Marcin Detyniecki
Ana García-Serrano
Andreas Nürnberger (Eds.)

LNCS 6535

Adaptive Multimedia Retrieval

Understanding Media and Adapting to the User

7th International Workshop, AMR 2009
Madrid, Spain, September 2009
Revised Selected Papers

Springer

Lecture Notes in Computer Science 6535

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Marcin Detyniecki
Ana García-Serrano
Andreas Nürnberger (Eds.)

Adaptive Multimedia Retrieval

Understanding Media and Adapting to the User

7th International Workshop, AMR 2009
Madrid, Spain, September 24-25, 2009
Revised Selected Papers

Springer

Volume Editors

Marcin Detyniecki
Université Pierre et Marie Curie
Paris, France
E-mail: marcin.detyniecki@lip6.fr

Ana García-Serrano
Universidad Nacional de Educación a Distancia
Madrid, Spain
E-mail: agarcia@lsi.uned.es

Andreas Nürnberger
Otto-von-Guericke Universität Magdeburg
Magdeburg, Germany
E-mail: andreas.nuernberger@ovgu.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-18448-2 e-ISBN 978-3-642-18449-9
DOI 10.1007/978-3-642-18449-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010942865

CR Subject Classification (1998): H.4, H.3, I.4, H.5.1, H.5, I.2, H.2

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book contains a selection of the revised contributions that were initially submitted to the International Workshop on Adaptive Multimedia Retrieval (AMR 2009). The workshop was organized by the Universidad Nacional de Educación a Distancia (UNED) in Madrid, Spain, during September 24–25, 2009.

The goal of the AMR workshops is to intensify the exchange of ideas between the different research communities involved in this topic, in particular those focusing on multimedia retrieval, human–computer interaction, machine learning and artificial intelligence, to provide an overview of current activities in their areas of expertise and to point out connections between them. In this spirit, the first three events were collocated with artificial intelligence-related conferences: in 2003 as a workshop of the German Conference on Artificial Intelligence (KI 2003); in the following year as part of the European Conference on Artificial Intelligence (ECAI 2004) and in 2005 co-located with the International Joint Conference on Artificial Intelligence (IJCAI 2005). Because of its success, in 2006 the University of Geneva, Switzerland, organized the workshop for the first time as a standalone event; and since then it has been so: AMR 2007 was organized by the Laboratoire d'Informatique de Paris VI (LIP6) in France and AMR 2008 by the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI) in Berlin.

In 2009 the workshop presented a multitude of ideas around two main topics: understanding the media content and adapting to the user. The diversity of aspects covered – ranging from theoretical work to practical implementations – provides a very rich and complementary set of perspectives. In the revised contributions contained in this edition the authors propose different ways to 'understand' medias as diverse as images, music, spoken audio and videos. *Understanding* is tackled by contributions that take small steps to reduce the semantic gap, e.g., by what one of the authors called 'pre-semantic features.' Further contributions focus on *describing* images, on *identifying* musical notes or spoken words, and on *summarizing* video content.

Adapting to the user was tackled from a variety of points of view. The object (to be adapted) ranged from the underlying similarity, over the type of interaction (e.g., browsing vs. querying), to the content to be presented (e.g., video summaries). Information about the user was collected with a variety of techniques, as for instance feedback or log analysis. Finally, all these challenges are linked and put into perspective in the special contribution of the invited speaker, which addresses the mining of networked media collections.

We believe that the above trends are representative and thus this book provides a good and conclusive overview of the current research in the area of adaptive multimedia retrieval. We would like to thank all members of the Program Committee for supporting us in the reviewing process, the workshop participants for their willingness to revise and extend their papers for this book, the sponsors for their financial help and Alfred Hofmann from Springer for his support in the publishing process.

August 2010

Marcin Detyniecki
Ana García-Serrano
Andreas Nürnberger

Organization

Program Chairs

Marcin Detyniecki	CNRS, Laboratoire d'Informatique de Paris 6, France
Ana García-Serrano	Universidad Nacional de Educación a Distancia, Madrid, Spain
Andreas Nürnberger	Otto-von-Guericke-Universität Magdeburg, Germany

Technical Chairs

Sebastian Stober	Otto-von-Guericke-Universität Magdeburg, Germany
Javier Artiles	Universidad Nacional de Educación a Distancia, Madrid, Spain

Local Organization

Rodrigo Agerri	Universidad Politécnica de Madrid, Spain
Javier Calle	Universidad Carlos III de Madrid, Spain
Víctor Fresno	Universidad Nacional de Educación a Distancia, Spain

Program Committee

Jenny Benois-Pineau	University of Bordeaux, LABRI, France
Stefano Berretti	Università di Firenze, Italy
Susanne Boll	University of Oldenburg, Germany
Eric Bruno	University of Geneva, Switzerland
Juan Cigarrán	Universidad Nacional de Educación a Distancia, Spain
Bogdan Gabrys	Bournemouth University, UK
Xian-Sheng Hua	Microsoft Research, Beijing, China
Alejandro Jaimes	Telefónica R&D, Spain
Philippe Joly	Université Paul Sabatier, Toulouse, France
Gareth Jones	Dublin City University, Ireland
Joemon Jose	University of Glasgow, UK
Stefanos Kollias	National Technical University of Athens, Greece

Stéphane Marchand-Maillet	University of Geneva, Switzerland
Trevor Martin	University of Bristol, UK
José María Martínez Sánchez	Universidad Autónoma de Madrid, Spain
Bernard Merialdo	Institut Eurécom, Sophia Antipolis, France
Jan Nesvadba	Philips Research, Eindhoven, The Netherlands
Nuria Oliver	Telefónica R&D, Spain
Gabriella Pasi	Università degli Studi di Milano Bicocca, Italy
Valery Petrushin	Accenture Technology Labs, Chicago, USA
Stefan Rüger	The Open University, Milton Keynes, UK
Simone Santini	Universidad Autonoma de Madrid, Spain
Raimondo Schettini	University of Milano Bicocca, Italy
Ingo Schmitt	University of Cottbus, Germany
Nicu Sebe	University of Amsterdam, The Netherlands
Alan F. Smeaton	Dublin City University, Ireland
Arjen De Vries	CWI, Amsterdam, The Netherlands

Supporting Institutions

Universidad Nacional de Educación a Distancia, Madrid, Spain
Otto-von-Guericke-Universität Magdeburg, Germany
Laboratoire d'Informatique de Paris 6 (LIP6), France
Centre national de la recherche scientifique (CNRS), France

Table of Contents

Invited Contribution

Grasping Multimedia Streams

Pinpointing Music

Adapting Distances

Understanding Images

Around the User

Mining Networked Media Collections

Stephane Marchand-Maillet, Donn Morrison, Eniko Szekely,
Jana Kludas, Marc Vonwyl, and Eric Bruno*

Viper group – University of Geneva, Switzerland
Stephane.Marchand-Maillet@unige.ch
http://viper.unige.ch

Abstract. Multimedia data collections immersed into social networks may be explored from the point of view of varying documents and users characteristics. In this paper, we develop a unified model to embed documents, concepts and users into coherent structures from which to extract optimal subsets and to diffuse information. The result is the definition information propagation strategies and of active guiding navigation strategies of both the user and document networks, as a complement to classical search operations. Example benefits brought by our model are provided via experimental results.

1 Introduction

Many current information management systems are centered on the notion of a query related to information search. This is true over the Web (with all classical Web Search Engines), and for Digital Libraries. In the domain of multimedia, available commercial applications propose rather simple management services whereas research prototypes are also looking at responding to queries. In the most general case, information browsing is designed to supplement search operations. This comes from the fact that the multimedia querying systems largely demonstrate their capabilities using query-based scenario (by Example, by concepts) and these strategies often show limitations, be it in their scalability, their usability or utility or their capabilities or precision. Multimedia search systems are mostly based on content similarity. Hence, to fulfill an information need, the user must express it with respect to relevant (positive) and non-relevant (negative) examples. From there, some form of learning is performed, in order to retrieve the documents that are the most similar to the combination of relevant examples and dissimilar to the combination of non-relevant examples. The question then arises of how to find the initial examples themselves.

Researchers have therefore investigated new tools and protocols for the discovery of relevant bootstrapping examples. These tools often take the form of browsing interfaces whose aim is to help the user exploring the information space in order to locate the sought items. Similarity-based visualization (see *e.g* [15,16]) organizes images with respect to their perceived similarities. Similarity

* This work is supported by the Swiss NCCR (IM)2 and the EU NoE Petamedia.

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 1–11, 2011.

is mapped onto the notion of distance so that a dimension reduction technique may generate a 2D or 3D space representation where images may be organized. A number of similar interfaces have been proposed to apply to the network of users or documents but most browsing operations are based on global hyperlinking (*e.g* Flickr or YouTube pages).

Another popular option is the use of keywords as a mean to apply an initial loose filtering operation over the collection. However, the possibility of responding keyword-based queries depends on the availability of textual annotation over the collection. To be scalable, this option must include a way of making best use of the shallow annotation provided over a subset of documents. Recent data organisation over the WWW, mixing large collections of multimedia documents and user communities offer opportunities to maintain such level of annotation and enable efficient access of information at large scale.

Here, we formalise a model for networked media collections (section 2) that unifies most operations made at the collection level as propagation of information (*e.g* annotation) within a multigraph (section 3). An example of developments exploiting these structures over documents only is presented in section 4.

2 Multidimensional Networked Data Modeling

We start with a collection $\mathcal{C} = \{d_1, \ldots, d_N\}$ of N multimedia items (text, images, audio, video,...). Traditionally, each document d_i may be represented by a set of features describing the properties of the document for that specific characteristic. In a search and retrieval context, it is expected that mainly discriminant characteristics are considered (*i.e* the characteristics that will make it possible to make each document unique w.r.t a given query). With each of these characteristics is associated a similarity measure computed over the document extracted features. Hence, given $\mathcal{C}$, one may form several similarity matrices $S^{[c]} = \left(s^{[c]}_{ij}\right)$, where the value of $s^{[c]}_{ij}$ indicates the level of similarity between documents d_i and d_j w.r.t characteristic c.

In our context, we consider each matrix $S^{[c]}$ as a weighted graph connectivity matrix. Since $S^{[c]}$ is symmetric, it represents the connectivity matrix of a complete non-oriented graph where nodes are documents. Collecting all matrices $S^{[c]}$ $\forall c$, we may therefore represent our collection as a multigraph acting over the node set $\mathcal{C}$.

This simple similarity-based mapping of the collection provides a useful dimensionless representation over which to act in view over efficient collection exploration. In [17], the High-Dimensional Multimodal Embedding (HDME) is presented as a way to preserve cluster information within multimedia collections. In our context, it forms a useful mapping for projection or dimension selection for visualization. It is also a way of enhancing our Collection Guiding principle proposed in [10].

Alternatively, documents may be attached a form of metadata taken from a knowledge base $\mathcal{B} = \{b_1, \ldots, b_M\}$ modeled again as a multi-graph over a set of M concepts b_j, acting as nodes bearing relationships between themselves. Distance

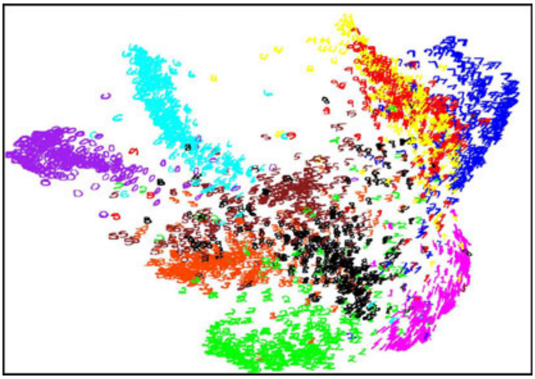


Fig. 1. Global representation of the MNIST handwritten digit image collection after cluster-preserving dimension reduction

relationships $D^{[s]} = \left(b_{ij}^{[s]}\right)$ between concepts apply here . The value $b_{ij}^{[s]}$ indicates how much concepts b_i and b_j are close to each other with respect to interpretation s. Examples of such relationships are tags (acting as concepts) whose distance is measured using any word distance measure (*e.g* based on WordNet).

In turn, documents and concepts are also associated with "tagging" relationships $T^{[s]} = \left(t_{ij}^{[s]}\right)$, where $t_{ij}^{[s]}$ evaluates the strength of association between document d_i and concept b_j, under a given perspective (*e.g* interpretation, language) s. Again, the notion of perspective allows for creating multiple relationships between documents and tags (including with respect to users who potentially authored these relationships, see next).

Consider finally a population $\mathcal{P}$ of P users $\mathcal{P} = \{u_1, \ldots, u_P\}$ interacting with the collection $\mathcal{C}$, thus forming a classical social network. Thus, users may be associated by inter-relationships (*e.g* the "social graph" , to use the term coined by FaceBook). Classical relationships such as "is a friend of" or "lives nearby" may be quantified for each pair of users. Matrices $P^{[v]} = \left(p_{kl}^{[v]}\right)$ may thus be formed, where the value of $p_{kl}^{[v]}$ indicates the strength of the proximity aspect v between user u_k and user u_l.

We then consider that any user u_k may have one or more relationships with a document d_i. For example, user u_k may be the *creator* of document d_i or u_k may have *ranked* the document d_i a certain manner. For each of these possible relationships, we are therefore able to form a matrix $R^{[v]} = \left(r_{ik}^{[v]}\right)$, where the value of $r_{ik}^{[v]}$ indicates the strength (or simply the existence) or relation v between document d_i and user u_k. Users may then be associated with concepts of the knowledge base $\mathcal{B}$ by similar relationships. Essentially, a user may be associated with a concept that describes some particulars of that user (*e.g* being a *student*) or, conversely a concept may be associated by a user because this user has used this concept to annotate documents. Hence, relationships $V^{[s]} = \left(v_{ij}^{[s]}\right)$ are created, where $v_{ik}^{[e]}$ weights a particular link between concept b_i and user u_k.

In summary, we obtain $(\mathcal{C}, S^{[c]}\ \forall c)$, $(\mathcal{B}, D^{[s]}\ \forall s)$ and $(\mathcal{P}, P^{[v]}\ \forall v)$ as multigraphs acting over document, concepts and user node sets and the graph $(G, E) = (\mathcal{C} \cup \mathcal{B} \cup \mathcal{P}, R^{[v]} \cup D^{[s]} \cup V^{[s]})$ as a multi-tripartite graph relating documents, concepts and user node sets. Figure 2 illustrates this representation.

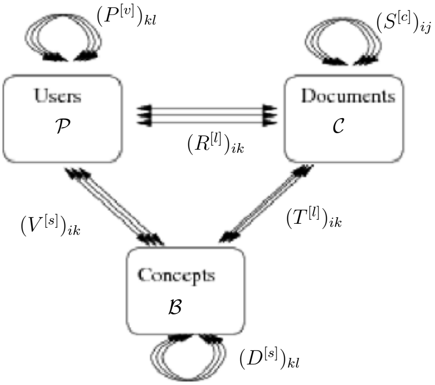


Fig. 2. The proposed graph-based modeling of a social network and associated annotated documents

Now, interestingly, this representation is a base tool for further network analysis and completion. Graph connectivity analysis of $(\mathcal{P}, P^{[v]}\ \forall v)$ for a given aspect v may tell us about coherence between parts of the population. Tools such as minimum vertex- or edge-cuts will indicate particular users or groups that are critical to maintain the connectivity and thus the coherence of the networked information structure.

This structure also allows for its own completion. Similar to what is proposed in [11], user interaction may be captured as one particular bipartite graph $(\mathcal{C} \cup \mathcal{P}, R^{[v]})$ and this information may be mined to enrich either inter-documents similarity $(\mathcal{C}, S^{[c]})$ (see section 3) or inter-user proximity $(\mathcal{P}, P^{[v]})$ to identify a community with specific interests (materialized by the interaction over certain groups of documents). When forming such new relationships, constraints for forming proper distance matrices (1:a) or similarity matrices (1:b) apply:

$$(a) \begin{cases} s_{ij}^{[c]} \geq 0 \\ s_{ij}^{[c]} = s_{ji}^{[c]} \\ s_{ii}^{[c]} = 0 \end{cases} (b) \begin{cases} 0 \leq s_{ij}^{[c]} \leq 1 \ \forall i, j \\ s_{ij}^{[c]} = s_{ji}^{[c]} \quad \forall i, j \\ s_{ii}^{[c]} = 1 \qquad \forall i \end{cases} \tag{1}$$

Similarly, recommender systems [7] will mine inter-user relationships and inter-document similarity to recommend user-document connections.

3 User Relevance Modeling

We are first interested in exploiting the graph G to mine a posteriori usage of interactive information systems, with the aim of using user interaction as a source of semantic knowledge that will enrich the descriptions of documents. In other words, with reference to our model in Figure 2, we will mine relationships $R^{[v]}$ (user-documents), in order to enhance the understanding of the structure of $T^{[s]}$ (documents-concepts) and $S^{[c]}$ (documents-documents).

Given a query-by-example retrieval system that affords relevance feedback, we can assume that, at any given stage, users will have invoked a set of L queries $\mathcal{Q} = \{q_1, ..., q_L\}$ over the set of documents $\mathcal{C}$. An $N \times L$ document-query relevance matrix $\mathcal{R}$ can then be defined, where each element

$$\mathcal{R}(i,j) = \begin{cases} +1 & \text{if the user marked document } d_i \text{ as positive in query } q_j, \\ -1 & \text{if the user marked document } d_i \text{ as negative in query } q_j, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The relevance matrix $\mathcal{R}$ thus formed is an instantiation of matrix $R^{[v]}$ defined in section 2. The proposed User Relevance Model formalizes a missing aspect of previous studies in long-term learning. Assumptions made in previous studies to generate and model artificial data ignore the concept-basis relationship between a document and a query, noise introduced due to user error, and oversimplify the user decision in judgment-making [14,6,2]. This new formalization permits the understanding of how long-term RF data is generated by users based on perceived concepts in the documents and queries.

The idea is then to examine the RF process to discover the underlying concepts present in the documents and queries. Essentially, we want to discover to what extent each concept $b_i \in \mathcal{B}$ exists in $d_i \in \mathcal{C}$ and $q_j \in \mathcal{Q}$.

Generally, decompositions made by latent-variable models are not unique and therefore the interpretation of the latent variables can be problematic [1]. However, the latent space present in the component matrices can be interpreted in light of the values of the rows and columns in the co-occurrence matrix.

For example, consider two images d_1 and d_2 depicting horses. Through a decomposition both documents are seen to have a high component of concept b_1. In the absence of further information, we could say that concept b_1 may represent something to do with horses.

Non-negative matrix factorization (NMF) offers a straightforward approach to the problem of discovering latent concepts from observed data. NMF, given a non-negative matrix $\mathcal{R}$, finds non-negative, non-unique factors giving:

$$\mathcal{R} \approx WH, \quad (3)$$

where $W \in R^{M \times K}$ and $H \in R^{K \times N}$ and such that $W \cdot H$ minimizes the Frobenius norm $||\mathcal{R} - WH||^2$ [9]. In our case, the resulting component matrix W yields a projection of the documents into the space defined by the latent basis vectors.

Another popular method is latent semantic analysis (LSA) which uses the singular value decomposition (SVD) to decompose the co-occurrence matrix into three components:

$$\mathcal{R} = U \Sigma V^T, \tag{4}$$

where U are the left singular vectors, Σ are the singular values, and V are the right singular vectors [4]. The decomposition is such that U and V are orthonormal, and Σ is a diagonal scaling matrix with values in decreasing order. By retaining only the first K singular values in Σ, we have an rank-K approximation of the original co-occurrence matrix:

$$\mathcal{R}_K \approx U_K \Sigma_K V_K^T, \tag{5}$$

where $U_K \in R^{M \times K}$, $\Sigma_K \in R^{K \times K}$, and $V_K \in R^{N \times K}$.

In choosing the number of latent variables appropriately in NMF and the SVD, we can conveniently extract the underlying concepts in the User Relevance Model. We can identify the concept weight matrices $T^{[s]} = \left(t_{ij}^{[s]}\right)$ as having the same dimensionality as W, U_K and H, V_K^T, respectively. In practice, we will never know the exact nature of the underlying concepts, and so M must be chosen such that it is large enough to capture diversity in the data yet small enough that some interpretable clustering is observed.

Because the singular value decomposition does not impose non-negativity assumptions on the co-occurrence matrix or the resulting component matrices, interpretation of the latent variables in U_K and V_K^T is not straightforward [8]. NMF is preferable in this sense, because the latent variables lend themselves to a modeling of non-negative concept weights, which we note leads to a probabilistic formulation [5].

