Recognition Evaluation of the US National Institute of Standards and Technology (NIST). Chapter 2 describes the achievements of TU Vienna-IFS at different tasks within the Music Information Retrieval Evaluation eXchange (MIREX) 2006, including Music Similarity Retrieval. Chapter 3 comprises a collaborative benchmark study on musical instrument classification of AUTH and TU Vienna-IFS.

In Part II, Chapter 4 describes the participation of a team from Bilkent University in the TRECVID 2006 campaign, in which they besides low-level video features also used textual speech transcripts for video classification and retrieval tasks. Eventually, Chapter 5 reports about the participation and achievements of the CEA LIST group at the Cross Language Evaluation Forum (CLEF) 2006.

Part I

Audio

Chapter 1

IRIT Participation in NIST Speaker Recognition Evaluation

Jerome Louradour, Khalid Daoudi
Institut de Recherche en Informatique de Toulouse (IRIT)

1.1 Introduction

NIST[1] is a non-regulatory federal agency within the U.S. Commerce Department's Technology Administration, whose mission is to promote innovation and industrial competitiveness by advancing measurement science, standards, and technology. Since 1997 [1], NIST Speech Group has organized annual Speaker Recognition Evaluations (NIST SRE). These evaluations are intended to be of interest to all researchers working on the general problem of speaker recognition. They have gained more and more interest in the last few years. To give an idea, 24 sites[2] participated in 2004 and the number of participants increased to 39 in 2006.

In the 2005 evaluations spontaneous and conversational telephone speech data was used, collected for the Mixer Corpus by the LDC[3], plus some multichannel data collected from several kinds of microphones. Although the data is mostly English speech, it includes some speech from additional languages, such as Spanish or Arabic, with bilingual speakers. In fact, NIST databases involve several strong sources of variability, so that speaker recognition systems are evaluated under realistic conditions of use.

[1]NIST: National Institute of Standards and Technology. `http://www.nist.gov`
[2]a site may be a gathering of several laboratories working in collaboration
[3]LDC: Linguistic Data Consortium

One of the main challenges is to handle the mismatch in recording conditions, as speakers may speak landline. Furthermore, the recording protocol imposes the conversation's theme to speakers, with a wide range of themes so as to guarantee a particularly high intraspeaker variability (related to semantic content and to emotion).

1.2 NIST Speaker Recognition Evaluation

1.2.1 The Speaker Verification Task

The reference task of NIST SRE is text-independent speaker verification, where the goal is to determine whether a sequence was pronounced or not by a "target speaker", without any constraint on the content.

IRIT took part in NIST SRE 2004 and 2005. Both years, the proposed systems showed competitive performance in the core test condition.

Unfolding of the Evaluation

NIST sends to all sites the same evaluation database and the same list of "trials" to perform. Each trial is a match between a target speaker and a test sequence, and participants cannot listen if the test sequence was pronounced by the target speaker. For each trials, they have to return a score and a binary decision. This decision has to be taken by comparing the score to a single threshold. The delay to give all the results (without knowing the answers) is about three weeks. So runtimes are not a criterion to evaluate the systems and consequently they are not mentioned during the result presentation at the NIST SRE workshop.

In the "core test" condition (required for all participants), every sequence is extracted from a 5 minutes long conversation and contains roughly 2 minutes of speech pronounced by the same speaker. Besides, only one sequence is avalaible to characterize each target speaker (one "training segment"). Laboratories may participate in other conditions, with more or less data to train speaker models, and with shorter or longer test segments. For instance, NIST SRE proposed 28 several conditions in 2004, and 15 several conditions in 2006. Unfortunately, relatively few laboratories have participated in these non-standard conditions up to now.

Evaluation Criterion

In order to evaluate and rank the submitted systems, performance is measured using a given Detection Cost Function (DCF) that has to be minimized.

This criterion is a combination of False Rejection (FR) and False Alarm (FA) rates:

$$\mathrm{DCF} = \underbrace{P_{Target}\boldsymbol{\tau}_{FR}}_{0.1}\mathrm{FR}\% + \underbrace{(1 - P_{Target})\boldsymbol{\tau}_{FA}}_{0.99}\mathrm{FA}\% \tag{1.1}$$

$P_{Target} = 0.01$ is the *a priori* probability of the specified target speaker. Parameters $\boldsymbol{\tau}_{FR} = 10$ and $\boldsymbol{\tau}_{FA} = 1$ are the relative costs of detection errors.

Besides, the overall relevance of output scores is also estimated with DET curves (Figs. 1.1 and 1.2), which represent the Detection Error Trade-off between the two types of detection errors (FA and FR) for any decision threshold, in a normal scale. When considering a single decision threshold as it is required in NIST evaluation, the DCF operating point belongs to the DET curve.

Note that in Figs. 1.1 and 1.2, also the approximative Equal Error Rates (EER) are mentioned. It corresponds to the point of the DET curve where FR% = FA%. Although the EER is not a suitable performance measure for fair comparison, it gives an idea of the classification error rates (if the fact that DET curves are roughly parallel is considered).

1.3 IRIT Participation 2004

The system presented in 2004 was based on a generative approach. It was derived from the state-of-the-art "UBM-GMM" systems, whose principle is well described in [3].

1.3.1 System Description

In this approach, the Universal Background Model (UBM) is trained offline on a large database involving a population of potential "impostors" (speakers different from the target speakers). In our system, two gender-dependent UBM were estimated using hours of speech from the NIST 2001 SRE database. Each GMM consists of 512 Gaussian components, and model parameters were learned using a tree-based VQ algorithm followed by EM iterations.

For each target speaker, a GMM is derived from the gender-corresponding UBM by adapting only mean vectors of Gaussian components, with a MAP criterion.

The output score is an average log-likelihood ratio (target speaker GMM / UBM), normalized using score statistics of the test sequence on some impostor GMM. This score normalisation is commonly called T-Norm.

The novelty of the submitted system lies in the scoring procedure, before the T-Norm. In order to improve performance, more importance was given to transitory zones of the speech for the computation of the non-normalized score. This trick to improve performance was motivated from a previous study, where IRIT showed that periods of transition between phones have a higher speaker-discriminative power [5]. These transitory zones were located using an efficient unsupervised segmentation algorithm of the speech signal in sub-phonetic units [4].

1.3.2 Results and Conclusions

Fig. 1.1 presents the results of all participants in NIST 2004 SRE: the IRIT DET curve is the black one. Even if the IRIT rank was 15 (out of 24 participants), and if there was a discernible gap between IRIT and the best systems, there are two facts to be considered when comparing the various systems:

1. NIST does not provide development database to the sites. The best participants have access to large databases from the LDC to tune their system.

2. The system presented by IRIT is only based on classical acoustic features, as the main motivation was to integrate a phonetic knowledge for a new scoring strategy. The best systems are in fact fusion of several baseline systems, each one using a particular aspect of speech: spectral, prosodic, phonetic and idiolectal features. The system presented by IRIT is only based on low-level spectral features.

1.4 IRIT Participation in 2005

The IRIT system presented in 2005 was based on a purely discriminative SVM approach, i.e. without any explicit probabilistic modeling.

1.4.1 System Description

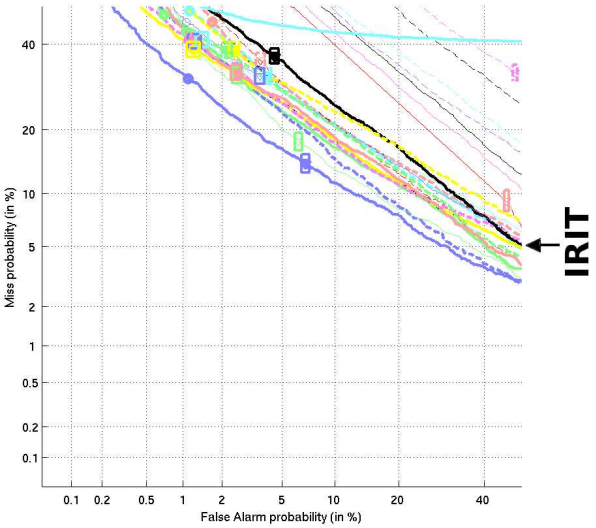
The key contribution of this system is the proposition of a novel sequence kernel for speaker verification [6, 7]. This kernel is a generalization of an efficient kernel in speaker verification: the GLDS kernel [8].

The basic principle of the new sequence kernel is to

- project all acoustic vectors in a high-dimensional vectorial space so as to make the data more separable,

	EER (%)	**DCF** ($\times 10^{-3}$)
best site	10	45.76
IRIT	18	81.47
worst site*	27	214.5

* without claimed bug



DET curves

Figure 1.1: IRIT position in NIST 2004 SRE

- then processing a normalization in this feature space so as to induce a Mahalanobis distance and make the SVM training algorithm more stable, and finally

- compute the average of dot products between all inter-sequence pairs of frames.

These steps are done implicitely to compute the kernel value, in an efficient way, with a technique which allows to compact information in a suitable way: the Incomplete Cholesky Decomposition.

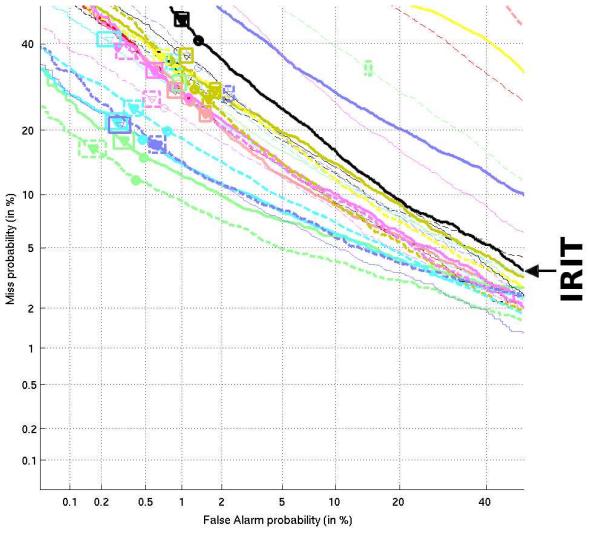
Finally, the sequence kernel is used to train target speaker SVM models using one-against-all scheme.