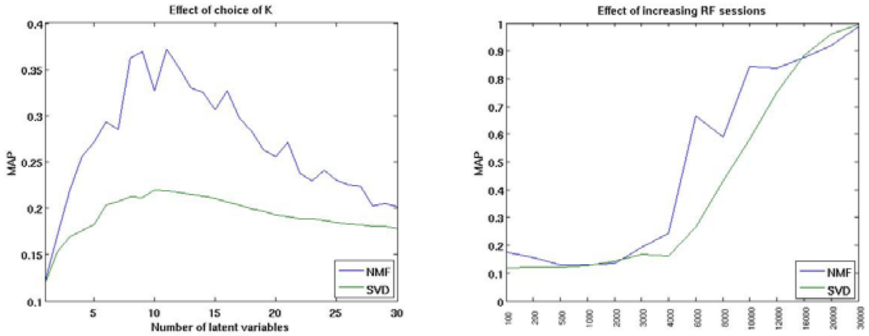
3.1 Experiments

Illustrative experiments are conducted on a small subset of the Corel image collection (see also [13]). The subset comprises 1,000 images uniformly spanning 10 categories (100 images per category). Although small, this dataset allows us to quickly and easily visualize performance. Document categories are contiguous in the matrix $\mathcal{R}$ and therefore similarly in all figures to make interpretation of the results easier.

All sessions of relevance feedback are generated according to the User Relevance Model described above. In other words, given the ground truth image categories, we generate a full relevance matrix and subsequently account for real-world sparsity and noise, yielding $\mathcal{R}$. Performance is measured using mean average precision (MAP). MAP emphasizes retrieving relevant documents first and provides a quantifiable measure of the clustering of the documents into latent concept classes.

The subplots of Figure 3 show the effects of varying various parameters in the User Relevance Model.

We know from previous work that the MAP should be at its maximum when the value of K is equal to the actual number of concepts underlying the data [12].



(a) Choice of number of latent variables (K)

(b) Number of RF sessions

Fig. 3. Parameters of the User Relevance Model are varied to show the MAP under different conditions

In Figure 3 (A), MAP is highest when $K \approx 10$. Figure 3 (B) demonstrates that the MAP increases as we collect more relevance feedback judgments (sessions). It is evident that the MAP approaches 1 as the number of relevance feedback sessions ML increases. A significant improvement in mean average precision is observed with as little as 6,000 RF sessions.

Figure 4 demonstrates the success of NMF's modeling of the concept weights from the User Relevance Model. The figure shows the highest ranked images for the particular concept. Figure 5 shows the reconstructed concept matrix W containing weights for each document. Columns corresponds to the bar plots

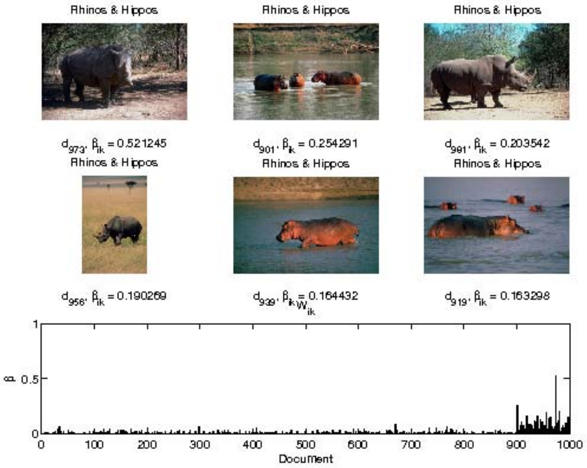


Fig. 4. Example underlying concept (bottom) with corresponding images depicting images of rhinoceri and hippopotami. Beneath each document is the corresponding concept weight.

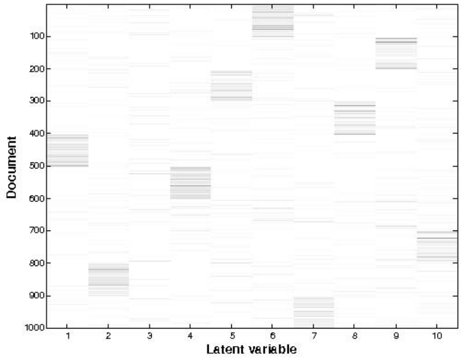


Fig. 5. Recovered document-concept relationships in W. Documents are clustered into the latent concepts. The concept in column 3 shows overlap between images. This is attributed to noise introduced in the User Relevance Model.

in Figure 4. Due to the inherent non-uniqueness of latent-variable models, the columns of W and the original document-concept matrix will not correspond. What is important to note is that the documents are grouped into similar clusters.

4 Multidimensional Data Exploration

We are now interested in defining exploration strategies over social networks. Based on the above modeling, we map this challenge onto that of defining optimal discrete structures in the multigraphs representing the social network. The objective is to complement the search paradigm with a navigation facility. We therefore assume that a search tool is used to position a user (called *client* to differentiate from users in the network) at a certain point within our multigraph by selecting a particular user or a particular document. From that point on, the navigation system should enable the client to move within a neighborhood, as defined by the connectivity structure, to explore the vicinity of this position. In other words, we wish to offer the client a view of where to navigate next and this view should be optimized from the information available. Further, this recommendation should be embedded within a global context so as to avoid cycles where the client stays stuck within a loop in the navigation path.

Our graph model is a suitable setup for this optimization. Formally, starting with a matrix $M = (m_{ij})$, where m_{ij} indicated the *cost* of navigating from item x_i to item x_j, we wish to find a column ordering o^* of that matrix that will minimize a certain criteria over the traversal of the items in this order. As a basis, we seek the optimal path that will minimize the global sum of the costs associated to the traversed edges. That is, we seek o^* as the ordering that will minimize the sum of the values above the diagonal so that

$$o^* = \arg\min_o \sum_{i \in o} m_{ii+1} \tag{6}$$

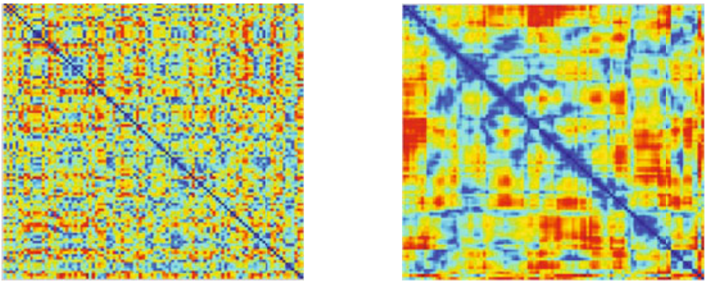


Fig. 6. An optimal reordering applied over a distance matrix (color similarity between 300 images)

The above is equivalent to solving the Symmetric Travelling Salesman Problem (S-TSP) over the complete graph with arc cost m_{ij}. The tour thus forms an optimal discrete structure to explore the complete set of nodes while minimizing the sum of the lengths of the edges traversed during the tour.

In our model, matrices $1 - S^{[c]} = \left(1 - s^{[c]}_{ij}\right)$ and $1 - P^{[v]} = \left(1 - p^{[v]}_{kl}\right)$ follow constraints (1:a) and are therefore suitable inputs for the S-TSP procedure. Figure 6 illustrates the effect of column-reordering on a set distance matrix. The values over the diagonal m_{ii+1} are taken as step costs during the navigation and their overall sum is minimized. We have applied the above principle over a collection of Cultural Heritage digital items composed of images (paintings, historical photographs, pictures,...) annotated with description and metadata. We have defined several browsing dimensions using visual, temporal and textual characteristics. The result is a Web interface presenting a horizontal browsing dimension as its bottom line. From each of these items, a complementary vertical path is displayed. Image size is used to represent the distance from the main focal item displayed at the center of the bottom line. In essence, our browser uses a strategy closes to that implemented in [3].

Figure 7 shows the interface with the focal point in the bottom blue box. Green vertical and horizontal arrows materialize the paths that may be followed. In the upper right corner (dashed red box), a table displays the summary sample of the collection of items. Clicking on any of these images (but the central one) brings it to the center and updates its context (*i.e* computed neighborhood). Clicking on the central image goes back to the search interface as a complement to the navigation mode and displays the full details (*e.g* metadata) of this particular document. The choice of tours followed along the horizontal and vertical axis is regulated by setting options at each step of the browsing, thus allowing “rotations” around the focal point to display any combination of dimensions.

Concerning the modeling of potential attached social network ($\mathcal{P}$), we are planning to include an interface enabling the browsing of population formed by

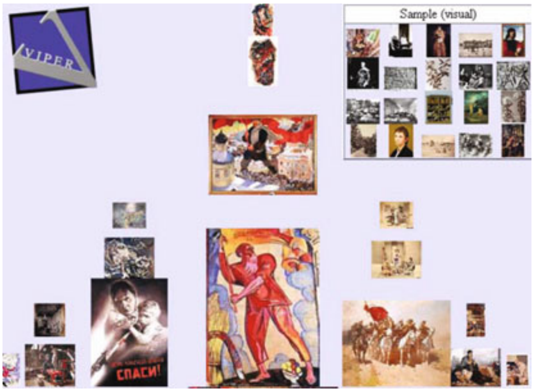


Fig. 7. The proposed browsing interface

the creators of the documents, in parallel with this document browsing interface. We have applied the very same principle with different features to browse meeting slide collections with respect to visual slide similarity (to identify reuse of graphical material), textual similarity (to relate presentations by topic and timeline (to simply browse thru the presentation). Again, a social of presentation authors may complement this document browsing tool.

5 Conclusion and Discussion

In this paper, we propose a modeling of data immersed into a social network based on multigraphs. We show how these multigraphs may be the base for defining optimal strategies for information discovery in complex and large documents collections, based on exploiting user interaction. Latent (topic) models are promising tools in this context. The choice of modeling may impose a particular model, based on the ease on interpretation of its parameters.

We also demonstrate that, via the definition of optimal structures, efficient exploitation of the network structure and contained data may be achieved. In particular, we advocate the use of the S-TSP for organizing a unique navigation path to be followed. Many graph-based structures are defined by NP-Complete problems. The problem of scalability thus forces proper approximations to be found.

Our work has been evaluated in all the steps of its engineering (*e.g* similarity computation). However, we still must complete our evaluation by the final usability of the produced interfaces. Initial demonstration sessions with credible tasks show encouraging interests from various classes of users (clients). Only quantitative tests over well-defined tasks and measures will tell us how much these interfaces are actually able to complement classical query-based search operations.

References

1. Bartholomew, D.J., Knott, M.: Latent variable models and factor analysis. Oxford University Press, Inc., New York (1999)
2. Cord, M., Gosselin, P.H.: Image retrieval using long-term semantic learning. In: IEEE International Conference on Image Processing (2006)
3. Craver, S., Yeo, B.-L., Yeung, M.: Multi-linearisation data structure for image browsing. In: SPIE Conf. on Storage and Retrieval for Image and Video DBs VII (1999)
4. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 4, 391–407 (1990)
5. Gaussier, E., Goutte, C.: Relation between plsa and nmf and implications. In: SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 601–602. ACM, New York (2005)
6. He, X., King, O., Ma, W.-Y., Li, M., Zhang, H.-J.: Learning a semantic space from user's relevance feedback for image retrieval. IEEE Transactions on Circuits and Systems for Video Technology 13(1), 39–48 (2003)
7. Hofmann, T.: Latent semantic models for collaborative filtering. ACM Transactions on Information Systems (TOIS) 22(1), 89–115 (2004)
8. Kabán, A., Girolami, M.A.: Fast extraction of semantic features from a latent semantic indexed text corpus. Neural Process. Lett. 15(1), 31–43 (2002)
9. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
10. Marchand-Maillet, S., Bruno, É.: Collection Guiding: A new framework for handling large multimediacollections. In: Audio-visual Content And Information Visualization In Digital Librairies, Cortona, Italy (2005)
11. Morrison, D., Bruno, E., Marchand-Maillet, S.: Capturing the semantics of user interaction: A review and case study. In: Emergent Web Intelligence. Springer, Heidelberg (2010)
12. Morrison, D., Marchand-Maillet, S., Bruno, E.: Semantic clustering of images using patterns of relevance feedback. In: Proceedings of the 6th International Workshop on Content-based Multimedia Indexing, London, UK, June 18-20 (2008)
13. Morrison, D., Marchand-Maillet, S., Bruno, E.: Modelling long-term relevance feedback. In: Proceedings of the ECIR Workshop on Information Retrieval over Social Networks, Toulouse, FR, April 6 (2009)
14. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Long-term learning from user behavior in content-based image retrieval. Technical report, Université de Genève (2000)
15. Nguyen, G.P., Worring, M.: Optimization of interactive visual similarity based search. ACM TOMCCAP 4(1) (2008)
16. Rubner, Y.: Perceptual Metrics for Image Database Navigation. PhD thesis, Stanford University (1999)
17. Szekely, E., Bruno, E., Marchand-Maillet, S.: High dimensional multimodal embedding for cluster preservation. Technical Report VGTR:0801, Viper - University of Geneva (2008)

Some Experiments in Evaluating ASR Systems Applied to Multimedia Retrieval

Julián Moreno[1], Marta Garrote[1], Paloma Martínez[1], and José L. Martínez-Fernández[2]

[1] Computer Science Department, Universidad Carlos III de Madrid, Avda. Universidad n 30,
28911, Leganés, Madrid, Spain
{jmschnei,mgarrote,pmf}@inf.uc3m.es
[2] DAEDALUS – Data, Decisions and Language S.A.
Avda. de la Albufera, 321
28031 Madrid, Spain
jmartinez@daedalus.es

Abstract. This paper describes some tests performed on different types of voice/audio input applying three commercial speech recognition tools. Three multimedia retrieval scenarios are considered: a question answering system, an automatic transcription of audio from video files and a real-time captioning system used in the classroom for deaf students. A software tool, RET (Recognition Evaluation Tool), has been developed to test the output of commercial ASR systems.

Keywords: Automatic Speech Recognition (ASR), Evaluation Measurements, audio transcription, voice interaction.

1 Introduction

There is a growing demand for services that improve access to information available on the web. The current trend in developing Information Retrieval (IR) systems focuses on dealing with any format (audio, video, images). These different formats not only appear in the objects collection to be searched but also in the user's queries. The existence of a huge amount of multimedia resources in the web requires powerful tools that allow the users to find them. These solutions exploit metadata related to the video, image or audio, using text based retrieval techniques. Although these techniques are advanced enough and show accurate results, other data formats, as video or audio, still need research. Techniques allowing a content based approach for these formats are still under development. The main goal is to retrieve a video, audio or image without using metadata or any other text related to the content.

For image contents, there are research efforts to make content based analysis, for example, the ImageCLEF track at the Cross Language Evaluation Forum[1]. Some of the image processing techniques exploit colour, brightness and other features to classify images, but some of them try to recognize shapes appearing in the image.

[1] http://www.clef-campaign.org

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 12–23, 2011.

Unfortunately, performance measures for this kind of analysis are still poor to allow some kind of widely used commercial application.

For audio contents, ASR techniques can be applied to produce textual transcriptions. In this way, conversion to text format is performed, in order to apply well known text retrieval techniques. This is the case of applications like Google Audio Indexer from Google Labs[2], which takes profit from audio transcription of videos using ASR technology, thus allowing to locate the point in a video or videos where the keyword written in the search box is mentioned. For the moment, this application only works for the English language and using videos of a specific domain (newscasts or politicians' talks). Nevertheless, there is a growing interest in the field of video and audio indexing. Google is not the only company developing products; other vendors in the market, such as Autonomy Virage[3], include tools to perform audio and video indexing.

In order to improve search and retrieval of audiovisual contents using speech recognition, it is necessary to evaluate the accuracy of ASR technology before using it for information retrieval applications.

The objective of this paper is twofold: firstly, evaluating the efficiency of speech recognition technologies in transcribing audio recordings from videos or audios to be indexed by an information retrieval system, such as a question answering system or a live subtitling application in a classroom; secondly, showing an evaluation tool, RET, developed to assist in testing the ASR technology in different application scenarios.

RET tool has been used in the evaluation of three ASR commercial products in (a) 160 short voice queries as input of a Question answering system in order to test a multimodal access (b) transcription of audio recordings from video resources with the aim of indexing them for further information extraction or information access and (c) live subtitling in an educational environment to help impaired students.

The paper is organized as follows: section 2 presents the related work; section 3 is devoted to explain the RET architecture and functionality, as well as the evaluation measures used; section 4 shows the experiments that have been performed in the three scenarios; an analysis of the results is shown in section 5; and finally, some enhancements are shown in section 6.

2 Related Work

The initial motivation which leads us to design and implement RET was the lack of a product covering all our needs. We were searching for a software tool that could provide us with measurements obtained from text comparison, to evaluate the efficiency of three speech recognition systems. There are several applications that have served as inspiration to solve our problem with the evaluation. One of these applications is DiffDoc [2], a texts comparison program which does not require a previous alignment: it does a direct comparison between files. The comparison process is similar to

[2] http://labs.google.com/gaudi

[3] http://www.virage.com/rich-media/technology/index.htm

RET's; both programs compare complex words and not simple characters, as most applications do (Winmerge [9], Altova Diffdog [1], Ultracompare [4]). An important advantage of DiffDoc [2] is the graphical interface, which shows the input files, allowing a visual comparison. The lack of numeric results after the comparison is the main disadvantage of this tool.

The *SCTK Scoring Toolkit* from National Institute of Standards and Technology of United States (NIST) [7] is a text comparison software specifically oriented to text-to-speech systems. The main differences between SCTK and RET are:

- RET interface displays original text (OT) and ASR output text (AOT), where the words that do not match are highlighted using colors. In this way, it is possible to carry out a qualitative study by linguists, apart from the quantitative one, in different application scenarios. It makes the application use easier.
- RET software supports several input formats such as XML (with or without temporal marks), TXT (plain text format, sentences format or TIME-TEXT format) and .SRT (**S**ub**R**ip Sub**T**itle files, with text and temporal marks). *NIST SCTK Scoring Toolkit* supports trn (transcript), txt (text), stm (segment time mark) and ctm (time marked conversation scoring) as input formats.
- The functionality of the algorithms used by both software tools is very similar. Regarding the input supported by the application, an adaptation of the algorithms functionality has been required. The algorithms of SCTK NIST and RET have not been compared as the input file formats from both systems are different and it was an unfeasible task.

3 Description of RET Tool

The RET architecture (Figure 1) is divided into three main modules, Graphical User Interface (GUI), IO File Manager and Text Comparator process.

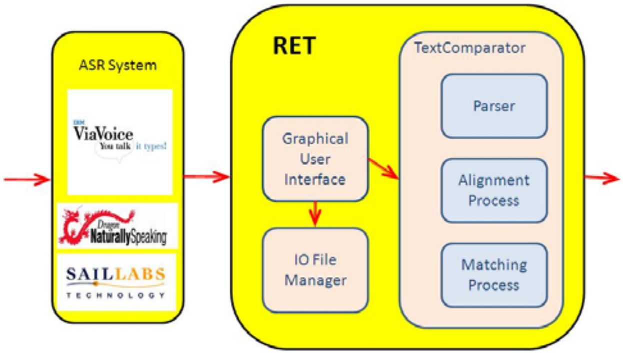
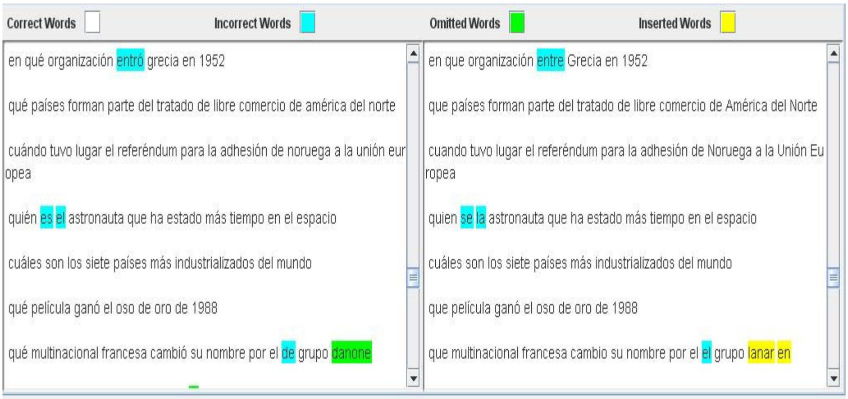


Fig. 1. Complete RET Tool Architecture



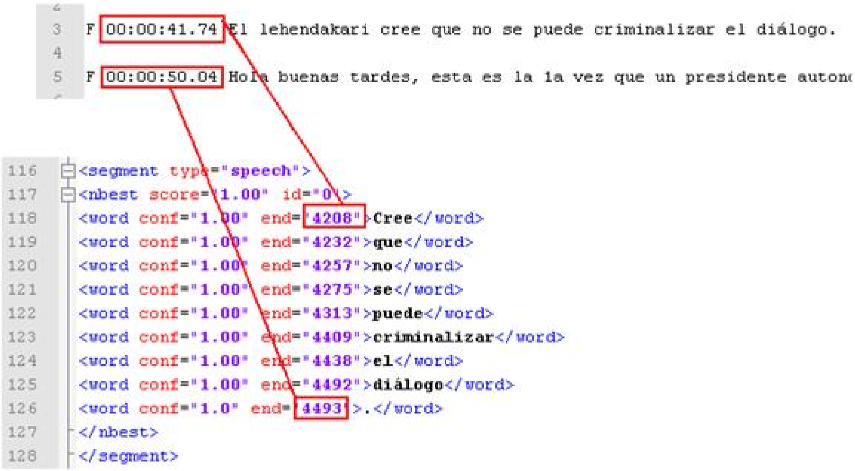
Correct Words □ Incorrect Words 🟦 Omitted Words 🟩 Inserted Words 🟨

en qué organización entró grecia en 1952	en que organización entre Grecia en 1952
qué países forman parte del tratado de libre comercio de américa del norte	que países forman parte del tratado de libre comercio de América del Norte
cuándo tuvo lugar el referéndum para la adhesión de noruega a la unión europea	cuando tuvo lugar el referéndum para la adhesión de Noruega a la Unión Europea
quién es el astronauta que ha estado más tiempo en el espacio	quien se la astronauta que ha estado más tiempo en el espacio
cuáles son los siete países más industrializados del mundo	cuáles son los siete países más industrializados del mundo
qué película ganó el oso de oro de 1988	que película ganó el oso de oro de 1988
qué multinacional francesa cambió su nombre por el de grupo danone	que multinacional francesa cambio su nombre por el el grupo lanar en

Fig. 2. RET Visual Results Image

Figure 2 shows the results of an evaluation example (on the left, the original text (OT) and on the right, the ASR output text (AOT)). A colours code is used to display the different types of errors: *white* for correct words, *blue* for incorrect words, *yellow* for inserted words and *green* for deleted words. Moreover, a tab with the numeric results, the numbers and a graphic bar chart are also shown.

TextComparator module compares OT and AOT files. It is made up of three different submodules: the parser, the alignment and the matching modules. Firstly, both files (OT and AOT) are parsed, obtaining two readable representations for the program to manage them. The procedure to obtain these objects is different depending on the format of the input file.



```
F 00:00:41.74 El lehendakari cree que no se puede criminalizar el diálogo.

F 00:00:50.04 Hola buenas tardes, esta es la 1a vez que un presidente autonó
```

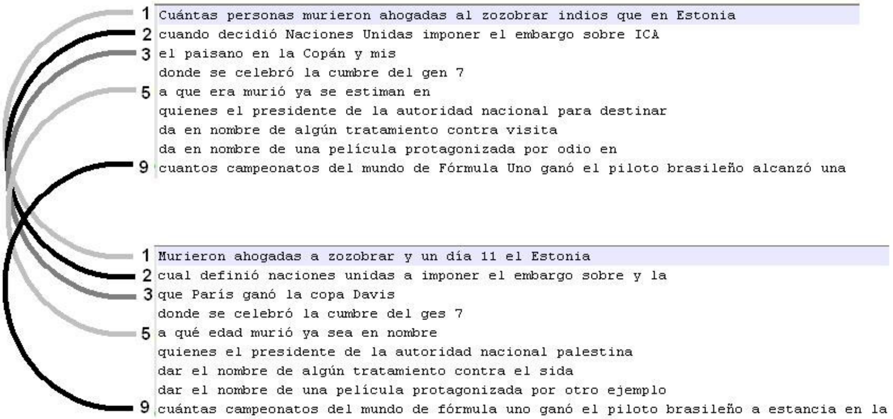
```
<segment type="speech">
<nbest score="1.00" id="0">
<word conf="1.00" end="4208">Cree</word>
<word conf="1.00" end="4232">que</word>
<word conf="1.00" end="4257">no</word>
<word conf="1.00" end="4275">se</word>
<word conf="1.00" end="4313">puede</word>
<word conf="1.00" end="4409">criminalizar</word>
<word conf="1.00" end="4438">el</word>
<word conf="1.00" end="4492">diálogo</word>
<word conf="1.0" end="4493">.</word>
</nbest>
</segment>
```

Fig. 3. Temporal alignment example

Once both files are parsed, the next step is aligning them. This process involves sentence matching of both texts, as, due to the different formats of the OT file (TXT) and AOT file (XML), the sequential order of elements of each sentence is not the same. Moreover, we must take into account those words that do not match up (because of any kind of recognizing error). This module aligns the texts obtained after parsing, and for this task, two different strategies are used: the first one aligns both texts by means of the temporal marks in the input file (an example is given in Figure 3); if there are no temporal marks, a positional alignment strategy is applied, to align by sentence or to obtain a plain text from the structured one.

The temporal alignment process takes the temporal marks from every sentence in OT and the temporal marks of every sentence in AOT. To align both texts, it is found a sentence in the AOT whose initial time is greater or equal to the initial time of OT sentence and whose final time is lower or equal to the final time of the OT sentence.

In the case of positional alignment, texts are aligned sentence by sentence. Figure 4 shows part of OT file and part of AOT file and their alignment. This means that the first sentence of OT is aligned with the first sentence of AOT, and so on.



```
1 Cuántas personas murieron ahogadas al zozobrar indios que en Estonia
2 cuando decidió Naciones Unidas imponer el embargo sobre ICA
3 el paisano en la Copán y mis
  donde se celebró la cumbre del gen 7
5 a que era murió ya se estiman en
  quienes el presidente de la autoridad nacional para destinar
  da en nombre de algún tratamiento contra visita
  da en nombre de una película protagonizada por odio en
9 cuantos campeonatos del mundo de Fórmula Uno ganó el piloto brasileño alcanzó una

1 Murieron ahogadas a zozobrar y un día 11 el Estonia
2 cual definió naciones unidas a imponer el embargo sobre y la
3 que París ganó la copa Davis
  donde se celebró la cumbre del ges 7
5 a qué edad murió ya sea en nombre
  quienes el presidente de la autoridad nacional palestina
  dar el nombre de algún tratamiento contra el sida
  dar el nombre de una película protagonizada por otro ejemplo
9 cuántas campeonatos del mundo de fórmula uno ganó el piloto brasileño a estancia en la
```

Fig. 4. Positional Alignment (Alignment by sentences)

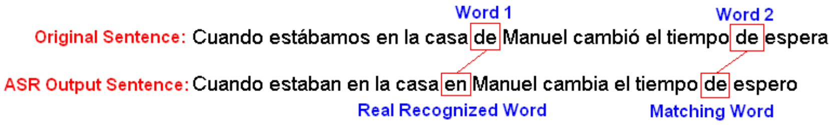
After the alignment, the comparison algorithm is applied. Both pre-processed texts (parsed and aligned) are compared in the matching module. The algorithm takes one word from the OT and compares it to the words from the AOT. When it finds a matching word, the algorithm makes a complementary search of matching words along both texts, which is necessary because of the specific domain (speech recognition). This algorithm avoids a matching between two words that should not match. The complementary search algorithm is explained in the next pseudo-code:

Algorithm 1. Complementary Matching Algorithm

Input: S1 list of words of OT, S2 list of words of AOT,
S1i word i from text S1, S2j word j from text S2,
A1 position of matched word in OT,
A2 position of matched word in AOT,
D1 distance in OT, D2 distance in AOT
Output: B Boolean indicating if the two compared words are the ones that must be compared, that is, if the result is positive (true) both words are correctly matched and if it is negative (false) the words must not be matched.

```
i = A1 {position of word from OT to use}
j = A2 {position of word from AOT to use}
repeat
    {find new matching word in OT}
    if (S1i equals S2j) then
        k = j
        repeat
            {find new matching word in AOT}
            if (S1i equals S2k) then
                    restart method with A1=i and A2=k
            end if
            k = k + 1
        while (k is minor than length of S2 AND k is minor than D2)
        j = k
    end if
    i = i + 1
while (i is minor than length of S1 AND i is minor than D1)
```

Figure 5 shows the application of the complementary search algorithm over two sentences: the first one is the OT sentence and the second one is the AOT sentence. The first step of the algorithm takes the word 'Cuando' from the OT sentence, which is compared to the first word in the AOT sentence, 'Cuando'. Both words are equal, so the algorithm counts that there is a correct word. Then, the next word in the OT sentence, 'estábamos', is taken. This word does not appear in the AOT sentence, so the counter for incorrect words is increased. The algorithm continues until the word 'de' in the OT sentence is reached (marked 'Word 1' in Figure 5). The following word in the AOT sentence is 'en', so the rest of the sentence is searched until the word 'de' is found (labeled Matching Word in figure 5). At this moment, the algorithm would indicate that the Matching Word is the transcription of Word 1, an incorrect matching. But the complementary matching algorithm continues checking if the word 'de' appears again in the OT sentence, to ensure that the matching is correct. It finds another word 'de' (labeled Word 2 in Figure 5) and it has to decide if it should be related to the Matching Word ('de' in the AOT sentence) instead of Word 1. In this situation, the algorithm searches for another 'de' word in the AOT sentence. It fails, so the



Original Sentence: Cuando estábamos en la casa de Manuel cambió el tiempo de espera

ASR Output Sentence: Cuando estaban en la casa en Manuel cambia el tiempo de espero

Fig. 5. Example of Complementary Search Process Profits

algorithm concludes that Word 2 should be linked to the Matching Word and not to Word 1. Thus, a transcription error for Word 1 should be counted. Of course, the algorithm can be parameterized to avoid dead ends while searching for matching words.

The standard measurements in speech recognition used to evaluate the ASR system are:

- *Correct* words rate = correct words (AOT) / words (OT) * 100.
- *Incorrect* words rate = incorrect words (AOT) / words (OT) * 100.
- *Omitted* words rate = not recognized words (wrong or right) (OT) / words (OT) * 100.
- *Inserted* words rate = inserted words (AOT) / words (OT) * 100.
- *Complete Correct Sentence Percentage* = number of sentences correctly recognized / number of sentences (OT).

SCTK NIST includes some other measurements besides these ones. For future versions, we planned to implement more measurements depending on specific user needs. Some examples of new measurements could be rates of specific words such as named entities, verbs, acronyms, etc.

4 Experiments

We introduce the experiments carried out with RET tool. All of them have been applied to the three different voice recognizers and with all the possible input formats.

The voice recognizers used for the experiments are IBM ViaVoice [3], Dragon Naturally Speaking [8] and Sail Labs' Media Mining Indexer (MMI) [6]. They are all commercial voice recognizers, and our aim was to compare them in order to choose the most appropriate one to accomplish tasks for other projects. ViaVoice 10.5 is a commercial speech recognizer that needs previous training by the user. It supports MS operating systems and incorporates keyboard shortcuts, specialized themes and the possibility of adding vocabulary. The output is word by word. We could not introduce audio files in ViaVoice, which meant a disadvantage to use it in some of the experimental scenarios. Dragon Naturally Speaking 9.0 shares many features with ViaVoice, but it does not accept specialized themes nor keyboard shortcuts. The Dragon Naturally Speaking output is phrase by phrase, needing several context words. ViaVoice and Dragon are speaker-oriented speech recognizers and they need a previous training process. However, we did not make a conventional training, but a multiuser one, using 10 different trainers, each one reading sentences from the basic text training provided by both programs. Finally, the main advantage of MMI version 5.0 is that it does not need previous training. The output is also phrase by phrase.

The scenarios where the program has been tested are two: a Question Answering System and an Audio-video Transcription System (divided in two sub scenarios), both in Spanish language (Castilian variety).

4.1 Question Answering System Scenario

As part of a biggest project on question answering, we tested the recognizers using as input 163 audio files containing questions read by 10 individuals (both sexes, different ages). They were short questions, asking information about important figures, celebrities, places, dates, etc. Some examples are: *Qué es BMW?(What is BMW?), Quién recibió el Premio Nobel de la Paz en 1989? (Who did win the Nobel Peace Prize in 1989?), Quién es la viuda de John Lennon? (Who is John Lennon's widow?), Cuándo se creó la reserva de ballenas de la Antártida? (When was the Antarctic whale reserve created?*. The recognizers were used to convert speech to text and later to send it to the question answering system.

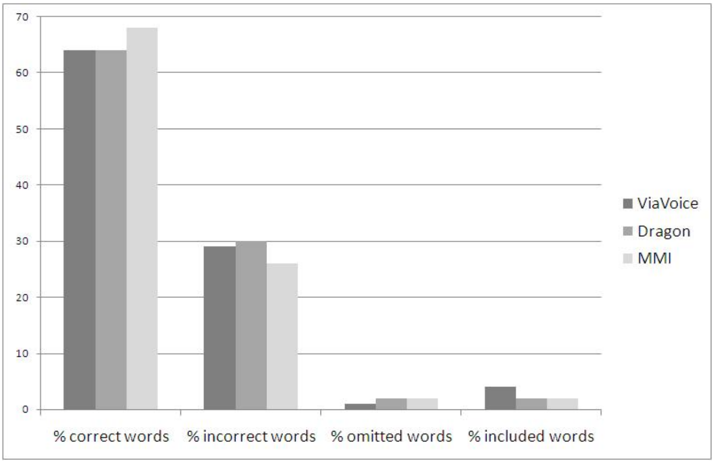


Fig. 6. Accuracy of the three speech recognizers in question answering scenario

The evaluation result of the recognition rate is shown in Figure 6. All systems are performing over a 60% of correct words rate. Structured texts have been used for testing[4].

The results of the evaluation provide numeric figures for the recognition rate, but not any accuracy value of the comparison between both texts. This can be seen in the graphical user interface. If we compare transcriptions with the OT using the visual results box, we can see that all speech recognizers are quite accurate due to the type of text (structured text).

[4] A structured text has a well-formed structure where sentences are systematically separated and can be easily parsed. An unstructured text has no defined structure or, even having a structure, the resulting separated sentences are too long to be considered.

4.2 Audio-Video Transcription

Video transcription for Information Retrieval

This work is focused on the use of a speech recognizer for making automatic transcriptions of audio and, subsequently, retrieving information from the resulting texts. For this task, the MMI was the chosen recognizer due to problems with ViaVoice to integrate audio files as input. As input, two newscasts video files were used; both of them last half an hour, and the difference between them is that while the first one is a national newscast, the 24h newscast addresses to an international audience. We made the comparison between the OT and the AOT from MMI to obtain measurements and assess the performance of the speech recognizer. The results are presented in Table 1.

Table 1. Results obtained with newscasts video files

	Newscast	Newscast 24h
% correct words	55	32
% incorrect words	32	48
% omitted words	7	9
% inserted words	3	9

In this case, the comparison process is carried out on structured texts, but these are formed by sentences long enough to be considered a mid-way point between structured and non-structured text scenarios. The comparison results are better than for the non-structured texts, but still worse than for the complete structured-texts. The difference between both results relies on the audio files used for the second test, which presented a higher noise level and this is reflected in the numeric results.

Real-time captioning system in a classroom

Another important scenario was a subtitling application for students with hearing impairment that transcribes the teacher's speech with the help of an ASR system, converting the spoken lesson into a digital resource. This content is available in real time for deaf students in form of captioning or as plain text, in paragraphs, where the user can navigate the whole transcription. A secondary task, apart from live subtitling, is the possibility of retrieving learning objects using subtitles to index video recorded in classrooms and helping students with disabilities in the learning process [5]. The evaluation has been carried out at the Carlos III University of Madrid during a 3th year subject of Computer Science degree called "Database Design". The teacher previously trained Dragon Naturally Speaking version 9 (DNS). Training duration was 30 minutes approximately, reading specific texts given by both ASR products. Additionally, specific vocabulary of "Database Design" subject was independently introduced and trained.

Four experiments were performed: (1) speech recognizer's basic model, (2) basic model and training, (3) basic model and specific vocabulary and (4) basic model, training and specific vocabulary. Figure 10 shows the figures provided by RET for the different tests.

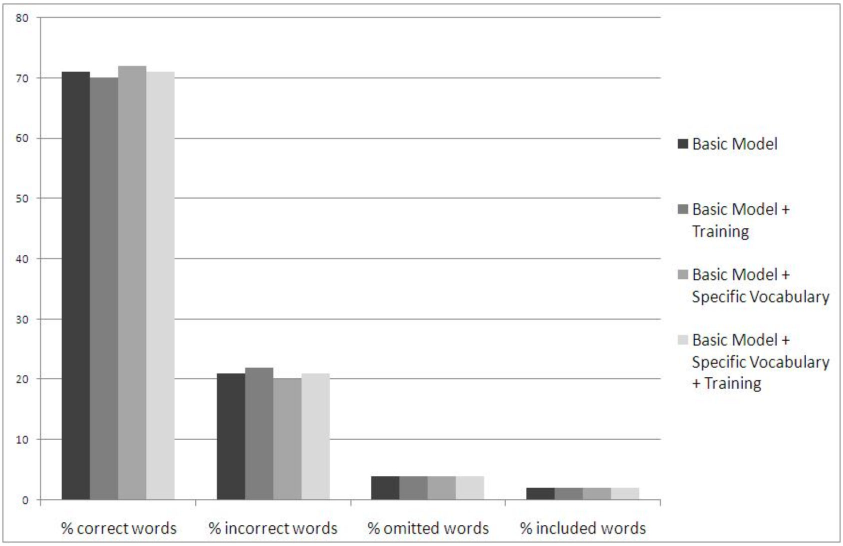


Fig. 7. Comparison of four tests in the real-time captioning scenario

The results obtained after the comparison show a high degree of accuracy for non-structured text, although it is usually poorer as the comparison process was not designed to work with this kind of texts.

As the algorithm does not work properly with non-structured texts, these results are due to a manual pre-processing of the texts, dividing them in two parts. Besides, the distances used in the 'complementary matching' algorithm were also adjusted to obtain the optimum value of the comparison results.

The scenario for this task (a classroom) involves dealing with spontaneous speech, even though the discourse is previously planned. This means the existence of typical elements of spontaneous speech as disfluences, self-interruptions, false starts, hesitations, all of which make the recognition process difficult. Owing to this fact, there is not much variation between the four tests, as training and vocabulary insertion do not provide better results. Moreover, keywords are not distinguished from stopwords, so, even introducing specific vocabulary, the total percentage does not improve as it is made up including stopwords.

5 Conclusions

Historically, the evaluation of ASR systems has been a quantitative evaluation, but also qualitative output is necessary and makes easier the task of testing an ASR system. Currently, SCTK software performs a quantitative evaluation, but it did not fit specific needs such as to work with XML file formats or a simple user interface to analyze transcription errors.

There are some features that distinguish RET software from SCTK. Firstly, the GUI is intuitive and friendly and makes the tool easier to use. The displayed results provide useful information, facilitating the interpretation task. RET supports different input file formats that fit our needs. And finally, measurements calculated by both systems are the same, but in our case we can easily increase the number of numeric results depending on specific needs.

The experiments are representative of the different text types which the software can deal with. For every experiment one of them has been used: (1) Structured text; (2) Unstructured text; and (3) Midway point text[5].

The numeric results from the ASR recognition rates, which are those given by the RET software, do not depend on the type of text. Texts features affects the quality of the comparison, being higher for structured texts, lower for midway-point texts and presenting the worst results for unstructured texts. This is consistent with the fact that the algorithm was design to work with structured texts and later adapted to deal with unstructured texts.

6 Future Work

As future work, one of the main improvements planned for RET is the increase of the number of evaluation measurements. Furthermore, several improvements are: (a) adding PoS (Part-of-speech) tagging to transcriptions to analyze which are the most problematic kind of words, for instance, named entities, verbs, acronyms, etc; (b) taking into account the length of words and sentences and (c) dividing the sentences into long and short, establishing a threshold to delimit them. Another important enhancement will be the creation of an output report with the comparison and the evaluation results.

Regarding the algorithms, future work aims the following:

- Allowing the user to keep a results history and establishing a fixed storing protocol for text comparison.
- Improving the comparison algorithm to manage continuous text (unstructured-text) or at least structured long texts. The use of punctuation marks could be useful for both the alignment and the comparison algorithm.
- Polishing the alignment algorithm, since increasing the accuracy of aligned texts, the comparison results will improve noticeably. Also, we must solve the problem of alignment for plain texts without temporal marks.
- Improving the 'Complementary Matching' algorithm to develop an automatic way to obtain the optimum values for the algorithm.

Acknowledgments. This research work has been supported by the Regional Government of Madrid under the Research Network MA2VICMR (S2009/TIC-1542) and by the Spanish Ministry of Education under the project BRAVO (TIN2007-67407-C03-01).

[5] Midway point text: structured text in which the text is long enough to be considered as unstructured one.