1.4.2 Results and Conclusions

Fig. 1.2 show the results of the new baseline IRIT system in NIST 2005 SRE. Results are promising, taking into account the aforementioned remarks on development databases and on system architecture.

	EER (%)	**DCF** ($\times 10^{-3}$)
best site	5	18.20
IRIT	12	56.53
worst site*	24	111.9

* without claimed bug



DET curves

Figure 1.2: IRIT position in NIST 2005 SRE

Bibliography

[1] A. Martin and M. Przybocki. "The 1997 speaker recognition plan," `http://www.nist.gov/speech/tests/spk/1997/sp_v1p1.htm`, 1997.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech*, 1997.

[3] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, 2000.

[4] R. Andre-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 36(1), 1988

[5] J. Louradour, K. Daoudi, and R. André-Obrecht, "Discriminative power of transient frames in speaker recognition," in *Proceedings of the 13th International Conference of Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[6] J. Louradour and K. Daoudi, "Conceiving a new sequence kernel and applying it to SVM speaker verification ," in *Proceedings 9th European Conference on Speech Communication and Technology (INTERSPEECH)*, 2005.

[7] J. Louradour, K. Daoudi, and F. Bach, "SVM speaker verification using an incomplete cholesky decomposition sequence kernel," in *Proceedings of IEEE Odyssey: The Speaker and Language Recognition Workshop*, 2006.

[8] W.M. Campbell, J.P. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

Chapter 2

TU Vienna-IFS Participation in MIREX 2006

Thomas Lidy, Andreas Rauber
Vienna University of Technology (TU Vienna-IFS)

2.1 Introduction

The Music Information Retrieval Evaluation eXchange (MIREX) is the annual scientific benchmarking forum held in conjunction with the International Conference on Music Information Retrieval (ISMIR). The MIREX campaign started in 2005, after several years of thorough discussions of how to realize a scientific evaluation in order to enable comparison of the many different algorithms and systems which are developed in the Music Information Retrieval research domain, and a pre-MIREX "Audio Description Contest" held at ISMIR 2004. MIREX is partitioned in a number of different tasks, such as Melody Extraction, Drum Detection, Genre Classification, Artist Identification, etc. The number and definition of tasks is adjusted annually, according to proposals of participants and/or new challenges that the Music Information Retrieval research domain is faced with. The MIREX forum is open to any individual or organization who wishes to participate.

The MIREX evaluations actively stimulate research and foster exchange between different research teams and enable a comparison of state-of-the-art techniques for individual Music Information Retrieval tasks.

2.2 MIREX 2006 Tasks

In MIREX 2006 eight different tasks were available to participate. TU Vienna-IFS participated in two of these tasks: Audio Music Similarity and Retrieval and Audio Cover Song Identification. The Audio Music Similarity and Retrieval task was to submit an audio feature extraction algorithm, to compute music similarity measures and to return a distance matrix from a music collection consisting of 5000 pieces, which was subsequently evaluated through human listening tests as well as objective statistics. TU Vienna-IFS submitted a new implementation of the Statistical Spectrum Descriptor (SSD) audio feature extractor and computed the distance matrix directly from feature space. Results from the human evaluation showed that the approach is among the top 5 algorithms which have no statistically significant differences. The evaluation of a number of objective statistics ranked the algorithm third in most of the cases. The TU Vienna-IFS submission was one of the two fastest in terms of total runtime, having the shortest distance computation time.

The approach has also been evaluated on Audio Cover Song Identification, where it was the best-performing "Audio Music Similarity and Retrieval" submission, outperformed however by 4 submissions which were specifically designed for the cover finding task.

2.3 Submitted Algorithm

A new implementation of Statistical Spectrum Descriptors (SSD) [2] was submitted to both MIREX tasks [3].

Audio files with 22.050 Hz sampling rate in mono format were provided in these MIREX 2006 tasks. After segmentation of an audio file into segments of 2^{17} samples (approx. 5.9 seconds), the first and the last segment were skipped, from the remaining segments, every third one was processed. A feature vector was then calculated for each of the remaining segments.

First, the spectrogram was computed with an FFT using a Hanning window with a size of 512 samples and 50 % overlap. Then, the spectrum was aggregated to 23 critical bands according to the Bark scale. The Bark-scale spectrogram was then transformed into the decibel scale and subsequently into the Sone scale, in order to approximate the loudness sensation of the human auditory system.

From this representation of a segment's spectrogram the following statistical moments are computed in order to describe fluctuations within the critical bands: mean, median, variance, skewness, kurtosis, min- and max-

value are computed for each critical band, forming the SSD feature set. The feature vector for an audio file is then constructed as the median of the SSD features of the extracted file segments.

The distance matrix was computed directly in the feature vector space using the cityblock metric as the measure of distance.

2.4 Audio Music Similarity and Retrieval

The task was to submit an audio feature extraction algorithm and subsequently compute music similarity measures from which a distance matrix should be produced, i.e. a matrix containing the distances between all pairs of music tracks in a music database[1]. Feature extraction algorithms, any models and their parameters had to be trained and optimized in advance without the use of any data which has been part of the MIREX test database. The music database comprised 5000 pieces of (Western) music from 9 genres in 22 kHz, mono, 16 bit Wave Audio format (including the tracks of the Audio Cover Song task - see Section 2.5 below). From the distance matrices, two forms of evaluations were performed: Evaluation based on human judgments and objective evaluation by statistic measures. Besides, the runtimes of the algorithms were recorded.

2.4.1 Human Evaluation

The primary evaluation focus of this MIREX 2006 task was on the judgements of the human evaluators. The human listening test was realized as follows:

60 songs were randomly selected as queries from the total of 5000 songs in the database. Each participating algorithm had to return the 5 most similar songs to the query (after filtering out the query itself, members of the cover song collection, as well as songs of the same artist as the query, in order to avoid the task to be an artist identification task). The results from all 6 participating algorithms then formed a list of 30 results per query, which had to be evaluated by human graders, who rated each retrieved song on two scales: one broad scale, stating whether the song is not, somewhat or very similar to the query song, and one fine-grained scale, where they had to score the retrieved songs on a real-value scale between 0 (not similar) and 10 (very similar). Each query/candidate list pair was evaluated by 3 different graders. 24 graders participated in the human evaluation, hence each person

[1] `http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval`

had to evaluate 7-8 query/candidate lists. The listening test was performed through the Evalutron 6000 web interface[2] created by the IMIRSEL team.

There were six participants in this task: Elias Pampalk (EP), Tim Pohle (TP), Vitor Soares (VS), Thomas Lidy & Andreas Rauber from TU Vienna-IFS (LR), Kris West - Transcription model (KWT), Kris West - Likelihood model (KWL).

From the human judgments both the fine-grained score and the broad scale have been evaluated: The score for the fine-grained scale has been computed as the mean of all human ratings. For the broad scale, several different scoring systems have been applied, with different weighting of the 'very similar' and/or 'somewhat similar' grades. A table with all the scores for these different measures is available from the MIREX 2006 Audio Similarity and Retrieval results page[3]. The 6 different scoring systems resulted in a consistent ordering of the submitted algorithms, also the fine-grained and the broad scale results were consistent. A significance test has been applied to the results of the human evaluation in order to determine whether they indicate significant differences between the performance of the algorithms. The Friedman test [1] was chosen because it is a non-parametric test which does not assume a normal distribution of the data. The Friedman test has been performed in Matlab with pairwise comparison of algorithms for each of the 60 queries, based on the fine-grained score. The results of the test at a confidence level of $p = 0.05$ showed that there are *no significant differences* between the top 5 algorithms (see Figure 2.1). Only the Likelihood algorithm by Kris West (KWL) performed significantly worse than three of the other algorithms. (The author however stated that there was a bug in his submissions.) As a consequence, there was no official ranking for this MIREX 2006 task.

2.4.2 Statistics

Computation of full distance matrices containing distances between all 5000 songs in the database enabled the computation of meta-data based statistics, such as: Average percentage of Genre, Artist and Album matches in the top 5, 10, 20 and 50 results, before and after artist filtering, Normalized average distance between examples of the same Genre, Artist or Album, Ratio of the average artist distance to the average genre distance, Number of times a song was similar to any of the 5000 queries, i.e. revealing songs that are always similar or never similar, Confusion Matrices, and more. One submission

[2] http://www.music-ir.org/evaluation/eval6000/

[3] http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval_Results

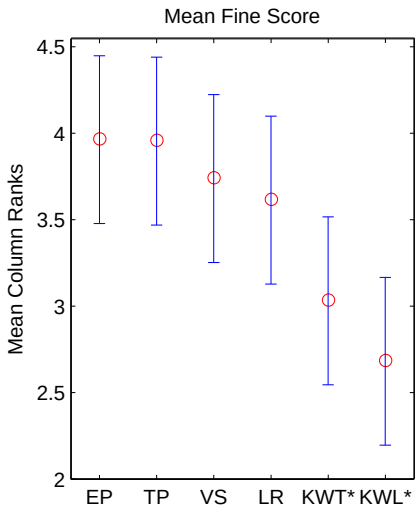
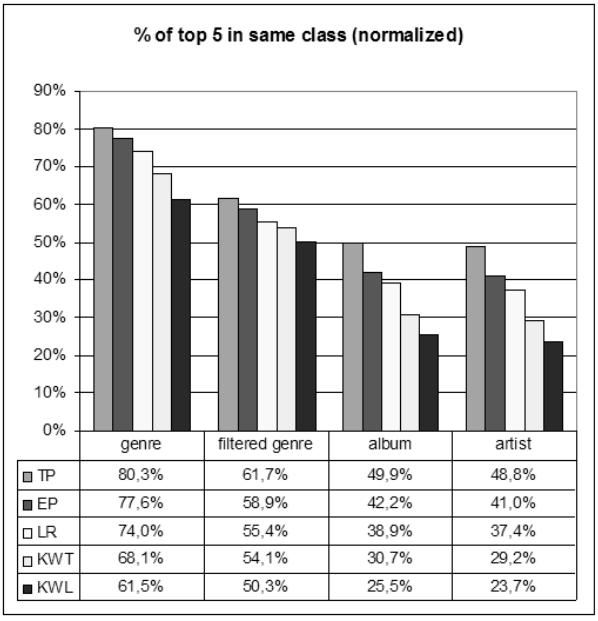


Figure 2.1: MIREX 2006 results from human listening tests, using the Friedman test. Circles mark the mean of the fine-grained human similarity scores, the lines depict the significance bounds at a level of p = 0.05.