References

1. Altova, Altova DiffDog, http://www.altova.com/products/diffdog/diff_merge_tool.html (viewed July 2010)
2. DiffDoc, Softinterface Inc., http://www.softinterface.com/MD/Document-Comparison-Software.htm (viewed July 2010)
3. IBM ViaVoice, http://www-01.ibm.com/software/pervasive/embedded_viavoice (viewed July 2010)
4. IDM Computer Solutions, Inc., UltraCompare, http:// www.ultraedit.com/loc/es/ultracompare_es.html (viewed June 2010)
5. Iglesias, A., Moreno, L., Revuelta, P., Jimenez, J.: APEINTA: a Spanish educational project aiming for inclusive education In and Out of classroom. In: 14th ACM–SIGCSE Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2009), Paris, July 3-8, vol. 8 (2009)
6. Media Mining Indexer, Sail Labs Technology, http://www.sail-technology.com/products/commercial-products/media-mining-indexer.html (viewed June 2010)
7. NIST, Speech recognition scoring toolkit (SCTK) version 1.2c. (2000), http://www.nist.gov/speech/tools (viewed May 2010)
8. Nuance Dragon Naturally Speaking, http://www.nuance.com/naturallyspeaking/products/whatsnew10-1.asp (viewed March 2010)
9. WinMerge, http://winmerge.org/ (viewed June 2009)

Instant Customized Summaries Streaming: A Service for Immediate Awareness of New Video Content

Álvaro García-Martín, Javier Molina, Fernando López, Víctor Valdés,
Fabrizio Tiburzi, José M. Martínez, and Jesús Bescós

Video Processing & Understanding Lab. – Universidad Autónoma de Madrid
{Alvaro.Garcia,Javier.Molina,F.Lopez,Victor.Valdes,
Fabricio.Tiburzi,JoseM.Martinez,J.Bescos}@uam.es

Abstract. This paper presents the Instant Customized Summaries Streaming service, a multimedia service able to deliver customized video summaries with a minimum delay after the original content has been uploaded to a video repository. Uploaded videos start being analyzed, via on-line real-time algorithms. As analysis results are available, video summaries are then generated also in an on-line scheme, adapted to different available terminal and finally streamed to the subscribed clients. The whole chain works in on-line mode so all the processes can be simultaneously executed without waiting for any of them to have finished its operation. A prototype of this service, accessible via a web interface, has been fully implemented.

Keywords: Video Adaptation, On-Line Processing, Video Summarization, Video On demand, Video Repositories, MPEG-7, MPEG-21.

1 Introduction

Nowadays users demand multimedia content adapted to the users' terminals and personalized according to their preferences. Some adaptation systems can be found such as [1], which performs MPEG-21 based adaptation and considers user preferences. In [2] a generic framework for analysis, adaptation and user preferences accomplishment is described. [3] introduced the affective content concept to enhance personalization. Several systems exist which consider video summarization and video streams adaptation but we are not aware of any system performing both operations in a synchronized way and what is more important doing it in on-line mode.

This paper presents a video personalization, summarization and streaming system that is able to provide access to a video repository via heterogeneous usage environments; the service provided has been named Instant Customized Summaries Streaming (ICSS) service. This service has been designed as an add-on (plug-in) to be used over an existing video repository. As a new video is uploaded to the repository, the ICSS launches an on-line summarization process, which allows to be streamed to the list of subscribed users while it is being generated. An adaptation mechanism is integrated within the system; currently performs adaptation to different user presentation preferences and terminal consumption capabilities, and it can be easily extended to additional usage environment restrictions.

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 24–34, 2011.

The most innovative aspect of the ICSS service is the possibility for the user to start watching a video summary as soon as the original unannotated content has been uploaded to the video repository, as the summary is being produced and adapted to the user's terminal and presentation preferences. The ICSS infrastructure can be further used to offer additional services over the video repository (e.g. providing enhanced personalized summaries based on annotations over the video repository).

In order to present the service, which is currently implemented over news videos, Section 2 presents a functional overview of the ICSS system, followed by Section 3, which presents the content representation model, Section 4 that details the different modules of the system, and Section 5 presenting the data flow among modules. After this presentation of the whole system, the current implementation of the ICSS service is presented in Section 6, before drawing some conclusions and future work in Section 7.

2 Functional Overview

The ICSS service is implemented by the ICSS System, which considers two actors:

- Content providers: users who upload videos to the ICSS Server.
- End users: content consumers informed of the availability of new videos as soon as they are uploaded.

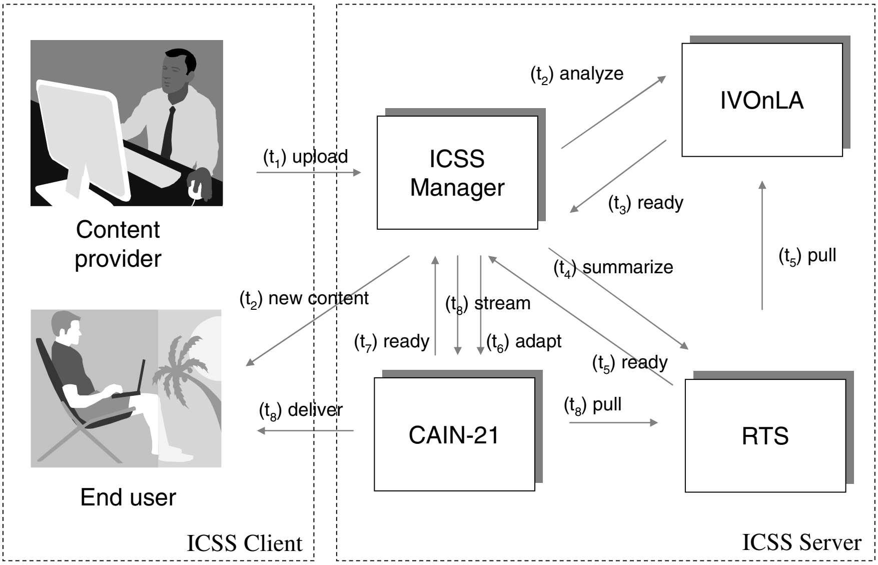


Fig. 1. Overview and functional decomposition of the ICSS System

The ICSS System can be roughly divided into the ICSS Client and the ICSS Server. The term ICSS Client refers to the set of web pages that the content providers and the end users operate in order to get access to the ICSS Server functionalities. The ICSS Server contains four fundamental modules (see Fig. 1.):

- ICSS Manager: this is a control module responsible of interacting with the users, requesting services to the other modules and managing their execution order.
- IVOnLA: refers to an Image and Video Online Analysis module. It provides a low-level yet semantically meaningful content-based description of the input media in real time and in an on-line mode. This description consists of a set of descriptors which intend to be used by any higher level analysis or processing module, in this case by the RTS.
- RTS: refers to a Real-Time Summarization module. It uses the IVOnLA descriptions to generate, in an on-line way, a video summary which is served to the CAIN-21.
- CAIN-21: refers to a Content Adaptation and INtegration adaptation module [4], which retrieves summarized video from the RTS module, adapts it to the terminal and user's preferences, and streams it to the consumer's terminal.

The ICSS System works as it follows: (t_1) the content provider starts by uploading a new content to the ICSS Manager. (t_2) The ICSS Manager requests the IVOnLA to start analyzing the content; in parallel, it notifies the clients the availability of new content. (t_3) When IVOnLA is ready to start serving low level descriptions, it informs the ICSS Manager. (t_4) The Manager launches the RTS. (t_5) The RTS starts retrieving descriptions from the IVOnLA and on-line generating the summary; it informs the ICSS Manager as soon as this summary is ready to start being served. (t_6) The Manager launches the CAIN-21. (t_7) Shortly afterwards, the CAIN-21 notifies to the ICSS Manager the availability to adapt the video summary. (t_8) Finally, either automatically or according to a user request, the Manager commands the CAIN-21 to start streaming the video summary. The RTS and CAIN-21 intercommunicate via a pull scheme, in which the requested unit consists on a frame and the associated audio samples.

The whole chain works in an on-line mode, so there is no need for any of the modules to finish its operation before launching the other modules. A detailed description of the type of data managed and served by each module is provided in Section 5.

3 Content Representation

The ICSS service makes use of MPEG-21 *Digital Items* (DIs) [5] to represent videos and metadata. As soon as a news item is uploaded to the video repository, both the video and its associated metadata are stored as a standard DI.

The original video is represented by means of a *Component* element. This *Component* makes reference to the video resource (*Resource* element) and its corresponding metadata (*Descriptor* element). Specifically, the *Descriptor* includes a *MediaInformationType* element (MPEG-7 part 5 [6]) with video format information. The genre of the item (such as politics, economics, sports...) is set via a *ClassificationType* element (MPEG-7 part 5).

The DI contains a *VariationType* element (MPEG-7 part 5) in order to relate the original *Component* of the DI with its variations: the video summaries. The *VariationRelationship* element indicates the kind of *Component*: *original* or *summarization*. In the current version only one original *Component* is permitted; however, it considers as many *Component* elements tagged *summarization* as available usage terminals contains the *Usage Environment Description Repository* (UED Repository) [4] that

CAIN-21 implements. Each summarized *Component* includes *MediaInformationType* elements with different screen sizes or video formats.

4 Server Modules Description

This section provides a detailed description of the modules integrated into the ICSS Server (See Fig. 2).

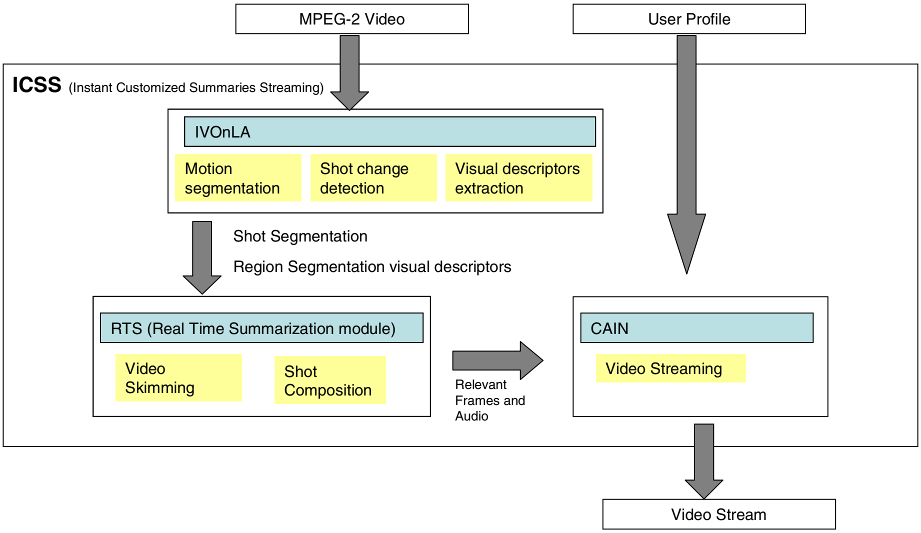

MPEG-2 Video
User Profile
ICSS (Instant Customized Summaries Streaming)
IVOnLA
Motion segmentation
Shot change detection
Visual descriptors extraction
Shot Segmentation
Region Segmentation visual descriptors
RTS (Real Time Summarization module)
Video Skimming
Shot Composition
Relevant Frames and Audio
CAIN
Video Streaming
Video Stream

Fig. 2. ICSS Server architecture and functional diagram of its modules

4.1 IVOnLA

This module is in charge of extracting a set of semantically meaningful descriptors from the new uploaded content as well as serving these descriptors to other modules. The analysis techniques employed are intended to achieve real-time performance by focusing on both on-line operation and MPEG compressed domain analysis.

On-line operation guarantees that no delay is added due to long-time media dependencies, eases achieving real-time performance and allows the synchronous integration of the analysis stage with the rest the modules in the ICSS System, which can therefore operate at the same time the content is being analyzed. In this sense, analysis follows a causal scheme: when extracting characteristics from frame *k,* only information coming from the frames *0* to *k-1* need to be available.

Compressed domain analysis has been adopted for two main reasons. First, the huge amount of data necessary for representing uncompressed high resolution video (such as Digital Television content: 720x576, 30 fps) makes its analysis extremely computationally demanding. Compressed domain approaches exploit the fact that relevant information is concentrated in a small fraction of the compressed video stream. A second reason is that a considerable amount of information derived during the coding process and available in the compressed stream (*e.g.* MPEG motion vectors) can be readily used in analysis algorithms, thus saving computation time.

In addition to some meaningful MPEG compressed domain information that might result useful for further analysis (DC images, AC coefficients and motion vectors) the set of features extracted by this module include video shot boundaries, a parametric characterization of the global motion field and a series of segmentation masks indicating moving objects and text captions.

Regarding shot detection, numerous proposals have shown to achieve a great level of accuracy in case of abrupt shot changes, yet gradual transitions are still an important challenge in this area; [12] includes a broad state-of-the-art review on this topic. Specifically, the shot detection mechanism used within the IVOnLA is based on the algorithm proposed in [14], in which the paralepipedic classification procedure has been replaced by an SVM [13] with a Radial Basis Function kernel.

Dominant motion is usually characterized by the parameters values of a certain model which best fits the optical flow of every frame. The *geometric* (4 parameters), the affine (6 parameters) and the bilinear (8 parameters) models are the most popular, since they can all reasonably represent most of the camera motion patterns. In case of compressed domain approaches [15][16][17] the MPEG motion vectors are considered as a reasonable estimation of the frame's optical flow; existing works mainly differ in the ways to achieve robustness against outlier vectors. For efficiency, IVOnLA performs global motion estimation simultaneously to motion objects segmentation, via an iterative fitting scheme.

There are many reported approaches to moving objects segmentation in the compressed domain [18][19][20]. The technique implemented in the IVOnLA is based on an efficient approach described in [18], in which robust dominant motion estimation and temporal coherence constraints are used to separate moving objects from the background. In order to improve the segmentation results of this technique several modifications over the base algorithm were proposed [21], namely a new scheme to better handle the intra-coded macroblocks, an improved mechanism to exploit the temporal coherence of the objects, a procedure to refine the motion masks with the colour information, and some initial ideas to handle dynamic backgrounds.

Detection of text captions in video frames exploits specific properties of these elements (strongest contrast between text and background, temporal persistency, static location, horizontal alignment,…). IVOnLA implements the approach described in [22] for caption detection in the compressed domain. Integration of our work on caption recognition is still pending due to the low efficiency of the freely available OCRs.

4.2 RTS

This module is in charge of the on-line/real-time generation of multimedia summaries.

The RTS automatically detects the anchorperson shots at the beginning of a news story, based on a fast face detection algorithm implemented in the OpenCV library[1] and features extracted by the IVOnLA module. These shots are kept for a later composition of a reduced size anchorman window overlapped over a video skim of the remaining shots (see Fig. 3).

[1] http://sourceforge.net/projects/opencvlibrary/

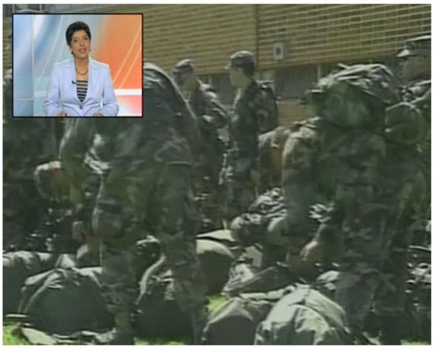


Fig. 3. Frame layout of a RTS generated summary

Video skimming is carried out in an on-line manner, so that partial results are generated while the original video is still being processed. This on-line approach implies a number of restrictions with respect to off-line ones: the small required delay, the progressive summary generation and the lack of complete information about the incoming video (i.e., specific content and length) have, in principle, a negative impact in the quality of the results and complicate the control of the summary length. The RTS follows the algorithm described in [9], which provides a flexible on-line summarization framework based on the construction of binary trees that model the potentially generable video summaries at a given time and rank them. This approach evolved from an instantaneous decision, inclusion or discard, for each incoming video fragment [10] to a limited buffering of *n* incoming fragments, which allows to choose the best of them. The higher is *n*, the better is the selection accuracy, at the expense of a higher complexity (2^n possible combinations of fragments must be evaluated for its inclusion) and delay in the summary generation.

For each incoming video fragment two nodes are appended to the previously existing nodes of the so called 'summarization tree' (see Fig. 4). Such nodes represent the inclusion or discard of each fragment in the output summary. Each branch represents a possible video summary which is scored according to the following criteria:

- Size: relative size of the resulting video summary, calculated by considering the number of inclusion/discard nodes on each branch.
- Continuity: considered to generate more pleasant video summaries. It is measured computing the ratio of consecutive video fragments included in the summary, which aims to generate video summaries as smooth as possible (avoiding too many discontinuities).
- Redundancy: the main purpose of most summarization approaches consists on redundancy elimination. For this purpose, the similarity of each video fragment with respect to other fragments included in the same tree branch is calculated.
- Activity: it is commonly considered to give a higher inclusion priority to fragments with high activity, which is related to a higher probability to include 'relevant' events. In this case the motion activity of the fragments included on each branch is averaged.

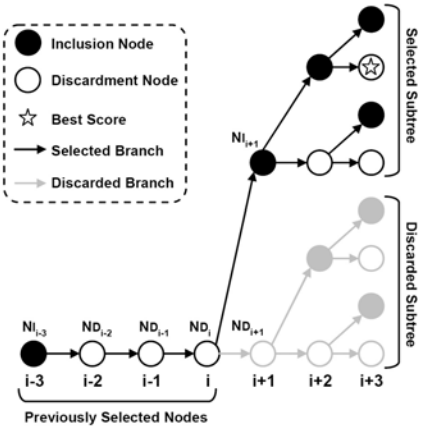


Fig. 4. Dynamic Summarization Tree Generation

Considering only small sub-trees instead of generating the complete summarization tree for the whole video allows to generate summaries in a progressive way and to limit the computational complexity. The depth of the summarization tree and the number of branches are limited by the selection of the subtrees with the higher score paths and the elimination of those with lower scores. Higher depth and tree leaf limit implies better quality summaries (as more possible summaries are evaluated) but also higher computational requirements (processing time and memory consumption). The proposed algorithm allows the usage of appropriate summarization tree parameters according to each situation. Further details can be found in [9], and information about the results obtained in TRECVID 2008 BBC Rushes Summarization Task in [11].

The implemented approach provides customization capabilities in terms of performance (computational cost vs. obtained summary quality) but also enables personalization in terms of generated summary characteristics. As aforementioned, summaries are generated based on size, continuity, redundancy and activity. Each feature is normalized and weighted to compute the score for each possible video summary. Weights can be user defined so as to balance the size of the output video skim, its smoothness, preference for static or high activity fragments and lack of redundancy. An additional relevance curve provided by a prototype On-Line Personalization Module (see section 7) has also been considered via an additional personalization weight, which might be prioritized respect to the previously considered criteria.

4.3 CAIN-21

This is a metadata-driven multimedia adaptation engine that facilitates the integration of multimedia adaptation operations and manages its execution. CAIN-21 [4] is an adaptation framework in which different multimedia adaptation tools can be integrated and tested. The source code along with an online demo of its functionality can be publicly accessed at [7].

Within the ICSS Server, CAIN-21 retrieves video from the RTS and delivers it to the user's terminal. Before the media stream delivery process starts, the capabilities of the available terminals should be accessible in the *UED Repository*.

If the generated media stream is not suitable for the terminal capabilities, an adaptation operation is performed as content is summarized and delivered. The CAIN-21 module also decides on the necessity of performing transcoding over the resulting summaries before delivery. Depending on the capabilities of the terminal, it decides on the more convenient delivery mechanism to employ. The result of such decision is a sequence of *Content Adaptation Tools* (CATs) along with each tool's parameters.

In order to perform the adaptation and streaming processes CAIN-21 makes uses of two CATs. *RawVideoCombinerCAT* retrieves video from the RTS and begins adaptation according to the terminal capabilities. Simultaneously *HttpVideoStreamServer* serves the adapted content through a web interface over HTTP protocol. The whole process is set up via the configuration parameters of each CAT.

5 Data Flow among Modules

Apart from the controlling messages described in Section 2, an inter-module communication scheme allows the exchange of processing data. This online data exchange is based on TCP/IP and works in a server/client way, so that modules act as servers for making their results available and as clients when they feed from other module results. In this section we describe the nature of the data served by each module.

The analysis module, IVOnLA, generates real time descriptions that are served to the RTS, for each frame, on the following packages:

- Light Descriptions: non bulky fixed-size structures with information about shot boundaries, motion activity and global motion.
- DC Images: images made up of the DC coefficients of the 8x8 DCT blocks (directly available in MPEG intracoded frames and interpolated in case of intercoded ones), which provide a low resolution representation of the video.
- Coarse Moving Object Masks, delineating frame areas whose motion differs from the frame dominant tendency along several frames, therefore spotting potential moving objects. Details of the employed algorithm are described in [21]
- Caption Detection Masks, isolating frame areas potentially enclosing text captions. These areas can later be used for text recognition or simply to ascertain frame relevance (considering, for example, incoming captions highly correlated with new starting events). A detailed description of the used algorithm is given in details are covered in [22]
- AC Images: The AC coefficients of the DCT blocks in the MPEG intracoded frames, which can be used to infer textural properties of frame regions.