(Vitor Soares, VS) has not been evaluated through these statistics, because the algorithm was not able to compute the full 5000 × 5000 distance matrix within the maximum time allowed for this MIREX 2006 task, which was 36 hours.

The results of this evaluation should be considered with caution, as the genre distribution in the music database was highly skewed: 50 % of the data was Rock music, 26.6 % Rap & Hip-Hop, 9.7 % Electronica & Dance, 5.3 % Country music and the remaining genres (Reggae, New Age, R & B, Latin and Jazz) were represented by 2 % or less, each. "Similar" songs, however, do not necessarily have the same genre label. This might be the reason why the ordering of the results from these statistics partly differs from the one of human listening results.

Figures 2.2 and 2.3 present the results of the percentages of how many within the retrieved 5 respectively 20 most similar songs had the same genre, artist or album as the query song. The numbers have been computed excluding the 330 cover songs and considering normalization for genres, artists or albums with less than 20 matches available in the database. The genre statistic is given before and after filtering out the query artist. The measurement of artist-filtered statistics is important, because many algorithms detect songs from the same artist as the most similar songs and unfiltered results evaluate mainly the capability of algorithms to identify artists. Further statistics for the top 10 and top 50 results are available from the Audio



	genre	filtered genre	album	artist
TP	80,3%	61,7%	49,9%	48,8%
EP	77,6%	58,9%	42,2%	41,0%
LR	74,0%	55,4%	38,9%	37,4%
KWT	68,1%	54,1%	30,7%	29,2%
KWL	61,5%	50,3%	25,5%	23,7%

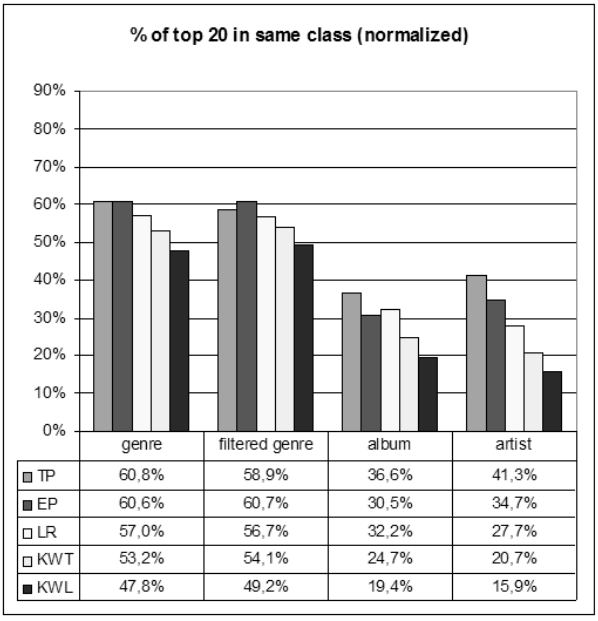
Figure 2.2: MIREX 2006 Audio Music Similarity and Retrieval: Average percentage of Genre (before and after artist filtering), Artist and Album matches in the *top 5* query results (normalized).

Music Similarity and Retrieval Statistics result web page[4]. In most of the cases the algorithm of TU Vienna-IFS was ranked third, with a result of 74 % in a 5-nearest-neighbor-like genre recognition task. Considering the percentage of top 20 album matches the algorithm was ranked second (c.f. Figure 2.3). The changing order of result ranking seems to be an indication of the non-significant differences between the algorithms as revealed by the human evaluation.

2.4.3 Runtimes

Computation times have been recorded individually for audio feature extraction and distance computation (except for the KWL model, where only the total time could be recorded). The runtimes were measured on Dual AMD Opteron 64 computers with 1.6 GHz and 4 GB RAM, running Linux (CentOS). The runtime of Soares' algorithm (VS) is not part of this comparison as it did not compute the full distance matrix. Pampalk's algorithm (EP) was the fastest in total (3 hours, 19 minutes) closely followed by the one of TU

[4] `http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval_Other_Automatic_Evaluation_Results`



	genre	filtered genre	album	artist
TP	60,8%	58,9%	36,6%	41,3%
EP	60,6%	60,7%	30,5%	34,7%
LR	57,0%	56,7%	32,2%	27,7%
KWT	53,2%	54,1%	24,7%	20,7%
KWL	47,8%	49,2%	19,4%	15,9%

Figure 2.3: MIREX 2006 Audio Music Similarity and Retrieval: Average percentage of Genre (before and after artist filtering), Artist and Album matches in the *top 20* query results (normalized).

Vienna-IFS (LR) (3 hours, 52 minutes) – c.f. Figure 2.4. The TU Vienna-IFS algorithm was by far the fastest one in distance matrix computation (2 minutes only), which is due to the direct computation of distances in feature space using a simple distance metric, namely the Cityblock metric. Other algorithms needed a factor of 25 to 193 more time for distance computation. The total runtime of the slowest participating algorithm was about 4 times the runtime of TU Vienna's.

2.5 Audio Cover Song Identification

The cover song database consisted of 30 different "cover songs" each represented by 11 different "versions", hence a total of 330 audio files. The cover songs represent a variety of genres (e.g., classical, jazz, gospel, rock, folk-rock, etc.) and the variations span a variety of styles and orchestrations.

Each of these cover song files has been used as a query and the top 10 returned items have been examined for the presence of the other 10 versions of the query file[5]. The 330 cover songs have been embedded within the 5000

[5] `http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song`

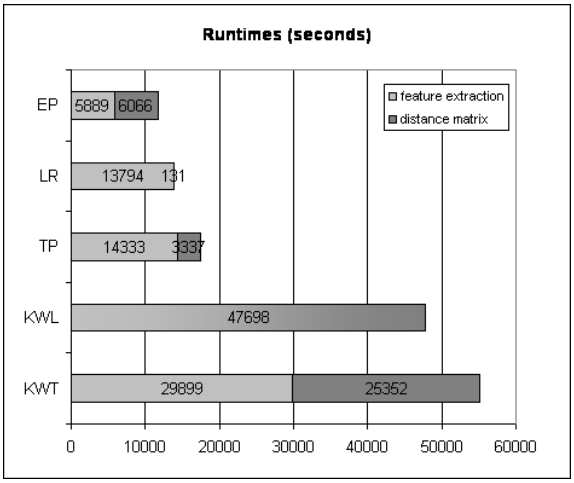


Figure 2.4: MIREX 2006: Runtimes of Audio Music Similarity algorithms in seconds (audio feature extraction and distance matrix computation).

songs database used for the Audio Music Similarity and Retrieval task which enabled an evaluation of the Similarity algorithms for the Cover Song task without any extra effort except for retrieving the cover song queries from the distance matrices. For the evaluation of the Cover Song task, however, a reduced data set of 1000 songs has been used to accommodate more complex systems which have been particularly designed and submitted for cover song identification.

There were four submissions with systems which have been particularly designed for cover song identification – Dan Ellis (DE), Christian Sailer & Karin Dressler (CS), Kyogu Lee (KL, 2 models) – and four systems which have been evaluated as by-product of the Audio Music Similarity and Retrieval task (TP, LR, KWT and KWL – see Section 2.4).

The total number of correctly identified cover songs – out of the 3300 potentially detectable covers – is depicted in Figure 2.5. It can be seen from the results in the figure, that the submission by TU Vienna-IFS (denoted 'LR') was the best-performing "Audio Music Similarity and Retrieval" algorithm, outperformed however by the four specific cover song identification systems. Further measures – the mean number of covers identified, the mean of maxima (average of best-case performance) and the mean reciprocal rank of the first correctly identified cover (MRR) – are provided in a table on the Audio Cover Song Identification web page[6]. A Friedman test has been run against the MRR measure and identified Ellis' system (DE) as the clear

[6] `http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song_Identification_Results`

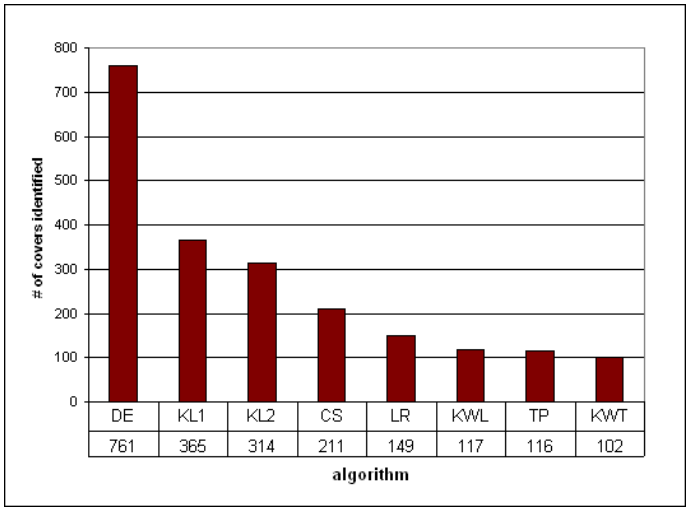


Figure 2.5: MIREX 2006 Audio Cover Song Identification results (total number of identified cover songs). Algorithms marked with * were specifically designed for the Cover Song Identification task.

winner of this task, while there was no significant difference between the 7 other algorithms.

2.6 Conclusions

The first large-scale human listening test for Music Similarity and Retrieval in MIREX showed, that the algorithm developed by TU Vienna-IFS is competing with state-of-the-art algorithms – no significant difference in performance was determined between the top 5 algorithms. It is also one of the two fastest algorithms, with by far the most efficient distance calculation. Different statistics have been derived from genre, artist and album assignments, which gave the algorithm the third rank in most of the cases, and second rank in one case.