The RTS module serves summarized video and audio with a frame resolution. This is, for each frame of the summarized video the CAIN-21 receives its information (visual and audio) to perform the streaming.

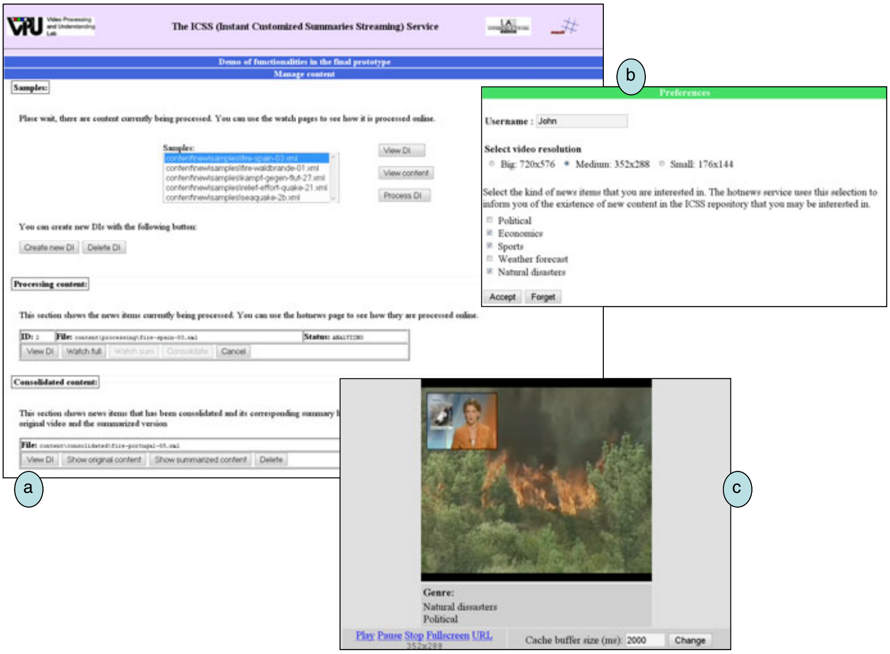
CAIN-21 deals with standard DIs. Unadapted DIs contain only one *Component* element with the original video. Summarized DIs contain both the original *Component* and its variations. The RTS module output produces variation elements containing raw video and raw audio. The purpose of CAIN-21 is to adapt these raw video *Component* elements to the terminal constraints. CAIN-21 generates one adapted *Component* element for each terminal profile. However, this adaptation (which is a time consuming operation) is only executed on-demand, i.e., if the client requests the video. In this moment the *RawVideoCombinerCAT* performs the real media adaptation.

6 Results

An ICSS prototype has been implemented using a web interface (see Fig. 5) to provide the content manager and the watching accesses to the ICSS service functionality. The content provider interface (see Fig. 5-a) allows managing the new content, the content being processed and the consolidated content. As an end user, before having access to content, you are asked to specify your user preferences (see Fig. 5-b). Finally, when the content is ready, a video player pops up from the internet navigator through the client interface (see Fig. 5-c).

By maintaining the name and preferences of the user, the ICSS prototype provides personalized genre-based consumption of summaries and adaptation to the current user's terminal.

In terms of computational efficiency a set of real-time (for input videos of 25fps and 720x576 pixels) working modules has been developed. This, combined with our online data exchange schema, permits the global system to work also in real-time. The modular design permits the distribution of the computational charge among different


The ICSS (Instant Customized Summaries Streaming) Service

Demo of functionalities in the final prototype

Manage content

Samples:

Plase wait, there are content currently being processed. You can use the watch pages to see how it is processed online.

Samples:
content\news\samples\fire-spain-03.xml
content\news\samples\fire-waldbrande-01.xml
content\news\samples\kampf-gegen-flut-27.xml
content\news\samples\relief-effort-quake-21.xml
content\news\samples\seaquake-2b.xml

You can create new DIs with the following button:

Processing content:

This section shows the news items currently being processed. You can use the hotnews page to see how they are processed online.

ID: 2 File: content\processing\fire-spain-03.xml Status: ANALYZING

Consolidated content:

This section shows news items that has been consolidated and its corresponding summary l
original video and the summarized version

File: content\consolidated\fire-portugal-05.xml

Preferences

Username : John

Select video resolution
○ Big: 720x576 ● Medium: 352x288 ○ Small: 176x144

Select the kind of news items that you are interested in. The hotnews service uses this selection to inform you of the existence of new content in the ICSS repository that you may be interested in.

□ Political
☑ Economics
☑ Sports
□ Weather forecast
☑ Natural disasters

Genre:
Natural dissasters
Political

Cache buffer size (ms): 2000

352x288

Fig. 5. ICSS Web Interface: a) Manager Interface. b) Client Preferences Interface. c) Client Interface.

computers, making possible scalability to any number of users. Furthermore, the inclusion of new modules (or the improvement of existing ones) was taken into account during the system design, resulting easy to accomplish new application requirements.

7 Conclusions and Future Work

The ICSS service is a multimedia service that enables instant on-line viewing of news items. From the end user's point of view, the most innovative option is the "hotnews" feature that enables watching summarized and adapted content as soon as it is available in the ICSS Server. From a research point of view, the ICSS is supported by innovative video analysis, summarization and adaptation algorithms integrated on a real-time, scalable and distributed framework.

The end user can provide preferences regarding the summarization and adaptation of the content. Currently, summarization algorithms only deal with news items, and the content provider must indicate the content domain before uploading it. In a near future summarization will be extended to other domains and domain classification will be done automatically. In the same way, adaptation algorithms currently perform a limited set of transformations which will be extended to broader users' restrictions.

Future work also considers the development of two new modules aimed to enhance the ICSS service: a Real-Time Visual Classification module (RTVC) which will be in charge of the detection of high level semantic concepts useful for performing tasks such as personalization or summarization; and an OnLine Personalization module, which will generate the aforementioned relevance curve associated to the original video, taking into account the classification information obtained by the RTVC as well as a set of user defined preferences.

Acknowledgments

This work has been supported by the European Commission (IST-FP6-027685 - Mesh), Spanish Government (TEC2007-65400 - SemanticVideo), the Comunidad de Madrid (S-0505/TIC-0223 - ProMultiDis-CM), the Consejería de Educación of the Comunidad de Madrid and the European Social Fund.

References

1. De Bruyne, S., De Schrijver, D., De Neve, W., Van Deursen, D., Van de Walle, R.: Enhanced shot-based video adaptation using MPEG-21 generic bitstream syntax schema. In: Proc. of IEEE Symposium on Computational Intelligence in Image and Signal Processing, CIISP 2007, pp. 380–385 (2007)
2. Chang, S.F., Vetro, A.: Video adaptation: Concepts, technologies, and open issues. Proceedings of the IEEE 93(1), 148–158 (2005)
3. Kim, S., Yoon, Y.: Universal video adaptation model for contents delivery using focus-of-choice model. In: Proc. of Second International Conference on Future Generation Communication and Networking, FGCN 2008, vol. 1, pp. 46–49 (2008)

4. López, F., Martínez, J.M., García, N.: CAIN-21: An extensible and metadata-driven multimedia adaptation engine in the MPEG-21 framework. In: Chua, T.-S., Kompatsiaris, Y., Mérialdo, B., Haas, W., Thallinger, G., Bailer, W. (eds.) SAMT 2009. LNCS, vol. 5887, pp. 114–125. Springer, Heidelberg (2009)
5. ISO/IEC 21000-2:2004, Information technology - Multimedia framework (MPEG-21) - Part 2: Digital Item Declaration (2004)
6. ISO/IEC 15938-5:2003, Information technology - Multimedia content description interface - Part 5: Multimedia description schemes (2003)
7. `http://cain21.sourceforge.net/`
8. Hauptmann, A., Yan, R., Lin, W.-H.: How many high-level concepts will fill the semantic gap in news video retrieval? In: Proc. of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007, pp. 627–634 (2007)
9. Valdés, V., Martínez, J.M.: Binary tree based on-line video summarization. In: Proc. of ACM Multimedia 2008 (2nd ACM TRECVid Video Summarization Workshop), pp. 134–138 (2008)
10. Valdés, V., Martínez, J.M.: On-line Video Skimming Based on Histogram similarity. In: Proc. of ACM Multimedia 2007 (1st ACM TRECVid Video Summarization Workshop), pp. 94–98 (2007)
11. Over, P., Smeaton, A.F., Awad, G.: The TRECVID 2008 BBC rushes summarization evaluation. In: Proc. of ACM Multimedia 2008 (2nd ACM TRECVid Video Summarization Workshop), pp. 1–20 (2008)
12. Cotsaces, C., Nikolaidis, N., Pitas, I.: Video Shot Detection and Condensed Representation. IEEE Signal Processing Magazine 23(2), 28–37 (2006)
13. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
14. Bescós, J.: Real-time Shot Change Detection over On-line MPEG-2 Video. IEEE Tr. On Circuits and Systems for Video Technology 14(4), 475–484 (2004)
15. Durik, M., Benois-Pineau, J.: Robust motion characterisation for video indexing based on mpeg2 optical flow. In: Proc. of International Workshop on Content-Based Multimedia Indexing, CBMI 2001, pp. 57–64 (2001)
16. Yu, T., Zhang, Y.: Retrieval of Video Clips using global motion information. Electronic Letters 37(14), 893–895 (2001)
17. Zhu, X., Elmagarmid, A.K., Xue, X., Wu, L., Catlin, A.C.: InsightVideo: Towards hierarchical video content organization for efficient browsing, summarization and retrieval. IEEE Trans. on Multimedia 7(4), 648–666 (2005)
18. Mezaris, V., Kompatsiaris, I., Boulgouris, N., Strintzis, M.: Real-Time Compressed-Domain Spatio-temporal Segmentation and Ontologies for Video Indexing and Retrieval. IEEE Tr. On Circuits and Systems for Video Technology 14(5), 606–621 (2004)
19. Venkatesh Babu, R., Ramakrishnan, K., Srinivasan, S.: Video Object Segmentation: A Compressed Domain Approach. IEEE Tr. On Circuits and Systems for Video Technology 14(4), 462–474 (2004)
20. Wang, Z., Liu, G., Liu, L.: A fast and accurate video object detection and segmentation method in the compressed domain. In: Proc. of the 2003 International Conference on Neural Networks and Signal Processing, vol. 2, pp. 1209–1212 (2003)
21. Escudero, M., Tiburzi, F., Bescos, J.: Mpeg video object segmentation under camera motion and multimodal backgrounds. In: Proc of. IEEE International Conference on Image Processing, ICIP 2008, pp. 2668–2671 (2008)
22. Márquez, D., Bescós, J.: A Model-Based Iterative Method for Caption Extraction in Compressed MPEG Video. In: Falcidieno, B., Spagnuolo, M., Avrithis, Y., Kompatsiaris, I., Buitelaar, P. (eds.) SAMT 2007. LNCS, vol. 4816, pp. 91–94. Springer, Heidelberg (2007)

Towards Reliable Partial Music Alignments Using Multiple Synchronization Strategies

Sebastian Ewert[1], Meinard Müller[2], and Roger B. Dannenberg[3]

[1] Bonn University, Bonn, Germany,
Multimedia Signal Processing Group
ewerts@iai.uni-bonn.de
[2] Saarland University and MPI Informatik,
Saarbrücken, Germany
meinard@mpi-inf.mpg.de
[3] Carnegie Mellon University, Pittsburgh, USA,
School of Computer Science
rbd@cs.cmu.edu

Abstract. The general goal of music synchronization is to align multiple information sources related to a given piece of music. This becomes a hard problem when the various representations to be aligned reveal significant differences not only in tempo, instrumentation, or dynamics but also in structure or polyphony. Because of the complexity and diversity of music data, one can not expect to find a universal synchronization algorithm that yields reasonable solutions in all situations. In this paper, we present a novel method that allows for automatically identifying the reliable parts of alignment results. Instead of relying on one single strategy, our idea is to combine several types of conceptually different synchronization strategies within an extensible framework, thus accounting for various musical aspects. Looking for consistencies and inconsistencies across the synchronization results, our method automatically classifies the alignments locally as reliable or critical. Considering only the reliable parts yields a high-precision partial alignment. Moreover, the identification of critical parts is also useful, as they often reveal musically interesting deviations between the versions to be aligned.

1 Introduction

As a result of massive digitization efforts, there is an increasing number of relevant digital documents for a single musical work comprising audio recordings, MIDI files, digitized sheet music, music videos, and various symbolic representations. In order to coordinate the multiple information sources related to a given musical work, various alignment and synchronization procedures have been proposed with the common goal to automatically link several types of music representations, [1,2,3,4,5,6,8,9,10,13,14,15,18,19,20,21,22,23,24]. In a retrieval context, this linking information allows for an easy and intuitive formulation of a query. For example, in [13] the query is created by selecting multiple bars in a score representation. As the score is linked to an audio recording the query

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 35–48, 2011.

in the score domain can be translated into a query in the audio domain, which can be used in an underlying audio retrieval system. This way, the user can make use of a semantically oriented high-level representation while the low-level representation needed only for technical reasons is hidden from the user.

In general terms, *music synchronization* denotes a procedure which, for a given position in one representation of a piece of music, determines the corresponding position within another representation. Even though recent synchronization algorithms can handle significant variations in tempo, dynamics, and instrumentation, most of them rely on the assumption that the two versions to be aligned correspond to each other with respect to their overall global temporal and polyphonic structure. In real-world scenarios, however, this assumption is often violated [11]. For example, for a popular song there often exists various structurally different album, radio, or extended versions. Live or cover versions may contain improvisations, additional solos, and other deviations from the original song [21]. Poor recording conditions, interfering screams and applause, or distorted instruments may introduce additional serious degradations in the audio recordings. On the other side, MIDI and other symbolic descriptions often convey only a simplistic view of a musical work, where, e. g., certain voices or drum patterns are missing. Furthermore, symbolic data as obtained from optical music recognition is often corrupted by recognition errors. In general, the synchronization of two strongly deviating representations of a piece of music constitutes an ill-posed problem. Here, without further model assumptions on the type of similarity, the synchronization task becomes infeasible.

In this paper, we address the problem of reliable partial music synchronization with the goal to automatically identify those passages within the given music representations that allow for a reliable alignment. Given two different representations of the same piece, the idea is to use several types of conceptually different synchronization strategies to obtain an entire family of temporal alignments. Now, consistencies over the various alignments indicate a high reliability in the encoded correspondences, whereas inconsistencies reveal problematic passages in the music representations to be aligned. Based on this automated local classification of the synchronization results, we segment the music representations into passages, which are then further classified as *reliable* and *critical*. Here, the reliable passages have a high confidence of being correctly aligned with a counterpart, whereas the critical passages are likely to contain variations and artifacts. The reliable passages can then be used as anchors for subsequent improvements and refinements of the overall synchronization result. Conversely, our automated validation is also useful in revealing the critical passages, which often contain the semantically interesting and surprising parts of a representation.

The remainder of this paper is organized as follows. In Sect. 2, we describe three conceptually different synchronization strategies. Then, in Sect. 3, we introduce our novel concept that allows for locally classifying the computed alignment results as reliable or critical. Finally, we report on our experiments in Sect. 4 and sketch some future work in Sect. 5. Further related work is discussed in the respective sections.

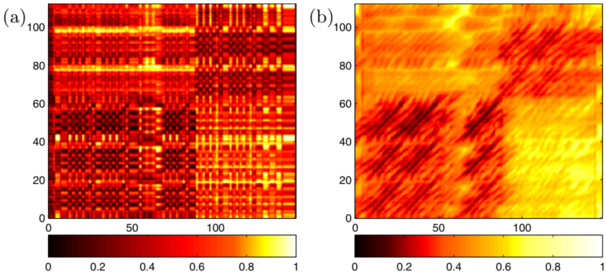


Fig. 1. Chroma based cost matrices for an audio recording (vertical axis) and a MIDI version (horizontal axis) of the song 'And I love her' by the Beatles. Times are given in seconds. **(a)** Cost Matrix **(b)** Smoothed Cost Matrix.

2 Music Synchronization Strategies

Most synchronization methods can be summarized in three simple steps. In a first step, the data streams to be aligned are converted to a suitable feature representation. Next, a local cost measure is used to compare features from the two streams. Based on this comparison, the actual alignment is then computed using an alignment strategy in a final step. In the following, we describe three conceptually different approaches for synchronizing a given MIDI-audio pair of a piece of music. Here, exemplarily using chroma features, we fix the parameters of the first two steps as described in Sect. 2.1 and focus on the third step, the alignment strategy (Sect. 2.2). As a first approach, we consider classical dynamic time warping (DTW), which allows for computing a global alignment path. We then introduce a recursive variant of the Smith-Waterman algorithm, which yields families of local path alignments. As a third approach, we use a partial matching strategy, which yields the least constrained alignment. While these three approaches share similar algorithmic roots (dynamic programming) they produce fundamentally different types of alignments. Intuitively, one may think of two extremes. On the one hand, DTW relies on strong model assumptions, but works reliably in the case that these assumptions are fulfilled. On the other hand, partial matching offers a high degree of flexibility, but may lead to alignments being locally misguided or split into many fragments. The Smith-Waterman approach can be thought of being in between these two extremes. As a complete description of the three alignment strategies would go beyond the scope of this paper, we summarize their properties while highlighting the conceptual differences among the approaches in Sect. 2.2. References to literature with exact implementation details are given as necessary.

2.1 Local Cost Measure

For comparing a MIDI file and an audio recording of the same song, we convert both representations into a common mid-level representation. Depending on the type of this representation the comparison can be based on musical properties like harmony, rhythm or timbre. Since our focus is on alignment strategies, we exemplarily fix one type of representation and revert to chroma-based music features, which have turned out to be a powerful tool for relating harmony-based music, see [8,10]. For details on how to derive chroma features from audio and MIDI files, we refer to the cited literature. In the subsequent discussion, we employ normalized 12-dimensional chroma features with a temporal resolution of 2 Hz (2 features per second).

Let $V := (v^1, v^2, \ldots, v^N)$ and $W := (w^1, w^2, \ldots, w^M)$ be two chroma feature sequences. To relate two chroma vectors we use the cosine distance defined by $c(v^n, w^m) = 1 - \langle v^n, w^m \rangle$ for normalized vectors. By comparing the features of the two sequences in a pairwise fashion, one obtains an $(N \times M)$-*cost matrix* C defined by $C(n, m) := c(v^n, w^m)$, see Fig. 1a. Each tuple (n, m) is called a *cell* of the matrix. To increase the robustness of the overall alignment procedure, we further enhance the structure of C by using a contextual similarity measure as described in [12]. The enhancement procedure can be thought of as a multiple filtering of C along various directions given by gradients in a neighborhood of the gradient $(1, 1)$. We denote the smoothed cost matrix again by C. For an example see Fig. 1b.