The algorithm was also evaluated on Audio Cover Song Identification together with three of the other Audio Music Similarity and Retrieval submissions and four submissions specifically designed for finding cover songs. It was the best on identifying covers out of the four Similarity algorithms, outperformed by the four specific Cover Song algorithms.

Bibliography

[1] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, December 1937.

[2] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proceedings of the Sixth International Conference on Music Information Retrieval*, pages 34–41, London, UK, September 11-15 2005.

[3] Thomas Lidy and Andreas Rauber. Computing statistical spectrum descriptors for audio music similarity and retrieval. In *MIREX 2006*, Victoria, Canada, October 8-12 2006.

Chapter 3

AUTH and TU Vienna-IFS Benchmark Study on Musical Instrument Classification

Emmanouil Benetos, Margarita Kotti, Constantine Kotropoulos
Aristotle University of Thessaloniki (AUTH)

Thomas Lidy, Andreas Rauber
Vienna University of Technology (TU Vienna-IFS)

3.1 Introduction

The need for analyzing musical content arises in different contexts. It has many practical applications, mainly for automatic music transcription, efficient data organization and search, and multimedia databases annotation. Automatic musical instrument classification is a fundamental step towards developing such applications. It is a research area which can also be applied to general sound recognition tasks. Challenges differ from the ones in automatic speech and speaker recognition areas. The problems addressed so far in musical content identification include classification of isolated instrument tones and classification of sound segments. Classifiers using isolated tones have a limited use in practical applications, while sound segment classifiers are of importance in a range of music information retrieval (MIR) systems.

This collaborative study investigates the use of several classification techniques and different sets of features extracted from audio data for effective musical instrument classification. In particular, a specific method for supervised classification based on non-negative matrix factorization techniques is proposed.

3.2 Musical Instrument Classification

3.2.1 Data Sets

In this work, the problem of automatic recognition of musical instrument segments is addressed. In the first set of experiments, 300 recordings from the University of Iowa (UIOWA) [1] database were used that form 6 instrument classes, i.e., bassoon, cello, flute, piano, soprano saxophone, and violin. In the second set of experiments, 20 musical instrument classes were employed in total. The number of recordings was raised to 1000. The additional instruments featured were: alto flute, alto saxophone, double bass, bass clarinet, bass flute, bass trombone, B♭ clarinet, E♭ clarinet, horn, oboe, tenor trombone, tuba, and viola. All recordings have a duration of about 20 seconds and are sampled at 44.1 kHz rate.

3.2.2 Experiments

In the first set of experiments, two distinct feature sets were extracted from the audio of the music instruments: the first one consisted of a combination of features originating from general audio data classification experiments and the MPEG-7 audio framework [2]. The second feature set used was Rhythm Pattern features developed within the SOMeJB project of TU Vienna-IFS, offering a time-invariant representation of fluctuation patterns on critical bands according to perception of the human auditory system [5]. In the second set of experiments, a sound description toolbox was developed, covering temporal, spectral, energy, perceptual, and harmonic audio properties.

For the first set of experiments, a supervised classifier was proposed, based on non-negative matrix factorization techniques. Non-negative matrix factorization (NMF) is a subspace method able to obtain a parts-based representation on objects by imposing non-negative constraints [3]. Given a $m \times n$ matrix $\mathbf{V}$, the goal is to approximate it as a product of the $m \times r$ matrix $\mathbf{W}$ (called basis matrix) and the $r \times n$ matrix $\mathbf{H}$ (called encoding matrix). Three NMF variants were examined, namely the standard, local and sparse NMF [4, 6]. A supervised classifier based on NMF techniques was proposed, which performs training on each class individually. The test data is afterwards projected onto each class basis matrix and the class label for the test recording is assigned to the class that maximizes the cosine similarity measure (CSM) of the class encoding matrix and the test encoding vector. For the second set of experiments, the supervised classifier performs Gram-Schmidt orthonormalization on the class basis matrices, thus projecting the test data on an orthogonal basis. In addition, multilayer perceptrons, radial basis func-

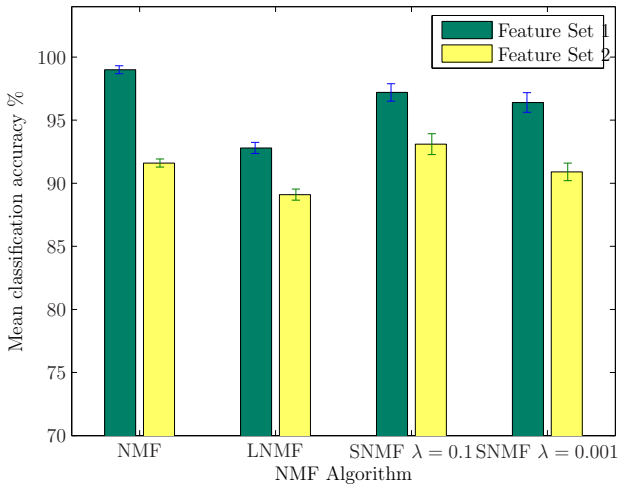


Figure 3.1: Mean classification results for the first set of experiments.

tion networks and and support vector machines have also been employed as classifiers for the expanded experiments.

Feature selection was applied to the feature sets used in both experiments, in order to select the feature subset that maximizes classification accuracy. The branch-and-bound search strategy with depth-first-search was employed. The first set of experiments was performed using 7-fold cross validation and the mean value of the classification accuracy and its standard deviation for the three NMF algorithms and the two feature sets. The second set of experiments was performed using 3-fold cross validation.

3.2.3 Results and Conclusions

is . and the results are

The results of the first set experiments, shown in Figure 3.1, indicate a high accuracy for the NMF and the Sparse NMF (SNMF) algorithms for the 1st feature set. The results outperform supervised classifiers based on hidden Markov models (HMMs) and Gaussian mixture models (GMMs) tested in [7].

The results of the second set of experiments, depicted in Figure 3.2, indicate that the SNMF outperforms all other NMF variants, as well as the neural network classifiers. Additional information about the performance of the SNMF algorithm is shown in the form of a confusion matrix (Table3.1), where it can be observed that most of the mis-classification occurs between instruments of the same family.

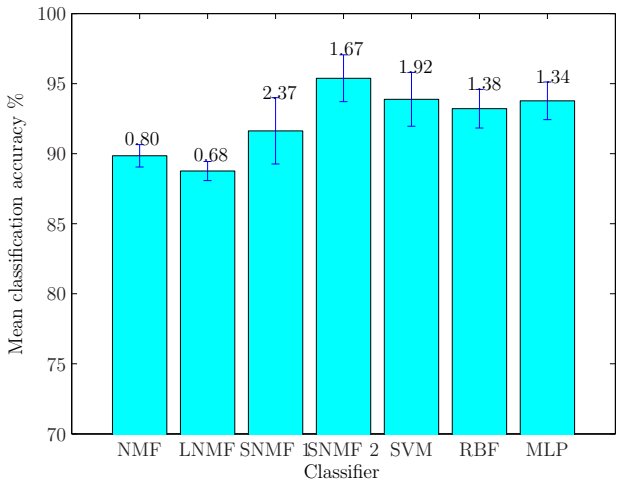


Figure 3.2: Mean classification results for the second set of experiments.

Table 3.1: Confusion matrix for a run of the SNMF ($\lambda = 0.001$) classifier.

Inst.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	**11**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	**10**	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
5	3	0	0	0	**7**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	**8**	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	1	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	**13**	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	**18**	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	1	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0	**13**	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**18**	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	**53**	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**39**	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**14**	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**15**

Bibliography

[1] University of Iowa Musical Instrument Sample Database, http://theremin.music.uiowa.edu/index.html.

[2] MPEG-7 overview (version 9), *ISO/IEC JTC1/SC29/WG11 N5525*, March 2003.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556-562, 2001.

[4] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-6, 2001.

[5] A. Rauber, E. Pampalk, and D. Merkl, "The SOM-enhanced JukeBox: organization and visualization of music collections based on perceptual models," *Journal of New Music Research*, Vol. 32, No. 2, pp. 193-210, June 2003.

[6] C. Hu, B. Zhang, S. Yan, Q. Yang, J. Yan, Z. Chen, and W. Ma, "Mining ratio rules via principal sparse non-negative matrix factorization," in *Proceedings of the IEEE International Conference on Data Mining*, 2004.

[7] E. Benetos, M. Kotti, C. Kotropoulos, J. J. Burred, G. Eisenberg, M. Haller, and T. Sikora, "Comparison of subspace analysis-based and statistical model-based algorithms for musical instrument classification," in *Proceedings of the 2nd Workshop on Immersive Communication and Broadcast Systems*, October 2005.

[8] E. Benetos, C. Kotropoulos, T. Lidy, and A. Rauber, "Testing supervised classifiers based on non-negative matrix factorization to musical instrument classification," in *Proceedings of the 14th European Signal Processing Conference*, September 2006.

Part II

Text

Chapter 4

Bilkent University at TRECVID 2006

S. Aksoy, P. Duygulu,
G. Akçay, E. Ataer, M. Baştan, T. Can, Ö. Çavuş, E. Doğrusöz, D. Gökalp, A. Akaydın, L. Akoğlu, P. Angın, G. Cinbiş, T. Gür, M. Ünlü

Department of Computer Engineering, Bilkent University

4.1 Introduction

This chapter describes the third participation of the RETINA Vision and Learning Group at Bilkent University to TRECVID. The team that participated to TRECVID included six undergraduate students and seven graduate students supervised by two faculty members and developed a system for automatic classification and indexing of video archives. This report summarizes the approaches that were submitted to TRECVID 2006, including one high-level feature extraction run as well as two manual and one interactive search runs. All of these runs have used a system trained on the common development collection. Only visual and textual information were used where visual information consisted of color, texture and edge-based low-level features and textual information consisted of the speech transcript provided in the corpus.

4.2 Preprocessing

In all of the runs, data provided with the TRECVID 2005 and 2006 corpora were used, such as shot boundaries, keyframes, speech transcripts and manual

annotation.

The speech transcripts were in free text form and required preprocessing. First, a part of speech tagger was used to extract nouns which are expected to correspond to object names. Then, a stemmer was applied and the stop words and also the least frequent words were removed to obtain a set of descriptive words. Finally, WordNet was applied in order to construct a hierarchy to extend the word set for each keyframe. For example, even when the word "sport" is not included in the speech transcript, if any kind of sport such as "basketball" or "soccer" appears in the speech transcript, that keyframe is also associated with the word "sport" to allow more flexibility during classification and retrieval.

Spatial content of images was modelled using grids. The low-level features based on color, texture and edge were computed individually on each grid cell of a non-overlapping partitioning of 352×240 video frames into 5 rows and 7 columns. Each resulting grid cell is associated with the statistics (mean and standard deviation) of RGB, HSV and LUV values of the corresponding pixels as the color features and the statistics of the Gabor wavelet responses of the pixels at 3 different scales and 4 different orientations as the texture features. Histograms of the gradient orientation values of the Canny edge detector outputs are used as the edge features. Orientation values are divided into bins with increments of 45 degrees and an extra bin is used to store the number of non-edge pixels.

This process results in 5 feature vectors for each grid cell with the following lengths: 6 for each of RGB, HSV and LUV statistics, 24 for Gabor statistics, and 9 for edge orientation histograms. Individual components of each feature vector are also normalized to unit variance to approximately equalize ranges of the features and make them have approximately the same effect in the computation of similarity.

4.3 High-Level Feature Extraction

The Bilkent team has developed a generic classifier that uses the low-level features described with the k-nearest neighbor rule. First, experiments were performed using different feature combinations and different k values on the TRECVID 2005 data. It was empirically decided to use tiled HSV histograms and Canny-based edge orientation histogram features. The k value was also chosen as 51. The examples provided with the common annotation were used to train the classifier for all high-level features.

It was observed that the resulting classifier was particularly effective for features such as weather, sport, studio setting, computer screen, and map.

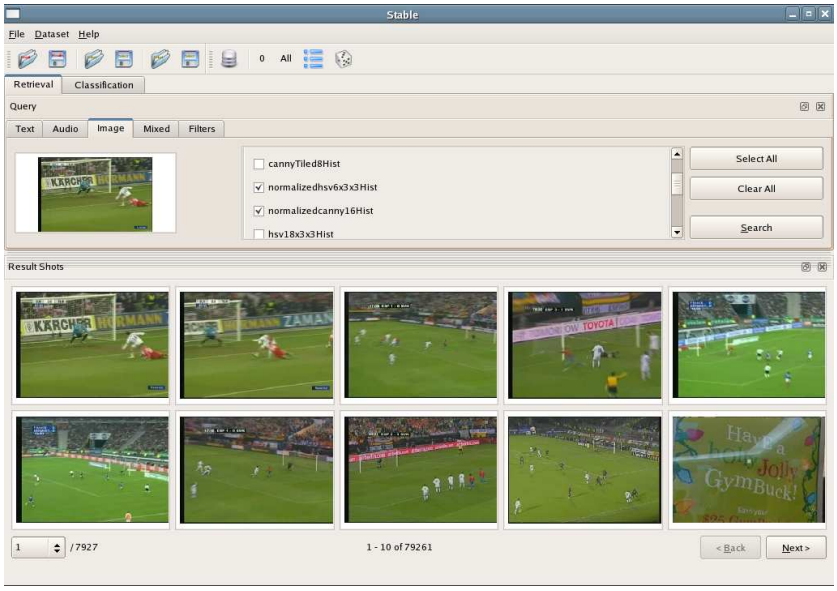


Figure 4.1: Manual search results using low-level color and edge-based features for a "soccer" query.

Evaluation of the classifier on TRECVID 2006 data also reflected this observation.

4.4 Search

The baseline manual run used processed ASR outputs. This run (Bilkent2) was particularly effective for topics that involve specific keywords such as "president", "Bush", "Chaney", "Rice", etc.

The second manual run used only low-level visual features that were also used in the high-level feature extraction task. The shots were sorted according to their distance to the query shot using these low-level features. This run (Bilkent1) was very effective for topics that have specific color content such as "soccer", "helicopter" (gray vehicle on blue/gray sky) and "flame". An example query is shown in Figure 4.1.

Finally, the interactive run (Bilkent3) that used both ASR keywords and low-level visual features combined the advantages of both features and gave satisfactory results for many more topics. Example queries are shown in Figures 4.2 and 4.3. These results show the potential of multi-modal analysis for future research.

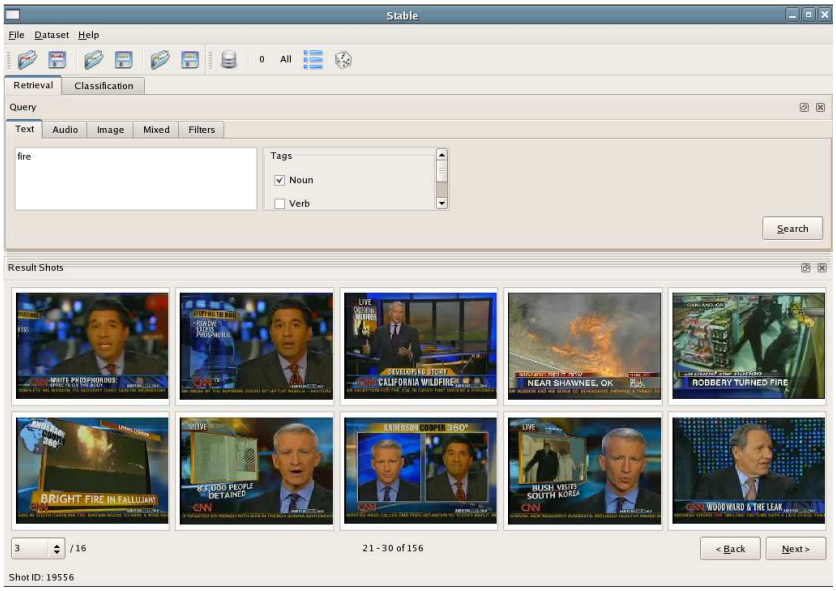


Figure 4.2: Interactive search results using the baseline (ASR-based) system for a "fire/flame" query.

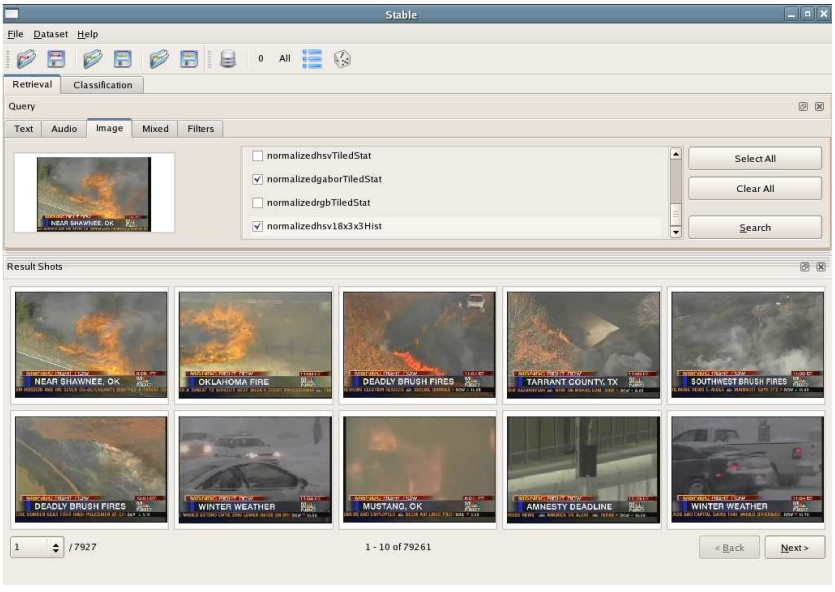


Figure 4.3: Interactive search results using low-level features for a "fire/flame" query.

4.5 Conclusions

The Bilkent participation to TRECVID 2006 consisted of one high-level feature extraction run, and two manual and one interactive search runs. The research group is currently working on extending the system with new low-level features, classifiers and novel methods for multi-modal fusion.

Chapter 5

CEA LIST Participation in CLEF 2006

Olivier Ferret, Romaric Besancon, Gregory Grefenstette
Commissariat à l'Energie Atomique (CEA LIST)

5.1 Introduction

CLEF stands for Cross Language Evaluation Forum. This forum runs text and image evaluation campaigns especially in cross lingual settings. Research in cross lingual information retrieval (CLIR) has been increasingly active over the last 10 years. The first CLIR campaigns were held within the Text Retrieval Conference TREC (trec.nist.gov) during the three years of campaigns 1997-1999. After those initial efforts, the European Union sponsored network of Excellence DELOS program sponsored the first European campaigns, called CLEF, starting in 2001. The CLEF campaigns are now held yearly, followed by CLEF conferences in various cities in Europe. Results from past campaigns can be found on the website: www.clef-campaign.org.

In the latest edition, CLEF 2006, the following tracks were held:

- **AdHoc** – multilingual document retrieval over a multilingual database of 2 million documents in 12 languages (DE,EN,ES,FI,FR,IT,NL,RU,SV, PT, BG and HU)
- **Question Answering** – providing specific answers over the same corpus (rather than just returning answer documents, the answer has to be extracted).
- **ImageCLEF** – retrieving images using a textual input query and a few sample images. Image databases included 28,000 historical images with captions [St Andrews collection], and three collections of over 50,000 annotated medical images.
- **WebClef** – 500 short queries in 11 languages over 2 million European government web pages.
- **GeoClef** – 25 geographical topics over the AdHoc collection.
- **Cross Language Speech Retrieval** – over 590 hours of continuous speech from the Shoah Archive.

The CEA LIST group from MUSCLE participated in the QA benchmark and ImageCLef combining both image treatment and text treatment as described below.