2.2 Alignment Methods

In the following, an alignment between the feature sequences $V := (v^1, v^2, \ldots, v^N)$ and $W := (w^1, w^2, \ldots, w^M)$ is regarded as a set $\mathcal{A} \subseteq [1 : N] \times [1 : M]$. Here, each cell $\gamma = (n, m) \in \mathcal{A}$ encodes a correspondence between the feature vectors v^n and w^m. Furthermore, by ordering its elements lexicographically $\mathcal{A}$ takes the form of a sequence, i. e., $\mathcal{A} = (\gamma_1, \ldots, \gamma_L)$ with $\gamma_\ell = (n_\ell, m_\ell)$, $\ell \in [1 : L]$. Additional constraints on the set ensure that only semantically meaningful alignments are permitted. We say that the set $\mathcal{A}$ is *monotonic* if

$$n_1 \leq n_2 \leq \ldots \leq n_L \text{ and } m_1 \leq m_2 \leq \ldots \leq m_L.$$

Similarly, we say that $\mathcal{A}$ is *strictly monotonic* if

$$n_1 < n_2 < \ldots < n_L \text{ and } m_1 < m_2 < \ldots < m_L.$$

Note that the monotonicity condition reflects the requirement of faithful timing: if an event in V precedes a second one this also should hold for the aligned events in W. A strictly monotonic set $\mathcal{A}$ will also be referred to as *match*, denoted by the symbol $\mathcal{M} = \mathcal{A}$. To ensure certain continuity conditions, we introduce step-size constraints by requiring

$$\gamma_{\ell+1} - \gamma_\ell \in \Sigma$$

for $\ell \in [1 : L - 1]$, in which Σ denotes a set of admissible step sizes. A typical choice is $\Sigma = \Sigma_1 := \{(1, 1), (1, 0), (0, 1)\}$ or $\Sigma = \Sigma_2 := \{(1, 1), (2, 1), (1, 2)\}$. Note that when using Σ_1 (Σ_2) the set $\mathcal{A}$ also becomes monotonic (strictly

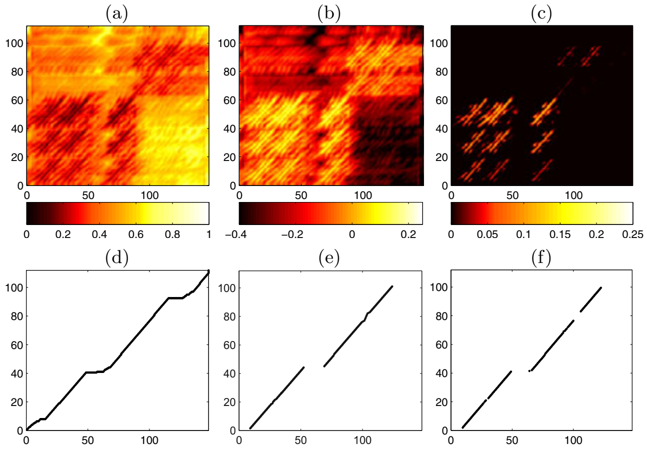


Fig. 2. Several techniques for the alignment of an audio recording (vertical axis) and a MIDI version (horizontal axis) of the song ‘And I love her’ by the Beatles. **(a)** Cost Matrix C. **(b)** Score Matrix S. **(c)** Thresholded Score Matrix $S_{\geq 0}$. **(d)** Optimal global path obtained via DTW based on matrix C. **(e)** Family of paths obtained via Smith-Waterman based on matrix S. **(f)** Optimal match obtained via partial matching based on matrix $S_{\geq 0}$.

monotonic). A set $\mathcal{A}$ that fulfills the step-size condition is also referred to as *path* denoted by the symbol $\mathcal{P} = \mathcal{A}$. As final constraint, the boundary condition

$$\gamma_1 = (1, 1) \text{ and } \ \gamma_L = (N, M),$$

ensures in combination with a step-size condition the alignment of V and W as a whole. If both the step-size as well as the boundary condition hold for a set $\mathcal{A}$, then $\mathcal{A}$ will be referred to as *global path* (or *warping path*) denoted by $\mathcal{G}$. Finally, a monotonic set $\mathcal{A}$ is referred to as *family of paths*, denoted by $\mathcal{F}$, if there exist paths $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K$ with $\mathcal{F} = \mathcal{A} = \bigcup_{k \in [1:K]} \mathcal{P}_k$.

If it is known a-priori that the two sequences to be aligned correspond to each other globally then a global path is the correct alignment model. Here, dynamic time warping (DTW) is a standard technique for aligning two given sequences and can be used to compute a global path [17,10]. In this context, the *cost* of an alignment $\mathcal{A}$ is defined as $\sum_{\ell=1}^{L} C(n_\ell, m_\ell)$. Then, after fixing a set of admissible step-sizes Σ, DTW yields an optimal global path having minimal cost among all possible global paths. In our experiments $\Sigma = \Sigma_1$ and $\Sigma = \Sigma_2$ yielded alignments of similar quality. However, here we choose $\Sigma = \Sigma_1$, since it leads to more reasonable results in cases where the assumption of global correspondence between the sequences is violated. For the subsequent discussion we use $\mathrm{A}(s, t)$ to refer to the segment in the audio recording starting at s seconds and terminating at t seconds. Similarly, $\mathrm{M}(s, t)$ refers to a MIDI segment. So listening to $\mathrm{M}(55, 65)$ of the song ‘And I love her’ (used as example in Fig. 2) reveals a short bridge in the song. However in the audio recording (taken from the Anthology release)

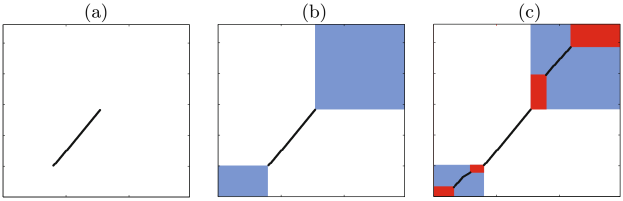


Fig. 3. First steps of our recursive Smith-Waterman variant. **(a)** Optimal path $\mathcal{P}$ derived via classical Smith-Waterman. **(b)** Submatrices defined via $\mathcal{P}$. **(c)** Result after the first recursion. Optimal paths have been derived from the submatrices and new submatrices (red) for the next recursive step are defined.

the bridge is skipped by the Beatles. Since DTW always aligns the sequences as a whole we find a semantically meaningless alignment between $\mathrm{A}(40, 42)$ and $\mathrm{M}(48, 65)$. A similar observation can be made at the beginning and the end of the optimal global path. Here, the intro and outro in the audio recording deviate strongly from those in the MIDI version.

In general, using DTW in the case that elements in one sequence do not have suitable counterparts in the other sequence is problematic. Particularly, in the presence of structural differences between the two sequences, this typically leads to misguided alignments. Hence, if it is known a-priori that the two sequences to be aligned only partially correspond to each other, a path or a family of paths allows for a more flexible alignment than a global path.

Here, the Smith-Waterman algorithm is a well known technique from biological sequence analysis [16] to align two sequences that only locally correspond to each other. Instead of using the concept of a cost matrix with the goal to find a cost-minimizing alignment, we now use the concept of a *score matrix* with the goal to find a score-maximizing alignment. To this end, we fix a threshold $\tau > 0$. Then, a score matrix S is derived from C by setting $S = \tau - C$. Fig. 2b shows a score matrix derived from the cost matrix shown in Fig. 2a using the threshold $\tau = 0.25$. The *score* of an alignment $\mathcal{A}$ is defined as $\sum_{\ell=1}^{L} S(n_\ell, m_\ell)$. Then, after fixing a set of admissible step-sizes Σ, the Smith-Waterman algorithm computes an optimal path having maximal score among all possible paths using dynamic programming similar to DTW (see [16,21]). Here, we found $\Sigma = \Sigma_2$ to deliver good alignment results.

We now introduce a recursive variant of the Smith-Waterman algorithm. In a first step, we derive an optimal path $\mathcal{P}$ as described above (see Fig. 3a). Then in a second step, we define two submatrices in the underlying score matrix S (see Fig. 3b). The first matrix is defined by the cell $(1, 1)$ and the starting cell of $\mathcal{P}$, and the second matrix by the ending cell of $\mathcal{P}$ and the cell (N, M). For these submatrices, we call the Smith-Waterman algorithm recursively to derive another optimal path for each submatrix (see Fig. 3c). These new paths define new submatrices on which Smith-Waterman is called again. This procedure is repeated until either the score of an optimal path or the size of a submatrix is below a given threshold. This results in a monotonic alignment set in form of a family of paths $\mathcal{F}$. Fig. 2e shows a family of two paths derived from the score matrix in Fig. 2b using our recursive Smith-Waterman variant. Here, the missing bridge in the audio as well as the different intros and outros in the audio and MIDI version are detected and, in this example, the recursive Smith-Waterman approach avoids a misalignment as in the DTW case (Fig. 2d).

This example illustrates another interesting property of the Smith-Waterman algorithm. Listening to A$(75, 83)$ and M$(99, 107)$ reveals a solo improvisation which is different in the audio and MIDI version, so they should not be aligned. Also, the corresponding area in the score matrix shows negative values. However, the Smith-Waterman algorithm aligns these two segments as part of the second path. The reason is that Smith-Waterman always tries to find the path with maximum score, and in this example the score for a longer path containing a few negative entries was higher than for a shorter path without negative entries. This property of the Smith-Waterman algorithm can be configured using *gap-penalty parameters*, see [21]. Essentially, these are used to define an additional weight for negative score entries. If the weight is high, then negative entries are emphasized and paths tend to be shorter and contain fewer entries with negative score. If the weight is low, paths tend to be longer and may contain longer sequences with negative score. We chose to use gap-penalty parameters being equivalent to a subtraction of 0.5 from all negative entries in the score matrix S. This is an empirically found value which worked best in our experiments.

However, even using gap-penalty parameters there is no control over the absolute length of sequences with negative score in an optimal path. If there is enough score to gain before and after a sequence with negative score, then this sequence will be bridged using Smith-Waterman and can become arbitrarily long in the resulting optimal path. So for the example depicted in Fig. 2e one could find a set of parameters to circumvent this misalignment, but in general these parameters would have to be set for each song individually. Here, a method referred to as *partial matching* allows for an even more flexible local alignment than Smith-Waterman (see [1,10,16]). The goal is to find a score-maximizing alignment, similar to the Smith-Waterman approach. But instead of using the Smith-Waterman score matrix S a thresholded version $S_{\geq 0}$ is used where every negative entry in S is replaced by zero (see Fig. 2c). The idea of this thresholding is to disallow the alignment of a pair of feature vectors if there is no score to gain from this alignment. Therefore, negative score can be ignored completely. In a sense, this concept is similar to finding the longest common subsequence (LCS) in string matching tasks. Based on this idea partial matching computes a score-maximizing optimal match. Again, this can be achieved efficiently using dynamic programming. See Fig. 2f for an example of an optimal match computed via partial matching based on the matrix shown in Fig. 2c. Here, the misalignment of the solo segments A$(75, 83)$ and M$(99, 107)$ as found in the Smith-Waterman case is not present. So partial matching, not enforcing any step-size or continuity conditions on the alignment, yields a more flexible alignment than the Smith-Waterman approach.

However, for other pieces, the lack of step-size constraints might lead to highly fragmented or even misguided alignments. As an example, we consider the song

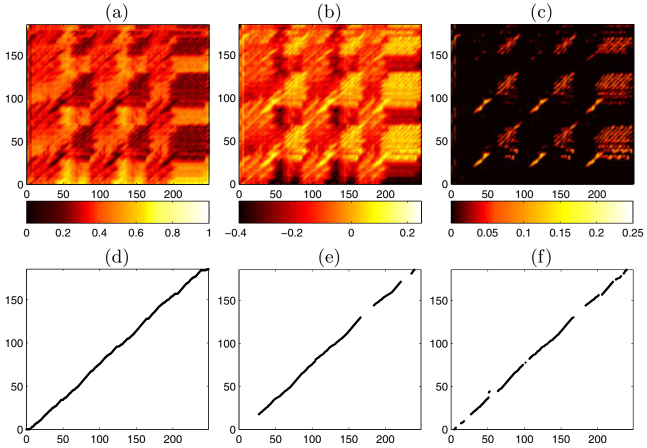


Fig. 4. Several techniques for the alignment of an audio recording (vertical axis) and a MIDI version (horizontal axis) of the Beatles song 'Lucy in the sky with diamonds', cf. Fig. 2

'Lucy in the sky with diamonds' by the Beatles (see Fig. 4). Here, the optimal match computed via partial matching is highly fragmented (Fig. 4f). This is caused by local deviations of the audio recording (taken from the anthology release) from the MIDI version. For example, A(133, 147) and M(165, 183) are semantically corresponding sections and should be aligned, but slightly differ in their arrangement. Here, the MIDI version features a very prominent bass line while in the audio recording the bass is only in the background. This leads to chroma representations for the audio and MIDI segment that differ strongly from each other thus explaining the low score in the corresponding area of the score matrix $S_{\geq 0}$, see Fig. 4c. Similar observations can be made comparing A(40, 47) and M(52, 62) as well as A(75, 82) and M(100, 110). However, the latter two pairs are correctly aligned by the recursive Smith-Waterman variant, see Fig. 4e. Here, a path is allowed to contain a few cells with negative score which helps to overcome local deviations in the feature representation. Nonetheless, using Smith-Waterman the segments A(133, 147) and M(165, 183) are still not aligned to each other. Here, the classical DTW approach, being forced to align the sequences as a whole, yields the best alignment result.

As illustrated by the examples shown in Figs. 2 and 4 each synchronization strategy sometimes yields reasonable and sometimes misguided and unsatisfying results. Therefore, without any definite a-priori knowledge about the input data none of the presented alignment methods can guarantee in general a reliable and musically meaningful alignment.

3 Proposed Method

In general, when the two sequences to be aligned correspond to each other in terms of their global temporal progression, the DTW procedure yields a robust

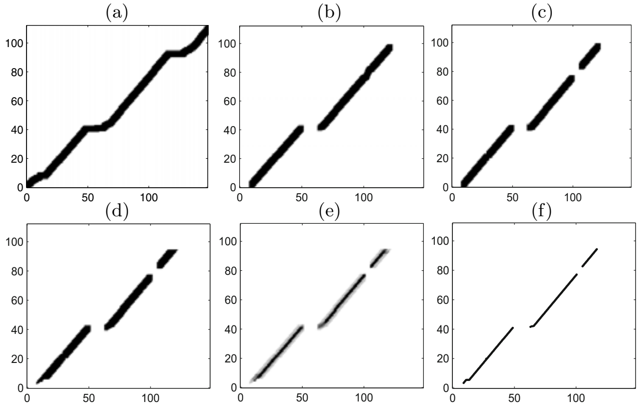


Fig. 5. Steps in our proposed method continuing the example shown in Fig. 2. **(a)-(c)** Augmented binary matrices for the optimal global path (DTW), family of paths (Smith-Waterman) and the optimal match (partial matching) using a tolerance neighborhood of 2 seconds. **(d)** Intersection matrix derived from (a)-(c). **(e)** Weighted intersection matrix. **(f)** Consistency alignment $\mathcal{C}$.

alignment result. On the other hand, if structural differences are present, the more flexible Smith-Waterman approach or the even more flexible partial matching procedure may yield more reasonable alignments than DTW. Now, if several strategies with different design goals yield similar alignment results, then there is a high probability of having semantically meaningful correspondences. Based on this simple idea, we present an automatic method towards finding passages in the MIDI and audio representations that can be synchronized in a reliable way. In contrast, this method also can be applied to identify critical passages, where the alignments disagree.

Given a MIDI-audio pair for a song, we start by computing an optimal global path using DTW, a family of paths using recursive Smith-Waterman, as well as an optimal match using partial matching. Next, we convert each alignment into a binary matrix having the same size as the cost matrix C. Here, a cell in the matrix is set to one if it is contained in the corresponding alignment and zero otherwise (see Figs. 2d-f). The next step is essentially a soft intersection of the three alignments. To this end, we augment the binary matrices by additionally setting every cell in the binary matrices to one if they are in a neighborhood of an alignment cell (see Figs. 5a-c). For Fig. 5, we used a large neighborhood of two seconds for illustrative purposes, while for the experiments in Sect. 4 we used a neighborhood of only one second. After that we derive an intersection matrix by setting each matrix cell to one that is one in all three augmented binary matrices (see Fig. 5d). The intersection matrix can be thought of as a rough indicator for areas in the cost matrix where the three alignment strategies agree with each other. However, this matrix does not encode an alignment that is constrained by any of the conditions described in Sect. 2.2. Therefore, to derive a final alignment result from this matrix, we first weight the remaining cells in the intersection matrix according to how often they are contained in one of the original three alignments (Fig. 5e). Then, interpreting the weighted matrix as a score matrix, we use partial matching to compute an optimal match $\mathcal{C}$ referred to as *consistency alignment* (Fig. 5f).

In the following, we call a segment in the audio recording (the MIDI version) *reliable* if it is aligned via $\mathcal{C}$ to a segment in the MIDI version (in the audio recording). Similarly, we call a segment *critical* if it is not aligned. Here, A(3, 39), A(39, 76) and A(83, 95) as well as M(8, 45), M(63, 99) and M(106, 117) are examples of reliable segments in the audio recording and in the MIDI version, respectively. However, the automatic detection of critical sections can also be very useful as they often contain musically interesting deviations between two versions. For example, consider the critical segment M(45, 63). This segment contains the bridge found in the MIDI that was omitted in the audio recording as discussed in Sect. 2.2. Here, our method automatically revealed the inconsistencies between the MIDI version and the audio recording. Similarly, the differences between the audio and the MIDI version in the intro, outro and solo segments have also been detected. Here, not relying on a single alignment strategy leads to a more robust detection of critical segments than using just a single approach. The reasons why a segment is classified as critical can be manifold and might be an interesting subject for a subsequent musical analysis. In this context, our approach can be thought of as a supporting tool for such an analysis.

4 Experiments

In this section, we report on systematically conducted experiments to illustrate the potential of our method. To this end, we used twelve representative pieces from the classical, popular, and jazz collection of the RWC music database [7]. For each piece, RWC supplies high-quality MIDI-audio pairs that globally correspond to each other. Hence, using classical DTW allows us to synchronize each MIDI-audio pair to obtain an accurate alignment that can be used as ground-truth. The synchronization results were manually checked for errors. For the experiment, we strongly distorted and modified the MIDI versions as follows. Firstly, we temporally distorted each MIDI file by locally speeding up or slowing down the MIDI up to a random amount between ±50%. In particular, we continuously changed the tempo within segments of 20 seconds of length with abrupt changes at segment boundaries to simulate ritardandi, accelerandi, fermata, and so on. Secondly, we structurally modified each MIDI file by replacing several MIDI segments (each having a length of 30 to 40 seconds) by sets of short 2 second snippets taken from random positions within the same MIDI file. In doing so, the length of each segment remained the same. These modified segments do not have any corresponding segments in the audio anymore. However, taken from the same piece, the snippets are likely to be harmonically related to the replaced content. Here, the idea is to simulate a kind of improvisation that fits into the harmonic context of the piece but is regarded as different between the audio and the MIDI version (similar to the differences found in A(75, 83) and M(99, 107), as discussed in Sect. 2.2). Finally, recall that we computed a ground-truth alignment via DTW between the original MIDI and the audio. Keeping

Table 1. Precision (P), Recall (R) and F-measure (F) for the alignment strategies DTW, recursive Smith-Waterman (rSW), Partial Matching (PM) and the consistency alignment (Con). **Left:** PR-values for modified MIDI-audio pairs. **Right:** PR-values for strongly modified MIDI-audio pairs.

	P	R	F
DTW	**0.63**	0.96	0.76
rSW	**0.85**	0.85	0.85
PM	**0.80**	0.92	0.85
Con	**0.93**	0.85	0.88

	P	R	F
DTW	**0.51**	0.95	0.66
rSW	**0.83**	0.84	0.83
PM	**0.74**	0.92	0.82
Con	**0.95**	0.84	0.89

track of the MIDI modifications, we derive a ground-truth alignment between the modified MIDI and the audio, in the following referred to as $\mathcal{A}^*$.

For each modified MIDI-audio pair, we compute the three alignments obtained by the three strategies described in Sect. 2.2 as well as the consistency alignment as described in Sect. 3. Let $\mathcal{A}$ be one of the computed alignments. To compare $\mathcal{A}$ with the ground-truth alignment $\mathcal{A}^*$, we introduce a quality measure that is based on precision and recall values, while allowing some deviation tolerance controlled by a given tolerance parameter $\varepsilon > 0$. The precision of $\mathcal{A}$ with respect to $\mathcal{A}^*$ is defined by

$$\mathrm{P}(\mathcal{A}) = \frac{|\{\gamma \in \mathcal{A} | \exists \gamma^* \in \mathcal{A}^* : ||\gamma - \gamma^*||_2 \leq \varepsilon\}|}{|\mathcal{A}|}$$

and the recall of $\mathcal{A}$ with respect to $\mathcal{A}^*$ is defined by

$$\mathrm{R}(\mathcal{A}) = \frac{|\{\gamma^* \in \mathcal{A}^* | \exists \gamma \in \mathcal{A} : ||\gamma - \gamma^*||_2 \leq \varepsilon\}|}{|\mathcal{A}^*|}.$$

Here, $||\gamma - \gamma^*||_2$ denotes the Euclidean norm between the elements $\gamma, \gamma^* \in [1 : N] \times [1 : M]$, see Sect. 2.2. Finally, the F-measure is defined by

$$\mathrm{F}(\mathcal{A}) := \frac{2\mathrm{P}(\mathcal{A})\mathrm{R}(\mathcal{A})}{\mathrm{P}(\mathcal{A}) + \mathrm{R}(\mathcal{A})}.$$

In our experiments, we used a threshold parameter ε corresponding to one second. The left part of Table 1 shows the PR-values averaged over all pieces for the four different alignment results. For example, when using DTW, the precision amounts to only $\mathrm{P} = 0.63$. The reason for this low value is that all time positions of the MIDI are to be aligned in the global DTW strategy, even if there is no semantically meaningful correspondence in the audio. When using Smith-Waterman or partial matching, the precision values become better. Note that the three alignments based on the three different strategies typically produce different (often random-like) correspondences in regions where the MIDI and audio differ. As a result, these correspondences are discarded in the consistency alignment yielding a high precision of $\mathrm{P} = 0.93$. This is exactly what we wanted to achieve by our proposed method, where we only want to keep the reliable information, possibly at the cost of a lower recall.