5.2 CEA LIST at ImageCLEF

The CEA LIST/LIC2M laboratory participated in ImageCLEF 2006 to perform experiments on merging strategies to integrate the results obtained from the cross-language text retrieval system and the content-based image retrieval (CBIR) system that are developed in our lab, using a simple merging strategy similar to the one used in previous ImageCLEF campaigns [1]. We use text and visual information from the queries: the title provided for the text retrieval system, and the example images for the CBIR system. Both systems are general-domain systems and are used independently on each part of the query. Then, a posteriori merging strategies are applied on the results provided by each system.

5.2.1 Retrieval systems

Both text retrieval system and CBIR systems are the same that we used in previous ImageCLEF campaigns [1]. The basic principles of the systems are presented here.

Multilingual Text Retrieval System

The multilingual text retrieval system has not been specially adapted to work on the text of the ImageCLEF corpora, and has simply been used as is: no special treatment has been performed to take into account the structure of the documents (such as title, description, location, date): all fields containing some text have been taken as is. The system works as follows:

Document and query processing: The documents and queries are processed through a linguistic analyzer that extracts relevant linguistic elements such as lemmas, named entities and compounds. The elements extracted from the documents are indexed into inverted files. The elements extracted from the queries are used as query "*concepts*". Each concept is reformulated into a set of search terms for each target language (in the case of imageClefPhoto, only one target language was used) either using a monolingual expansion dictionary (that introduces synonyms and related words), or using a bilingual dictionary.

Document Retrieval: Each search term is searched in the index, and documents containing the term are retrieved. All retrieved documents are then associated with a concept profile, indicating the presence of query concepts in the document. This concept profile depends on the query concepts, and is language-independent (which allow merging results from different languages). Documents sharing the same concept profile are clustered together, and a weight is associated with each cluster according to its concept profile and to the weight of the concepts (the weight of a concept depends on the weight of each of its reformulated term in the retrieved documents). The clusters are sorted according to their weights and the first 1000 documents in this sorted list are retrieved.

Content-based Image Retrieval System

The content-based image retrieval system we used in ImageCLEF 2006 is the system PIRIA (Program for the Indexing and Research of Images by Affinity) [3], developed in our lab. The query image is submitted to the system, which returns a list of images ranked by their similarity to the query image. The similarity is obtained by a metric distance that operates on every image signatures. These indexed images are compared according to several classifiers: principally *Color, Texture* and *Form* if the segmentation of the images is relevant. The system takes into account geometric transformations and variations like rotation, symmetry, mirroring, etc. PIRIA is a global one-pass system, feedback or "relevant/non relevant" learning methods are not used.
Color Indexing: This indexer first quantifies the image, and then, for each quantified color, it computes how much this color is connex. It can also be described as a border/interior pixel classification [5]. The distance used for the color indexing is a classical L2 norm.

Texture Indexing: A global texture histogram is used for the texture analysis. The histogram is computed from the Local Edge Pattern descriptors [2]. These descriptors describe the local structure according to the edge image computed with a Sobel filtering. We obtain a 512-bins texture histogram, which is associated with a 64-bins color histogram where each plane of the RGB color space is quantized into 4 colors. Distances are computed with a L1 norm.

Form Indexing: The form indexer used consists of a projection of the edge image along its horizontal and vertical axes. The image is first resized in 100x100. Then, the Sobel edge image is computed and divided into four equal sized squares (up left, up right, bottom left and bottom right). Then, each 50x50 part is projected along its vertical and horizontal axes, thus giving a 400-bins histogram. The L2 distance is used to compare two histograms.

5.2.2 Search and Merging Strategy

Both systems are used independently to retrieve documents from textual and visual information. For the CBIR results, since queries contain several images, a first merging has been performed to obtain a single image list from the results of each query image: the score associated to result images is set to the max of the scores obtained for each query image.

Results obtained by each system are then merged using a weighted sum of the scores obtained by each system. To make results from the different systems comparable, we tried several normalization functions, presented in Table 1, where i is the weight associated with the scores of the ith system, $RSVmax$ is the the highest score obtained for a query, $RSVmin$ the lowest score, $RSVavg$ the average score and RSV the standard deviation of the scores. These functions have for instance been tested by [4] for data fusion in the multilingual tracks of previous CLEF campaigns. The submitted runs used the *normRSV* merging function.

sumRSV	$\sum \alpha_i * RSV_i$
normRSVMax	$\sum \alpha_i * RSV_i / RSV_{max}$
normRSV	$\sum \alpha_i * (RSV_i - RSV_{min}) / (RSV_{max} - RSV_{min})$
Zscore	$\sum \alpha_i * [(RSV - RSV_{avg}) / RSV_\delta + (RSV_{avg} - RSV_{min}) / RSV_\delta]$

Table 1: Different weightings for score merging

Based on the results from the previous campaigns, we also considered a conservative merging strategy: we use the results obtained by one system only to reorder the results obtained by the other, the score of a document is modified using the same merging coefficient.

5.2.3 Results for the ImageCLEF Photo task

We used, for the text retrieval part, textual queries in English, Spanish, French and German. We used English and German as independent target languages (as the annotations in both languages refer to the same images and form more an aligned corpus, it did not seem interesting to use both languages as a single multilingual corpus). We only submitted runs with the English target language. For the CBIR system, we tested the color and texture indexers.

CBIR results		
indexer	map	relret
color	**0.0468**	**961**
texture	0.0363	887

text results for English		
topics	map	relret
eng	**0.1427**	**1835**
spa	0.1416	1427
fre	0.1031	1380
ger	0.1009	1067

text results for German		
topics	map	relret
eng	0.0916	**1445**
spa	0.127	1394
fre	0.117	1302
ger	**0.145**	1381

We present in Table 2 the results obtained by the CBIR system alone and the text system alone.

Table 2: Comparative results for the CBIR system and the image system alone and the text system alone

We present in Table 3 the results obtained by the merging of the two systems, with a *normRSV* merging schema, and for different values of α (α being the weight associated with the text results, 1 − α to the image results). In this merging, we used the English topics with the English annotations and the color indexer. The runs submitted for merged results used α = 0.7.

Results are given for the mean average precision (*map*) and the number of relevant documents retrieved (*relret*).

	eng		rre		spa		spa	
α	map	relret	map	relret	map	relret	map	relret
1	0.1427	1835	0.1031	1380	0.1416	1427	**0.1009**	1067
0.9	**0.146**	**1951**	0.108	1536	**0.144**	1702	0.0969	**1397**
0.8	0.146	1927	**0.112**	**1545**	0.14	1714	0.097	1399
0.7	0.14	1864	0.109	1536	0.136	1713	0.0978	1399
0.6	0.128	1807	0.105	1533	0.129	**1715**	0.0948	1399
0.5	0.112	1750	0.1	1499	0.125	1707	0.0906	1400

Table 3: Comparative results for in merging strategies

We see from these results that this simple a-posteriori merging of text and image results based on a weighted sum of the scores can increase the mean average precision and number of relevant documents retrieved (best value of α is around 0.9 or 0.8).

However, the gain is still small, due to the fact that CBIR results are surprisingly quite poor. On one hand, we are investigating in more details the flaws of the indexers for this new image corpus. On the other hand, a first analysis of the results show that merging the CBIR results for the example images before merging with the text results is not a good idea: when several example images are given, they can provide different aspects of what the results should look like, and therefore can be as different as possible, in the range of relevant images. The merging of results base on purely visual similarity can be irrelevant in this case. The analysis of the CBIR results show that the rate of common images in the results for the different example images of a same topic does not exceed, in average, 16 to 18%. Table 4 present the results obtained using only one example image for each topic. The best example image has been taken (according to the reference), and the gain on mean average precision in this case is more than 11%. The problem still remains to find the best image example (in our results, the average precision obtained for each image example does not correlate well with the average score given by the CBIR system).

	eng	
α	map	relret
1	0.1427	1835
0.9	0.151	**1974**
0.8	0.156	1955
0.7	**0.159**	1914
0.6	0.156	1887

Table 4: Results merging text results with each image example

Another solution to this problem can be to consider each result of an example image as an independent result, and merge all results according to the same schema. In this case, the weight associated to the text result is α and the weight associated to each CBIR result is α/n where n is the number of example images. Table 5 present the results obtained with this method. The gain in

this case is around 8%, but this method does not need to determine a priori the best example image.

	eng	
α	map	relret
1	0.1427	1835
0.9	0.149	**1990**
0.8	0.153	1962
0.7	**0.154**	1915
0.6	0.149	1851

Table 5: Results merging text results with all image example

5.2.4 Conclusion of CEA LIST's participation in ImageCLEF

The experiments performed by the CEA LIST in the ImageCLEF 2006 campaign show that merging results from different media may increase the performance of a search system: a well-tuned a posteriori merging of the results obtained by two general purpose systems (no particular adaptation of the systems was made for the two tasks) can improve the mean average precision. An analysis of the CBIR results show that merging the results obtained for different example images can increase the noise in global results since example images are often chosen to be visually different to show several aspects of possible relevant images. Some solutions are proposed to cope with this aspect, such as taking only one example image (the best), or using all example image, but merging each with text results (not between them). Both solutions lead to better results in term of mean average precision.

More sophisticated solutions could be considered, such as working on the image analysis to try to determinate the similarities of the example images and find similar images in the collection based on these similarities, instead of considering each example image independently.

5.3 CEA LIST in Question Answering at CLEF QA 2006

The OEdipe question-answering system developed by the CEA-LIST/LIC2M was primarily a baseline system that is progressively extended by adding new modules in order to both enlarge its capabilities and improve its results. This incremental approach allows to carefully test the impact and the interest of each added module. The first version of the OEdipe system [7] was a passage-based system that has been developed for the French EQUER evaluation campaign [6]. It has been extended for CLEF-QA 2005 [11] to extract short answers from passages. Our results in this evaluation showed that our strategy for processing factoid questions was reasonably successful but that our few heuristics for answering definition questions were not sufficient. Hence, for our participation to the French monolingual track of the CLEF-QA 2006 campaign, we developed a new module in the OEdipe system that is designed to deal with definition questions. This module is based on the automatic learning of patterns to extract short answers for definition questions.