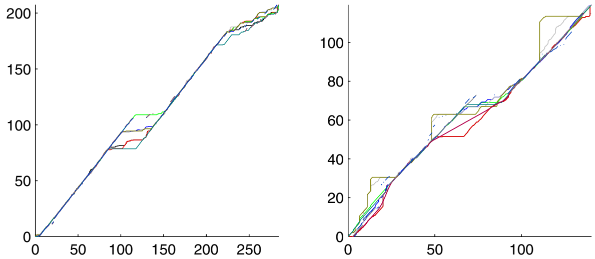


Fig. 6. Various alignments for two Beatles songs using conceptually different synchronization approaches **Left:** 'While My Guitar Gently Weeps' **Right:** 'Norwegian Wood'

In a second experiment, we modified the MIDI version even further by not only distorting and replacing randomly chosen MIDI segments as described above, but by inserting additional MIDI snippet segments. These additional structural modifications make the synchronization task even harder. The corresponding PR-values averaged over all pieces are shown in the right part of Table 1. As the task is harder now, the quality measures for all three alignment strategies drop except for our consistency alignment, where the precision (P = 0.95) essentially remains the same as in the first experiment.

5 Conclusions and Future Work

In this paper, we introduced a novel method for locally classifying alignments as reliable or critical. Here, our idea was to look for consistencies and inconsistencies across various alignments obtained from conceptually different synchronization strategies. Such a classification constitutes an essential step not only for improving current synchronization approaches but also for detecting artifacts and structural differences in the underlying music material. In the latter sense, our approach may be regarded as a supporting tool for musical analysis.

To cope with the richness and variety of music, we plan to incorporate many more competing strategies by not only using different alignment strategies but also by considering different feature representations, various feature resolutions, several local cost (similarity) measures, and different enhancement strategies for cost (score) matrices. Here, additional alignment strategies can be based on approaches algorithmically different from the ones presented here, like HMM-based methods as known from online score following and computer accompaniment [2,18,20], but also on approaches describes in Sect. 2 in combination with varying values for important parameters like τ or Σ. Fig. 6 shows first illustrating results for two Beatles songs, where we computed a large number of alignments using many different synchronization approaches. Despite of significant differences in structure, timbre, and polyphony, the consistencies of the various alignments reveal the reliable passages that can then serve as anchor for subsequent improvements and refinements of the overall synchronization result.

Acknowledgements. The first author is funded by the German Research Foundation (DFG CL 64/6-1). The second author is with the Cluster of Excellence on *Multimodal Computing and Interaction* at Saarland University. The third is partially supported by the National Science Foundation (0534370).

References

1. Arifi, V., Clausen, M., Kurth, F., Müller, M.: Synchronization of music data in score-, MIDI- and PCM-format. Computing in Musicology 13 (2004)
2. Cont, A.: Real time audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical hmms. In: Proc. IEEE ICASSP, Toulouse, France (2006)
3. Dannenberg, R., Hu, N.: Polyphonic audio matching for score following and intelligent audio editors. In: Proc. ICMC, San Francisco, USA, pp. 27–34 (2003)
4. Dannenberg, R., Raphael, C.: Music score alignment and computer accompaniment. Special Issue, Commun. ACM 49(8), 39–43 (2006)
5. Dixon, S., Widmer, G.: MATCH: A music alignment tool chest. In: Proc. ISMIR, London, GB (2005)
6. Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T., Okuno, H.G.: Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In: ISM, pp. 257–264 (2006)
7. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: Rwc music database: Popular, classical and jazz music databases. In: ISMIR (2002)
8. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: Proc. IEEE WASPAA, New Paltz, NY (October 2003)
9. Kurth, F., Müller, M., Fremerey, C., Chang, Y., Clausen, M.: Automated synchronization of scanned sheet music with audio recordings. In: Proc. ISMIR, Vienna, AT (2007)
10. Müller, M.: Information Retrieval for Music and Motion. Springer, Heidelberg (2007)
11. Müller, M., Appelt, D.: Path-constrained partial music synchronization. In: Proc. IEEE ICASSP, Las Vegas, USA (2008)
12. Müller, M., Kurth, F.: Enhancing similarity matrices for music audio analysis. In: Proc. IEEE ICASSP, Toulouse, France (2006)
13. Müller, M., Kurth, F., Damm, D., Fremerey, C., Clausen, M.: Lyrics-based audio retrieval and multimodal navigation in music collections. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 112–123. Springer, Heidelberg (2007)
14. Müller, M., Kurth, F., Röder, T.: Towards an efficient algorithm for automatic score-to-audio synchronization. In: Proc. ISMIR, Barcelona, Spain (2004)
15. Müller, M., Mattes, H., Kurth, F.: An efficient multiscale approach to audio synchronization. In: Proc. ISMIR, Victoria, Canada, pp. 192–197 (2006)
16. Pevzner, P.A.: Computational Molecular Biology: An Algorithmic Approach. MIT Press, Cambridge (2000)
17. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series (1993)
18. Raphael, C.: Automatic segmentation of acoustic musical signals using hidden markov models. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 360–370 (1998)

19. Raphael, C.: A hybrid graphical model for aligning polyphonic audio with musical scores. In: Proc. ISMIR, Barcelona, Spain (2004)
20. Schwarz, D., Orio, N., Schnell, N.: Robust polyphonic midi score following with hidden markov models. In: International Computer Music Conference, Miami (2004)
21. Serrà, J., Gómez, E., Herrera, P., Serra, X.: Chroma binary similarity and local alignment applied to cover song identification. IEEE Transactions on Audio, Speech and Language Processing 16, 1138–1151 (2008)
22. Soulez, F., Rodet, X., Schwarz, D.: Improving polyphonic and poly-instrumental music to score alignment. In: Proc. ISMIR, Baltimore, USA (2003)
23. Turetsky, R.J., Ellis, D.P.W.: Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses. In: Proc. ISMIR, Baltimore, USA, pp. 135–141 (2003)
24. Wang, Y., Kan, M.-Y., Nwe, T.L., Shenoy, A., Yin, J.: LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics. In: MULTIMEDIA 2004: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 212–219. ACM Press, New York (2004)

Note Recognition from Monophonic Audio: A Clustering Approach

Rainer Typke

Austrian Research Institute for Artificial Intelligence (OFAI), Vienna

Abstract. We describe a new method for recognizing notes from monophonic audio, such as sung or whistled queries. Our method achieves results similar to known methods, but without any probabilistic models that would need to be trained. Instead, we define a distance function for audio frames that captures three criteria of closeness which usually coincide with frames belonging to the same note: small pitch difference, small loudness fluctuations between the frames, and the absence of non-pitched frames between the compared frames. We use this distance function for clustering frames such that the total intra-cluster costs are minimized. Criteria for clustering termination include the uniformity of note costs. This new method is fast, does not rely on any particular fundamental frequency estimation method being used, and it is largely independent of the input mode (singing, whistling, playing an instrument). It is already being used successfully for the "query by humming/whistling/playing" search feature on the publicly available collaborative melody directory Musipedia.org.

1 Introduction

1.1 Goal: Recognizing Notes from Monophonic Audio

This paper describes a new method for finding note boundaries in monophonic audio recordings. Unlike existing note onset detectors, our method bases its decisions on whole groups of notes instead of only looking very locally for changes in the spectrum. Also, our method does not use or need any training, which makes it very generally applicable to singing, whistling, or a range of musical instruments.

1.2 Related Work

Models: Matti Ryynänen and Anssi Klapuri [12,14,13] participated very successfully in various MIREX tasks that involved finding boundaries between notes (e. g., "Query by Singing/Humming" 2008). Their method uses Hidden Markov Models (HMM) for modeling notes. The HMMs need to be trained using audio recordings which are similar to those one wants to process later. Toh, Zhang, and Wang [5] use Gaussian Mixture Models (GMMs) to distinguish between onset

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 49–58, 2011.

frames and non-onset frames. They build onset detectors using those GMMs and decide on note onsets by picking peaks.

Local spectral changes: Paul Brossier et al. [4,3] published various note onset detectors and made them available as part of a program called aubio[1]. Note onsets are detected by using a "detection function derived from one or a few consecutive spectral frames of a phase vocoder. The detection function increases at the beginning of the note attacks. Peak- picking is required to select only the relevant onsets."

Harte, Sandler, and Gasser [9] describe a detection function for harmonic changes in polyphonic music. They map 12-bin chroma vectors to the interior space of a 6-dimensional polytope such that close harmonic distances appear as small Euclidean distances. Harmonic changes are detected by calculating a "Harmonic Change Detection Function" that determines, for each frame, the Euclidian distance between the tonal centroids of the previos and the next frame.

Rule-based system: Goffredo Haus et al. [10] described a complex, rule-based method for detecting notes in audio recordings. They employ rules for detecting noise; if a recording is deemed noise-free enough, other distortions such as vibrato are dealt with, and the recording is analyzed on three levels ("frame", "block", and "note"). Unfortunately, this method was never compared to others in MIREX.

Discontinuity detection: Collins [7] uses a pitch detector, removes glitches, suppresses vibrato, and then detects note onsets based on discontinuity of pitch. De Mulder et al. [8] base segmentation on loudness instead of pitch. Pauws [11] also focuses on onset detection, but looks at multiple curves at once (energy, surf, and high-frequency content).

Clustering: Camacho [6] uses a clustering approach to distinguish between pitched and unpitched segments. He determines the pitch strength trace and determines clusters by locally maximizing distances between centroids.

1.3 Motivation

With our new method, we aim for a performance similar to that of Ryynänen's HMM models, but with more general applicability and without the need for training any models. We are willing to sacrifice some real-time properties that methods such as Brossier's have. In a query-by-humming setting such as on Musipedia.org, it is acceptable if notes are detected during a time frame of a few fractions of a second immediately after the user has finished singing, whistling, or playing a query. Therefore, the entire query can be taken into consideration for analysis. Due to the wide variety of input modes one has to expect on a public website – people will sing in many different ways, whistle, sometimes even play

[1] http://aubio.org

instruments –, it is very useful if one can avoid models whose training would mean that note features would be taken into account which differ with these different input modes.

2 Starting Point: Pitch and Intensity Curves

Instead of raw audio, we use a combination of a pitch curve and an intensity (loudness) curve as input for our algorithm. Not only does this relieve us from reinventing the wheel, as there is already a large body of literature describing how to estimate fundamental frequencies, but it also opens up our method to possible later improvements in fundamental frequency estimation. Any better method for extracting a pitch curve from audio recordings can be combined with our note detector and should immediately lead to better results.

We used a method from the early nineties for the purpose of pitch extraction: Boersma's autocorrelation algorithm [2] in the form that is implemented in Version 5.0.29 of Praat [1]. Praat was also used to extract the desired intensity curve.

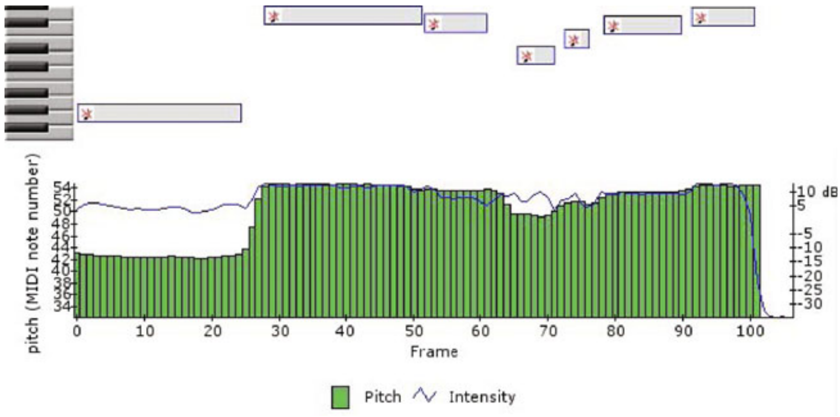


Fig. 1. Bottom: pitch curve (shown with bars) and intensity curve (line). Top: the notes recognized from these two curves. This query was sung by a visitor of Musipedia.org; you may listen to it – and see how it is analyzed – at http://www.musipedia.org/ed_qu.php?hash=9ad1849e978b75db92eaf7592f8d8174 or http://shrtlnk.net/rainbow (click on "Debug Information" for listening and for viewing intermediate results).

Since we ignore everything other than pitch and intensities, our method works equally well for a large range of input modes – singing, whistling, and musical instruments can all be treated in the same way after converting the signal into these two curves. An obvious disadvantage of this approach, however, is that we can no longer recognize borders between sung notes in cases where the lyrics provide the only clue. That is, if consecutive notes are tied together, sung with the same pitch, and sung without a dip in loudness to divide them.

3 Clustering

In this section, we will describe how we extract a list of notes with their onset times, durations, and pitches from a pitch curve and an intensity curve. Both curves are given as lists of numbers with one number per audio frame. We used a frame width of 0.15 s.

3.1 Pitchless Frames Mark Note Boundaries

At least if a pause is long enough (and with our frame size, a pause of one frame is already long enough), human listeners will perceive groups of audio frames without pitch as not belonging to any note. Therefore, we reduce the note recognition problem to that of segmenting groups of consecutive frames which are all pitched.

3.2 Clustering According to Pitch and Intensity

Instead of directly detecting note onsets (and thus concentrating on the beginnings of notes), we use a more holistic approach and cluster frames which fit well together.

Distance function for frames. In order to attach a cost to a group of frames, we first define a distance function. The function calculates the cost of two given frames being part of the same note. This cost should be high if the pitches of the two frames differ strongly, and it should also be high if there is an intensity reduction, followed by an intensity increase, between them. Finally, the existence of non-pitched frames between them should increase the cost. It is clear that non-pitched frames should always lead to a note split. Therefore, we weight the cost component for non-pitched notes strongly enough for always outweighing the other components; this is, for instance, achieved with factor 12 (see below), but we could just as well have chosen factor 99. This leaves us to deal with two more components, intensity and pitch. Since we want a method that is generally applicable, we do not want to determine good weighting factors for individual singers, thus we need to normalize the intensity component before applying weights. Pitch does not need to be normalized because all singers should already agree on how big, for instance, a minor second is. We normalize the intensity component based on the value range found in the whole audio recording. This lets us treat singers or whistlers with personal preferences for larger or smaller absolute intensity differences equally. After normalization, we adjusted the weights for pitch and intensity such that pitch usually dominates intensity. Especially if intervals above the range of a typical vibrato are involved, we know that there must be a new note event; for intensity changes, it is less clear whether we are confronted with a new note or just with tremolo or an unintended recording artifact.

Based on these requirements and principles, we define the distance function **fdist** as follows. F is the set of all frames, f_1 and f_2 are two frames to be compared. The function **intdist** finds the longest and deepest intensity dip between two given frames; by "intensity dip", we mean a sequence of frames with strictly decreasing intensity, followed by a sequence of frames with strictly increasing intensity. For this largest intensity dip, the function returns the depth (largest observed intensity difference) divided by its width (time distance between the beginning and end of the dip). The function **fpitchless** simply returns the number of frames with no pitch between the given frames.

$$\begin{aligned}\mathbf{fdist}(f_1, f_2) = \mathbf{intdist}(f_1, f_2) / \max_{a,b \in F} \mathbf{intdist}(a, b) + \\ |\,\mathbf{fpitch}(f_1) - \mathbf{fpitch}(f_2)| / 12 + \\ 12 * \mathbf{fpitchless}(f_1, f_2)\end{aligned} \qquad (1)$$

Given two frames, function **fdist** returns the sum of the normalized size of any intensity dip between them, plus the pitch difference (given in octaves), plus a large penalty for any frames without pitch between them. Note that the first component, $\mathbf{intdist}(f_1, f_2) / \max_{a,b \in F} \mathbf{intdist}(a, b)$, is zero for any group of frames that does not contain a dip. Therefore, for arbitrarily large groups of consecutive frames with no intensity dip and no pitchless frames, the only non-zero component is the pitch component.

We then define the **fcluster_cost** of a frame set C simply as:

$$\mathbf{fcluster\_cost}(C) = \sum_{f_1, f_2 \in C} \mathbf{fdist}(f_1, f_2) \qquad (2)$$

This definition of the cost of a cluster has the effect that our algorithm does not necessarily attach higher costs to larger groups of frames – as long as the pitch is stable, even very large groups of frames can have low costs, as long as they do not contain pitchless frames or a dip in intensity.

This definition also leads to robustness against vibrato and, to some degree, also tremolo. Any small variation in pitch within a note will only lead to small increases of the pitch component of the cost; as long as the vibrato pitch range stays well below a minor second, which it usually should, a pitch difference of a minor second or more will always have a much higher influence on the pitch cost component than vibrato. Increasing robustness against tremolo is probably best done by somewhat smoothing the intensity curve; but even if this smoothing fails to completely remove the tremolo, a small tremolo-induced intensity dip still does not necessarily make it impossible to detect the intensity change that is typical for a note onset since we always look at the largest intensity dip we can find between two frames.

Clustering algorithm. For clustering frames, we use an iterative algorithm for finding new note borders that maximally reduce the sum of all cluster costs.

```
Input: A list of frames with a pitch value and an intensity value for every frame.
Output: A list of notes with onset times, durations, and pitches.
  Initial clustering: Create one cluster from every group of consecutive frames with a defined pitch.
  repeat
    Calculate the cost fcluster_cost for each existing cluster.
    Split the most expensive cluster in two such that the total cost of the two new clusters is minimal.
  until the cluster costs are uniform or the most expensive cluster is cheap.
  Report each cluster as a note: Onset time of the first frame, duration calculated from the number of frames, pitch as the median pitch of all frames in the cluster.
```

A good border for splitting a cluster is found efficiently as follows: a prefix array and a postfix array for the cluster are constructed. Each array element in the prefix array contains the sum of frame pair costs from the beginning of the cluster until the array element in question. The postfix array is constructed in a similar way, but with the sum of frame pair costs from the array element until the end of the cluster. The cluster is then split at the point where the sum of corresponding elements of the prefix and postfix arrays is minimal – this will typically find the optimum border with linear effort.

Clustering terminates when cluster costs are either "uniform" or "cheap". By this, we mean that either the standard deviation of cluster costs falls below 0.3 or the most recent cost reduction falls below a quarter of the median note cost. These two thresholds are both not absolute, but adapt automatically to variations such as the presence or absence of tremolo or vibrato. Tremolo or vibrato drives up the costs of all notes, which means that the final result should contain more expensive notes than a query that is sung without any tremolo or vibrato. In both cases, we can simply stop splitting clusters once the most recent gain is small in comparison to the average note cost, or the cluster costs are uniform.

3.3 Improvement: Glissando Removal

A few experiments with sung queries from the real world have shown that note recognition can be improved by running the clustering algorithm until the largest cluster is quite small (a few frames) and then setting the pitch to zero for improbable miniature notes, that is, notes which are just a few frames in length but contain a large pitch variation. Such clusters are most likely not meant to be notes, but are just glissandi. After setting the pitch to zero for these glissando frames, the algorithm is restarted from scratch.

4 Evaluation

A subjective evaluation is easy – one can simply go to http://www.musipedia.org/query_by_humming.html, whistle, play, and sing tunes, and immediately see the quality of the transcription. When doing that, one should, of course, rule out any unnecessary problem such as oversampling or an unsuitable volume setting by replaying one's query after it has been recorded, especially if the transcription contains errors.

However, it is also desirable to evaluate the note recognition algorithm in a more objective way, to compare it to the state of the art, and to find its strengths and weaknesses. We chose Ryynänen's method, which is based on note event modeling, for our comparison because it consistently performed well at the MIREX competitions. Also, Ryynänen made transcription examples available[2]. We analyzed 14 of Ryynänen's 15 examples using our new method and counted the errors for both methods. We counted false positives (notes reported by the method which a human cannot hear in the sung input) and false negatives (sung notes that are not recognized by the method). We skipped one of Ryynänen's examples (the second instance of "Brother, can you spare a dime", ID 37) because even a human cannot reliably detect which notes the singer intended, which means that false positives and false negatives cannot be meaningfully counted for this one example. Since we throw away lyrics before detecting notes, one should expect that our new method will run into difficulties if repeated notes are marked exclusively by lyrics and neither by intensity dips nor rests between notes; therefore, we counted the false negatives for repeated notes separately.

Figure 2 shows the percentages of false positives and false negatives as box-and-whisker plots. 14 sung queries with a total of 518 notes were considered.