5.3.1 Overview of the OEdipe system

We present in this section the main features of the OEdipe system. More details can be found in the report of the previous campaign [11]. Figure 1 illustrates the architecture of the OEdipe system. This architecture is a standard pipeline architecture: the question is first submitted to a search engine to retrieve a restricted set of documents. The linguistic processing provided by the CEA LIST LIMA (LIc2m Multilingual Analyzer) analyzer [8] is used to normalize words and extract named entities from both question and documents. A deeper analysis of the question is performed to identify the expected type of the answer and the focus of the question.

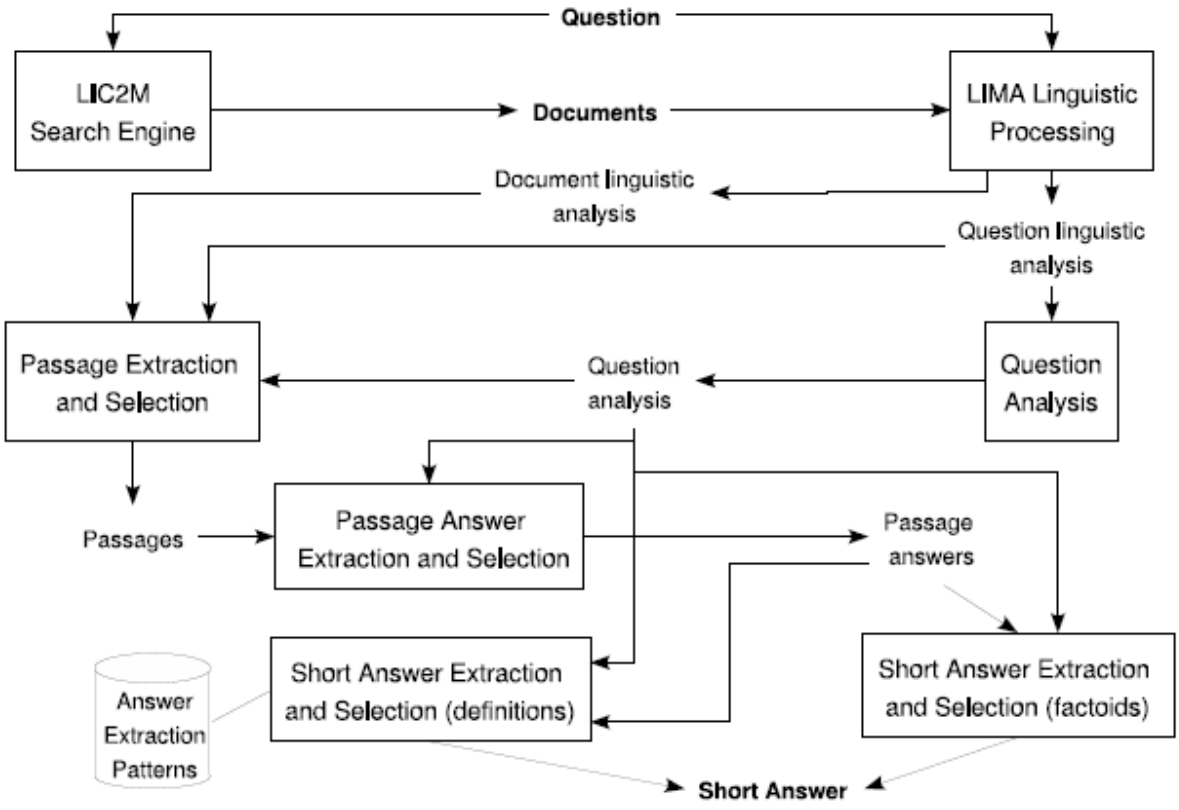


Figure 1. Architecture of the OEdipe system

A two-step gisting process is applied to the documents retrieved by the search engine to locate where answers are the more likely to be found: passages are first delimited using the density of the question words in documents, which is a basic criterion but is not computationally expensive; passage answers with a maximal size of 250 characters are then extracted from these passages by

computing a score in a sliding fixed-size window. Finally, short answers are extracted from the passage answers depending on their expected type.

More precisely, the OEdipe system is composed of the following modules:

Search Engine As in CLEF-QA 2005, we used the CEA LIST LIC2M search engine, that was already evaluated through the Small Multilingual Track of CLEF in 2003 and 2004 [9] [10]. Each question is submitted to the search engine without any pre-processing as the LIC2M search engine applies the LIMA linguistic processing to its queries. The only difference with CLEF-QA 2005 concerns the number of retrieved documents. In CLEF-QA 2005, a predefined number of documents (equal to 25) was retrieved. But the LIC2M search engine is concept-based and returns classes of documents that share the same query concepts (*i.e.* multi-terms and named entities): classes are ranked, but documents inside classes are considered equivalent and are not ranked. Hence, retrieving a given number of documents implies an arbitrary cut inside a class. We tried to avoid this arbitrary selection by using a variable number of documents (between 20 and 50) and applying the following algorithm, where *random(s,n)* is a function that randomly selects *n* elements from a set *s*. For CLEF-QA 2006, an average number of 33 documents by question were selected by this algorithm.

$selectedDocuments \leftarrow \{\}$
$i \leftarrow 1$
while $card(selectedDocuments) < 20 \ \wedge \ i \leq card(classes)$ **do**
    $currentClass \leftarrow classes[i]$
    $i \leftarrow i + 1$
    if $card(selectedDocuments) + card(currentClass) \leq 50$
      then
        $selectedDocuments \leftarrow selectedDocuments \cup currentClass$
      else
        $randSelDocsNb = 50 - card(selectedDocuments)$
        $selectedDocuments \leftarrow selectedDocuments \cup random(currentClass, randSelDocsNb)$
    fi

Linguistic Processing The linguistic processing of questions and documents is performed by the LIMA linguistic analyzer. Only a subset of its features[1] is used to normalize words (using a POS-tagger and lemmatizer), identify content words and extract named entities. The named entities recognized by LIMA are restricted to the MUC named entities [15], that is to say *persons*}, *locations*, organizations, *dates and times* and *numerical measures*, plus *events* and *products*.

[1] The LIMA analyzer can also perform term extraction or syntactic analysis but these capabilities are not exploited in OEdipe yet

Question Analysis In OEdipe, the question analysis module performs three different tasks:

- *identification of the expected type of the answer*: the result of this task determines the strategy applied by OEdipe for extracting answers. If the expected type of the answer corresponds that a type of named entities that can be recognized by LIMA, OEdipe searches in the selected document passages for the named entity of that type whose context is the most compatible with the question. Otherwise, it assumes that the question is a definition question and applies specific linguistic patterns to extract possible answers

- *identification of the focus of the question*: the focus of a question is defined as the part of the question that is expected to be present close to the answer. This task is part of the new module developed for definition question[2] and is achieved using a specific set of patterns, implemented as finite-state automata. Here are two examples of such patterns

  [Qu']::[est] [ce] [@Que] [$L_DET] *{1-30} [?]:FOCUS:
  [Qui]::[être$L_V] *{1-30} [?]:FOCUS

  The first rule identifies *Atlantis*} as the focus of the definition question "*Qu'est-ce que l'Atlantis*?" (What is Atlantis?), while the second rule extracts *Hugo Chavez* as the focus of the question "*Qui est Hugo Chavez*?" (Who is Hugo Chavez?).

- *selection and weighting of the significant words of the question*: the significant words of the question are its content words and their weight is obtained by looking up their normalized information in a reference corpus to evaluate their specificity degree.

Passage Extraction and Selection This module first delimits candidate passages by detecting the document areas with the highest density of question words. Then, a score is computed for each delimited passage depending on the number and the significance of the words of the question it contains. Candidate passages are ranked according to this score and passages whose the score is lower than a predefined threshold are discarded.

Passage Answer Extraction and Selection: A passage answer is extracted from each selected passage by sliding a window over the passage and computing a score at each position of the window according to its content and the expected type of the answer. The size of this window is equal to the size of the passage answers to extract (250 characters in our case). The position of the window is anchored on content words for definition questions and on named entities that are compatible with the expected answer type for factoid questions. A predefined number of passage answers is selected according to their score.

[2] It is currently performed only for definition questions but from a more general point of view, it could also be useful for improving the processing of factoid questions.

Short Answer Extraction and Selection: When the expected type of the question is a named entity, each passage answer is centered on a named entity of that type. Hence, the extracted short answer is directly that named entity. Its score is equal to the score of the passage answer. The extraction of short answers for definition questions is presented in details in the next section. As for passage answers, each short answer is given a score and all short answers are ranked according to this score;

5.3.2 Answering Definition-type Questions

As mentioned before, the main improvement of the OEdipe system for CLEF-QA2006 concerns its ability to answer questions whose the expected answer is not a named entity, and more specifically definition questions such as *What is X*? or *Who is X*? *What/Who* questions have proved to be difficult both because they are generally short and the search of an answer cannot be focused by a specific type of elements such as a named entity. Hence, trying to answer this kind of questions with a basic approach leads to poor results (as it was illustrated by OEdipe's results at CLEF-QA 2005).

Most of the question answering systems rely on a set of handmade linguistic patterns to extract answers for that kind of questions from selected sentences [20]. Some work has been done to learn such patterns from examples, following some work in the Information Extraction field. One of the first attempt was from Ravichandran and Hovy [18], who proposed a strategy that combines the use of the Web and suffix trees. Mining the Web for learning such patterns is also the solution adopted by [13],. Jousse et al. [19] tested several machine learning algorithms for extracting patterns and Cui et al. [12] proposed a new algorithm for learning probabilistic lexico-syntactic patterns, also called *soft patterns*.

Works such as [18] or [19] have proved that building a set of question-answer examples is quite easy, especially from the Web. This is why we chose to rely on lexico-syntactic patterns learnt from examples for answering definition questions. Moreover, this approach appears to be both more flexible and less costly when a question answering system must be extended to new domains.

5.3.3 Learning of definitional patterns

The algorithm we used to learn linguistic patterns for extracting answers to definition questions is an extension of the Ravichandran and Hovy's algorithm[18]. This extension, proposed by Ravichandran in [17], allows learning multilevel patterns instead of surface patterns: multilevel patterns can refer to different levels of linguistic information. These kinds of patterns have been used in various applications to extract different kind of information: semantic relations to populate knowledge bases in [16] and [14], answers in question answering systems or factual relations between entities in information extraction. In our case, the induction of patterns is done from a set of example answers to definition questions. The basic element of a pattern can be the surface form of a word, its part of speech (POS) or its lemma. These three levels of information

are obtained using the LIMA linguistic analyzer. More precisely, the overall procedure for building up a base of patterns dedicated to the extraction of definitional answer is the following:

1. Building a corpus of example answer sentences. Unlike [18] or [13], our corpus of example answers was not built from the Web but came from the results of the EQUER evaluation and from the previous CLEF-QA evaluations. For each definition question of these evaluations, all the sentences containing a correct answer to the question were extracted.
2. Application of the LIMA linguistic analyzer to all the answer sentences to get the three levels of linguistic information.
3. Abstraction of answer sentences. This abstraction consists in replacing in each answer sentence the focus of the question by the tag <*focus*> and the short answer to the question by the tag <*answer*>.
4. Application of the multilevel pattern-learning algorithm between each pair of sentences (see below).
5. Selection of the top *P* patterns on the basis of their frequency.

The multilevel algorithm for the induction of patterns is taken from [17]. It is composed of two steps: the first one consists in calculating the minimal edit distance between the two sentences to generalize[3] that are necessary to transform one sentence into the other one}; the second one extracts the most specific multilevel pattern that generalizes the two sentences. Some of the obtained alignments are then completed by adding two wildcard operators: ({*}s{*}) represents 0 or 1 instance of any word while ({*}any_word{*}) represents exactly 1 instance of any word.

Here are some examples of the definitional patterns induced by this algorithm[4]

<answer> (<focus>
<focus> (<answer>
<focus> , le <answer>
<focus> , un <answer>
<focus> être L_DET_ARTICLE_INDEF <answer>
<focus> , (*any_word*) <answer>

5.3.4 Application of patterns to extract answers

The definition patterns resulting from the learning method described in the previous section were integrated into the OEdipe system by applying them after the selection of passage answers. The exact procedure is the following:

[3] The edit distance between two sentences is equal to the number of edit operations (insertion, deletion and substitution

[4] Patterns' lexical items are translated as: *le=the, un=a-an, être=to be*

- Instantiation of definitional patterns: the <*focus*> tags in patterns is replaced by the focus of the question, as identified by the dedicated rules of the question analysis module (see above)
- Application of the LIMA linguistic analyzer: each passage answer is analyzed to get the three levels of linguistic information possibly used in patterns.
- Extraction of short answers: each instantiated pattern is aligned with the passage answer. This alignment starts from the focus of the window and is check word by word until the <*answer*> tag is reached in the pattern. If this alignment succeeds, the noun phrase in the passage answer that corresponds to the <*answer*> tag in the pattern is kept as a possible short answer.
- Selection of the top *A* short answers on the basis of their frequency.

The set of short answers extracted by using definition patterns are ranked according to their number of matching patterns. The short answer with the highest value is returned as the answer to the considered question.

5.3.5 Evaluation

Only one run was submitted in the CLEF-QA 2006 evaluation for the OEdipe system. For the 187 factoid (F), definition (D) and temporally restricted (T) questions, OEdipe returned 30 right answers, 3 unsupported answers and 6 inexact answers, which gives an overall accuracy of 16%. Moreover, the detection of a lack of answer by OEdipe was right for only one question out of four. For the 10 list questions, 7 right answers were returned, which corresponds to an average precision of 0.1633 (see [21] for explanations about evaluation measures).

	Factoid (F+T)		**Definition (D)**	
	# right answers	**accuracy**	**# right answers**	**accuracy**
CLEF-QA 2005	28	18.7	0	0.0
CLEF-QA 2006	15	10.3	15	36.6

Table 6. Comparison of the distributions of OEdipe's right answers at CLEF-QA 2005 and CLEF-QA 2006

At a quick glance, the results of OEdipe at CLEF-QA 2006 seem to be comparable, with a slight improvement, to its results at CLEF-QA 2005, where its overall accuracy was equal to 14% with 28 right answers for the F, D and T questions. However, Table 6 shows that the distributions of right answers are quite different for the two evaluations. The use of definitional patterns leads to a very significant improvement for definition questions but results for factoid questions, which were processed by exactly the same version of OEdipe as the CLEF-QA 2005 version, significantly decrease.

The improvement for definition questions was expected since a specific module was developed for that kind of questions but there is no clear explanation of the decrease of results for factoid

questions, except that they were perhaps more difficult than CLEF-QA 2005 factoid questions. Moreover, Table 7 shows that the question analysis module is not responsible of this decrease of OEdipe's results for factoid questions since its accuracy for CLEF-QA 2006 questions is higher than for CLEF-QA 2005 questions. It should also be noted that the focus was correctly identified for all the definition questions.

question type	# questions	# incorrect types	accuracy (2006)	acc. (2005)
Factoid (F+T)	146	9	93.8	86.0
Definition (D)	41	4	90.2	100

Table 7. Detailed results of the question analysis module for CLEF-QA 2006 and comparison with CLEF-QA 2005

5.3.6 Conclusion of CEA LIST's participation in CLEF QA

We have given an overview of the CEA LIST version of the OEdipe system that participated to the French monolingual track of the CLEF-QA 2006 evaluation and we more particularly detailed its new aspects with regards to the CLEF-QA 2005 version, that is to say the learning from examples of lexico-syntactic patterns and their application to extract short answers for definition questions. These new aspects proved to be interesting but there is still room for improvements. The first of them is certainly to integrate the syntactic analysis of LIMA in OEdipe, which would particularly enable to take each noun phrase as a whole and make extraction patterns more general. The second main improvement is to extend the use of extraction patterns to the processing of factoid questions. The results of OEdipe for that kind of questions are significantly lower than its CLEF-QA 2005 results and an explanation of this fact still have to be found. However, it is clear that the use of such patterns should be an interesting way to focus the search of a named entity answer in passages.

Bibliography

[1] Romaric Besancon and Christophe Millet. Data fusion of retrieval results from different media:experiments at ImageCLEF 2005. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Gareth J.F.Jones, Michael Kluck, Bernardo Magnini, Henning Muller, and Maarten de Rijke, editors, Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Springer, 2005.

[2] Ya-Chun Cheng and Shu-Yuan Chen. Image classification using color, texture and regions. Image and Vision Computing, 21(9), September 2003.

[3] Magali Joint, Pierre-Alain Moellic, Patrick Hede, and Pascal Adam. PIRIA : A general tool for indexing, search and retrieval of multimedia content. In SPIE Electroning Imaging 2004, San Jose, California USA, January 2004.

[4] Jacques Savoy and Pierre-Yves Berger. Report on CLEF-2005 evaluation campaign: monolingual, bilingual, and GIRT information retrieval. In Carol Peters, editor, Working Notes for the CLEF 2005 Workshop, Vienna, Austria, 2005.

[5] Renato. O. Stehling, Mario. A. Nascimento, and Alexandre X. Falcao. A compact and efficient image retrieval approach based on border/interior pixel classification. In CIKM ’02: Proceedings of the eleventh international conference on Information and knowledge management, McLean, Virginia, USA, 2002.

[6] Ayache, C., Grau, B., Vilnat, A., Campagne d’évaluation EQueR-EVALDA : Evaluation en Question- Réponse. Actes de l’Atelier EQueR-EASY de TALN’05, Dourdan, France, 2005

[7] A. Balvet, M. Embarek, O. Ferret, " Minimalisme et question-réponse : le système Œdipe ", Taln 2005, Dourdan, France, June 6-10, 2005

[8] Romaric Besancon and Gael Chalendar (de). L’analyseur syntaxique de LIMA dans la campagne d’évaluation EASY. In actes de la 12e conference annuelle sur le Traitement Automatique des Langues Naturelles, TALN 2005, Dourdan, France, June 2005

[9] Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard, Hubert Naets: Concept-Based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003. CLEF 2003: 174-184, 2003

[10] Romaric Besançon, Olivier Ferret, Christian Fluhr: Integrating New Languages in a Multilingual Search System Based on a Deep Linguistic Analysis. CLEF 2004: 83-89 2005

[11] Romaric Besançon, Mehdi Embarek, Olivier Ferret: The OEDipe System at CLEF-QA 2005. CLEF 2005: 337-346

[12] Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. In Proceedings of SIGIR 2005.

[13] Yongping Du, Xuanjing Huang, Xin Li, Lide Wu: A Novel Pattern Learning Method for Open Domain Question Answering. IJCNLP: 81-89, 2004

[14] Mehdi Embarek and Olivier Ferret. Extraction de relations sémantiques à partir de textes dans le domaine médical. JOBIM 2006, Bordeaux, France, 2006

[15] R. Grisham and B. Sundheim, Design of the MUC-6 evaluation , proc. of the MUC-6 conference, Columbia, MD, 1995

[16] Pantel, Patrick, Deepak Ravichandran, and Eduard Hovy: Towards Terascale Knowledge Acquisition. Proceedings of the COLING conference, Geneva, Switzerland. 2004.

[17] D. Ravichandran. Terascale Knowledge Acquisition. PhD thesis, University of Southern California, 2005

[18] D. Ravichandran and E. Hovy. Learning surface text patterns for a Question Answering system. Proceedings of the 40th ACL conference. Philadelphia, PA, 2002

[19] F. Jousse, I. Tellier, M. Tommasi and P. Marty, Learning to Extract Answers in Question Answering: Experimental Studies, Actes de CORIA'05, 85-100, 2005

[20] M. M. Soubbotin and S. M. Soubbotin Patterns and potential answer expressions as clues to the right answers. Proceedings of the Tenth Text REtrieval Conference, 175-182, 2001

[21] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In F. Borri In C. Peters, editor, Results of the CLEF 2004 Cross-Language System Evaluation Compaign, Bath, U.K., 2004.