Table 1. FP = false positives, FN = false negatives, FN-R = false negatives without counting note repetitions

ID	Total	—Ryynänen—		—Our method—		
	Notes	FP	FN	FP	FN	FN-R
26	30	0	0	0	0	0
33	41	0	0	1	2	2
34	59	1	1	0	16	0
54	45	0	4	0	3	3
64	31	0	2	0	8	0
67	44	0	8	0	11	0
72	28	3	0	0	6	0
77	29	0	3	1	2	0
85	44	0	0	0	7	0
96	32	0	0	0	1	1
99	23	1	0	1	1	1
111	40	1	2	0	6	6
115	28	3	2	2	3	3
137	44	3	3	0	2	1

[2] http://www.cs.tut.fi/sgn/arg/matti/demos/monomel.html

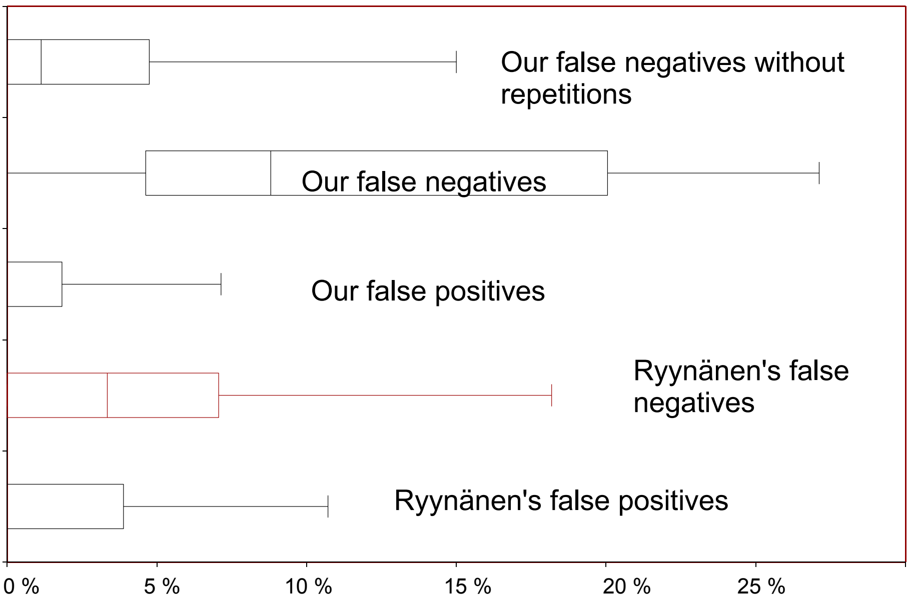


Fig. 2. False negatives (with/without counting note repetitions) and false positives for our method (top) and Ryynänen's note event modeling method (bottom). The vertical lines in the box-and-whisker plots mark the arithmetic mean, the boxes contain the middle 50 % of data points, and the whiskers indicate the non-outlier maximum and minimum values. Apart from the expected poor recognition of note repetitions which are only perceivable if one takes lyrics into consideration, our method shows a performance comparable to the state of the art.

The top three box-and-whisker plots show the results for our method, while the bottom two show Ryynänen's false negatives and false positives. The raw data are shown in Table 1.

One can see that apart from our expected poor performance for note repetitions that can solely be detected based on lyrics (which we purposely ignore), the methods perform roughly equally well – we even produced slightly fewer false positives. This is surprising because Ryynänen's detector was trained using the same data which were used for this evaluation, while our method was not optimized for these data. In particular, the variable parameters of our method, namely frame size, weights for the distance function, and the termination criteria of the cluster algorithm, were set based on real-world Musipedia queries, not on Ryynänen's data which were used for evaluation. Our method performs equally well for notes for which Ryynänen's detector would need to be trained separately, such as whistled notes or notes played on string instruments (we tried a viola, for example). Also, our method is noticeably simpler.

5 Conclusions

We presented a new, simple method for recognizing boundaries between notes in whistled or sung monophonic queries. Note onsets are not detected locally; instead, optimum borders between clusters of frames are found by minimizing intra-cluster distances between frames. Training is not necessary, and the method supports a wide range of input methods such as singing, whistling, and musical instruments.

Our evaluation shows that compared to Ryynänen's note event model method, we lose some accuracy for recognizing repeated notes because our method does not yet take lyrics into consideration for note separation. However, we gain the ability to process a much wider range of instruments (not only singing but also whistling and various instruments), we avoid the need for training models on data, and we gain a great deal of simplicity. For input methods other than singing, the current lack of lyrics processing obviously is not a problem.

It would be desirable to do an objective performance comparison also for queries that were played on various musical instruments or whistled. Unfortunately, we do not have access to a collection of such queries and performance data for other methods, but our subjective evaluation on www.musipedia.org/query_by_humming.html indicates that our performance for whistling or string instruments exceeds the performance for singing.

The performance of our method could still be increased in various ways:

- One could improve the recognition of repeated notes by introducing a new spectrum-based cost component to our frame distance function: the more the spectrum changes between frames, the higher the distance should be. This could introduce the possibility of splitting notes based on lyrics, or based on typical note onset noises for musical instruments, without sacrificing too much of the general applicability.
- There are a few parameters for which a systematic optimization could help; these parameters include:
    - the weights of the pitch and intensity components for the frame distance function,
    - the termination criteria ("uniformity" and "cheapness") for clustering,
    - the frame size.
- Also, better pitch and intensity detection would automatically improve the performance.

Paul Brossier's methods for onset detection (and similar methods) have an obvious advantage: they do not need to look ahead very far and are thus suitable for real-time onset detection, while our clustering method as described in this paper expects to see the whole recording. However, our method could easily be modified such that it only needs to see enough frames for the next one or two notes before making a decision. It would simlpy need to be run on a sliding window, and notes would be detected with a delay of about one or two notes.

Acknowledgements

R. Typke gratefully acknowledges support by the Austrian Research Fund (FWF), Project Number M1027-N15. Many thanks also to the reviewers for their valuable comments.

References

1. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (version 5.0.29) [computer program] (2008), http://www.praat.org/ (retrieved August 4, 2008)
2. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings 17, 97–110 (1993)
3. Brossier, P., Bello, J.P., Plumbley, M.D.: Fast labelling of notes in music signals. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona (2004)
4. Brossier, P., Bello, J.P., Plumbley, M.D.: Real-time temporal segmentation of note objects in music signals. In: Proceedings of the International Computer Music Conference (ICMC 2004), Miami, Florida, USA (2004)
5. Wang, Y., Toh, C.C., Zhang, B.J.: Multiple-feature fusion based onset detection for solo singing voice. In: ISMIR (2008)
6. Camacho, A.: Detection of pitched/unpitched sound using pitch strength clustering. In: ISMIR, pp. 533–537 (2008)
7. Collins, N.: Using a pitch detector for onset detection. In: ISMIR (2005)
8. de Mulder, T., Martens, J.-P., Lesaffre, M., Leman, M.M., de Baets, B., de Meyer, H.: An auditory model based transcriber of vocal queries. In: ISMIR (2003)
9. Harte, C., Sandler, M., Gasser, M.: Detecting harmonic change in musical audio. In: Audio and Musical Computing for Multimedia Workshop 2006 (in conjunction with ACM Multimedia) (2006)
10. Haus, G., Pollastri, E.: An audio front end for query-by-humming systems. In: Proceedings of the International Conference on Music Information Retrieval, ISMIR (2001)
11. Pauws, S.: CubyHum: a fully operational query by humming system. In: ISMIR, pp. 187–196 (2002)
12. Ryynänen, M.: Probabilistic Modelling of Note Events in the Transcription of Monophonic Melodies, Master's thesis, Tampere University of Technology (2004)
13. Ryynänen, M., Klapuri, A.: Query by humming using locality sensitive hashing for MIREX 2008 (2008), MIREX abstract, http://www.music-ir.org/mirex/2008/abs/QBSH_ryynanen.pdf
14. Ryynänen, M., Klapuri, A.: Transcription of the singing melody in polyphonic music (2006), MIREX abstract, http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AME_ryynanen.pdf

On Similarity Search in Audio Signals Using Adaptive Sparse Approximations

Bob L. Sturm[1] and Laurent Daudet[2]

[1] Department of Architecture, Design and Media Technology
(http://media.aau.dk) Aalborg University Copenhagen
Lautrupvang 15, 2750 Ballerup, Denmark
bst@create.aau.dk

[2] Université Paris Diderot, Paris 7
Institut Langevin (LOA), UMR 7587
10, rue Vauquelin, 75231 Paris Cedex 05, France
laurent.daudet@espci.fr

Abstract. We explore similarity search in data compressed and described by adaptive methods of sparse approximation, specifically audio signals. The novelty of this approach is that one circumvents the need to compute and store a database of features since sparse approximation can simultaneously provide a description and compression of data. We investigate extensions to a method previously proposed for similarity search in a homogenous image database using sparse approximation, but which has limited applicability to search heterogeneous databases with variable-length queries — necessary for any useful audio signal search procedure. We provide a simple example as a proof of concept, and show that similarity search within adapted sparse domains can provide fast and efficient ways to search for data similar to a given query.

1 Introduction

Similarity search in time series databases is typically addressed with techniques of feature extraction and nearest neighbor search methods within an economically indexable space of much smaller dimension than the original data (e.g., [1,2]). One creates feature vectors that describe time series using a variety of linear and non-linear transforms, such as a truncated discrete Fourier transform [1], dyadic wavelet transforms [2], or even transforms of transforms, such as Mel-frequency cepstral coefficients (MFCCs) [3]. For databases of sampled audio, the use of MFCCs has proven useful for a variety of tasks, for instance, fingerprinting and music identification, content classification, and cover song identification (e.g., [3,4,5]). There also exists a variety of distance measures in, as well as indexing schemes for, feature spaces, which allow economic ways to search and retrieve data based on a query [1,4].

Features created from compressed audio data, e.g., MPEG-1, layer 3 have been used in, e.g., tasks of content classification [6], but the descriptiveness and applicability of these features have limits due to the transforms employed [7]. Indeed, such schemes for audio compression are designed first and foremost to

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 59–71, 2011.

compress audio data to perceptual transparency, not to accurately describe it. A compelling alternative is given by methods of sparse approximation [8,9]. Such approaches can provide audio compression schemes competitive with standard ones at low bit rates [10], while possessing features with much better time and frequency resolution [7]. Additionally, a sparse representation is hierarchical in that the most energetic signal content is generally represented before the least energetic. It is also adaptive since terms are selected and placed in a representation where they fit the data the best — as opposed to a transform, e.g., Fourier, that uses the same basis functions no matter the signal content. What is more, sparse approximation has scalable precision; in other words, approximations are progressively made more precise by including more terms. This means that at levels of low-resolution one can obtain an idea of the general behaviors of the data — a property exploited by time series similarity matching using wavelets [2]. All these observations suggest that sparse approximation can be useful for data compression, as well as for economically and efficiently searching it. Initially, however, there is a high computational price to pay in sparse approximation as the decomposition procedures are complex. A decomposition only needs to be done once though, to generate the signal representation.

Jost et al. [11] have explored similarity search within sparse and descriptive domains created by sparse approximation. This approach, however, restricts all data to be the same size, and is not invariant to translation — rendering it useless for similarity search in a database of heterogenous audio signals. In this paper we elaborate on this method for use in the task of similarity search in audio signals. We propose a translation-invariant method that permits subsequence search, and show with a simple example that similarity search within the sparse domain can be effective, even when a query is corrupted by noise. Our work demonstrates that sparse approximation, in addition to being useful for audio signal compression [10], provides an accurate and searchable description of audio data that is useful for similarity search, and is scalable to large databases. This ultimately eliminates the need for a database of features separate from the compressed data itself.

The outline of this paper is as follows. The next two subsections review sparse approximation, and the procedure of similarity search of Jost et al. [11]. Section 2 contains our elaborations of this method, specifically addressing its shift-variance and restrictions on signal size, within the context of similarity search in audio signals. The third section provides a proof of concept demonstrating that subsequence similarity search within a sparse domain is possible and gives sensible results. We also provide a brief analysis of the computational complexity of this approach.

1.1 Sparse Approximation

Consider the discrete finite-length signal $\mathbf{x} \in \mathbb{R}^K$, and a full-rank matrix (dictionary) composed of unit-norm vectors (atoms) from the same space $\mathbf{D} \in \mathbb{R}^{K \times N}$, where usually $N \gg K$. Sparse approximation attempts to find a solution $\mathbf{s} \in \mathbb{R}^N$ to the constrained problem

$$\min_{\mathbf{s}} \#\{|\mathbf{s}| > 0\} \text{ subject to } ||\mathbf{x} - \mathbf{D}\mathbf{s}||_2^2 \leq \delta \tag{1}$$

that is, find the vector $\mathbf{s}$ with the largest number of zero elements that results in a squared error of at most $\delta \geq 0$. Essentially, sparse approximation is the modeling of a function using few atoms from the dictionary. The pursuit $\mathscr{P}_{\mathbf{D}}\{\mathbf{x}\}$ decomposes $\mathbf{x}$ in terms of the dictionary $\mathbf{D}$ such that if $\mathbf{s}' \leftarrow \mathscr{P}_{\mathbf{D}}\{\mathbf{x}\}$, then $||\mathbf{x} - \mathbf{D}\mathbf{s}'||_2^2 \leq \delta$. This may or may not be the solution to (1), but is at least much more sparse than $\mathbf{x}$ itself (i.e., $\#\{|\mathbf{s}| > 0\} \ll K$). We can express the linear combination $\mathbf{D}\mathbf{s}'$ in terms of the $n \overset{\Delta}{=} \#\{|\mathbf{s}'| > 0\}$ non-zero elements in $\mathbf{s}'$:

$$\mathbf{x} = \mathbf{H}(n)\mathbf{a}(n) + \mathbf{r}(n) = \hat{\mathbf{x}}(n) + \mathbf{r}(n) \tag{2}$$

where $\mathbf{a}(n)$ is a length-n vector created by removing all $N - n$ zeros from $\mathbf{s}'$, $\mathbf{H}(n)$ are the corresponding columns of $\mathbf{D}$, and $\mathbf{r}(n) = \mathbf{x} - \hat{\mathbf{x}}(n)$ is the error satisfying $||\mathbf{r}(n)||_2^2 \leq \delta$.

Many methods have been proposed to solve (1) exactly or approximately. Basis Pursuit [9] proposes to relax the sparsity constraint ($\#\{|\mathbf{s}_i| > 0\}$) to the ℓ_1-norm ($||\mathbf{s}||_1$), which can then be solved using convex optimization. Matching Pursuit (MP) [8] is a greedy iterative descent algorithm that builds a solution $\mathbf{s}$ iteratively based on a measure of similarity between an intermediate residual and atoms in the dictionary. Of the two, the iterative approach of MP provides solutions that are suboptimal with respect to the ℓ_1-norm, but is much less computationally complex, and can provide representations that are simultaneously sparse and descriptive of high-dimensional audio signals [10,7].

1.2 Image Similarity Search Using Sparse Approximations

Jost et al. [11] have used sparse approximation to facilitate similarity search in a database of images. Consider performing a pursuit with a full-rank dictionary $\mathbf{D}$ on a query signal $\mathbf{x}_q \in \mathbb{R}^K$, $||\mathbf{x}_q||_2^2 = 1$, $\{\mathbf{H}_q(n_q), \mathbf{a}_q(n_q), \mathbf{r}_q(n_q)\} \leftarrow \mathscr{P}_{\mathbf{D}}\{\mathbf{x}_q\}$, as well as on a set of signals of the same dimension $\mathcal{Y} \overset{\Delta}{=} \{\mathbf{y}_i \in \mathbb{R}^K, ||\mathbf{y}_i||_2^2 = 1\}_{i \in \mathcal{I}}$

$$\mathcal{Y}_s \overset{\Delta}{=} \left\{ \{\mathbf{H}_i(n_i), \mathbf{a}_i(n_i), \mathbf{r}_i(n_i)\} \leftarrow \mathscr{P}_{\mathbf{D}}\{\mathbf{y}_i\} \right\}_{i \in \mathcal{I}}. \tag{3}$$

We wish to find the elements of $\mathcal{Y}$ that are similar to the query, e.g., those elements that have a minimum Euclidean distance to $\mathbf{x}_q$. In this respect, the $\mathbf{y}_i \in \mathcal{Y}$ most similar to $\mathbf{x}_q$ is given by solving

$$\min_{i \in \mathcal{I}} ||\mathbf{y}_i - \mathbf{x}_q||_2^2 = \max_{i \in \mathcal{I}} \langle \mathbf{x}_q, \mathbf{y}_i \rangle = \max_{i \in \mathcal{I}} \mathbf{y}_i^T \mathbf{x}_q. \tag{4}$$

However, to avoid taking $|\mathcal{I}|$ inner products for a query, Jost et al. [11] solve this more efficiently by translating it to the sparse domain. Using the sparse representations of each signal, one can express (4) as

$$\max_{i\in\mathcal{I}} \langle \mathbf{H}_q(n_q)\mathbf{a}_q(n_q), \mathbf{H}_i(n_i)\mathbf{a}_i(n_i)\rangle = \max_{i\in\mathcal{I}} \mathbf{a}_i^T(n_i)\mathbf{H}_i^T(n_i)\mathbf{H}_q(n_q)\mathbf{a}_q(n_q)$$
$$= \max_{i\in\mathcal{I}} \mathbf{a}_i^T(n_i)\mathbf{G}_{iq}\mathbf{a}_q(n_q) = \max_{i\in\mathcal{I}} \sum_{m=1}^{n_i}\sum_{l=1}^{n_q}[\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{ml} \quad (5)$$

where $[\mathbf{B} \bullet \mathbf{C}]_{ml} = [\mathbf{B}]_{ml}[\mathbf{C}]_{ml}$ is the Hadamard product ($[\mathbf{B}]_{ml}$ is the mth row of the lth column of $\mathbf{B}$), $\mathbf{G}_{iq} \overset{\Delta}{=} \mathbf{H}_i^T(n_i)\mathbf{H}_q(n_q)$ is a matrix with elements from the Gramian of the dictionary, i.e., $\mathbf{D}^T\mathbf{D}$, and the outer product of the weights $\mathbf{A}_{iq}$ is defined

$$\mathbf{A}_{iq} \overset{\Delta}{=} \mathbf{a}_i(n_i)\mathbf{a}_q^T(n_q). \quad (6)$$

We can write the sum in (5) in terms of a sum involving $1 \le M < \min(n_q, n_i)$ atoms from each representation, as well as a remainder term

$$S_{iq}(M) \overset{\Delta}{=} \sum_{m=1}^{M}\sum_{l=1}^{M}[\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{ml} \quad (7)$$
$$R_{iq}(M) \overset{\Delta}{=} \sum_{m=1}^{n_i}\sum_{l=M+1}^{n_q}[\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{ml} + \sum_{m=M+1}^{n_i}\sum_{l=1}^{M}[\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}]_{ml}. \quad (8)$$

For a given M we will find $\{S_{iq}(M)\}_{i\in\mathcal{I}}$, estimate the remainders $\{R_{iq}(M)\}_{i\in\mathcal{I}}$, and compare bounds to eliminate signals that are not similar to the query. We estimate these bounds by first making the assumption that the elements of $\mathbf{a}_i(n_i)$ and $\mathbf{a}_q(n_q)$ are ordered in terms of decreasing magnitude, and that they decay exponentially, e.g., for the query signal

$$0 < |[\mathbf{a}_q(n_q)]_m| \le Cm^{-\gamma},\ m = 1, 2, \ldots \quad (9)$$

(and similarly for each $\mathbf{a}_i(n_i)$) with $C, \gamma > 0$ depending in a complex way on the signal, the dictionary, and the pursuit. Such decay is guaranteed for MP since for all full-rank dictionaries the residual energy is monotonic decreasing [8]. Since $|[\mathbf{G}_{iq}]_{ml}| \le 1$ because the dictionaries consist of unit-norm atoms, then we can find a loose upper bound on the remainder (8)

$$R_{iq}(M) \le C^2 \sum_{m=1}^{n_i}\sum_{l=M+1}^{n_q}(ml)^{-\gamma} + C^2 \sum_{m=M+1}^{n_i}\sum_{l=1}^{M}(ml)^{-\gamma} \quad (10)$$

Defining this upper bound as $\widetilde{R}_{iq}(M)$, we obtain the following loose upper and lower bounds on (5) such that $L_{iq}(M) \le \langle \mathbf{x}_q, \mathbf{y}_i\rangle \le U_{iq}(M)$:

$$L_{iq}(M) \overset{\Delta}{=} S_{iq}(M) - \widetilde{R}_{iq}(M) \quad (11)$$
$$U_{iq}(M) \overset{\Delta}{=} S_{iq}(M) + \widetilde{R}_{iq}(M). \quad (12)$$

For MP and a full-rank dictionary, these bounds converge to the true value of $\langle \mathbf{x}_q, \mathbf{y}_i \rangle$ as M, n_i and n_q approach infinity.

Given that we have estimated C and γ from the set of representations, the problem of finding the most similar database element to the query can be done iteratively over M in the following way [11]. We first assume that every element of $\mathcal{Y}$ is similar to the query, and then reduce the number of good matches by successively considering pairs of atoms from each representation. Starting with $M = 1$, we compute the sets $\{L_{iq}(1)\}_{i \in \mathcal{I}}$ and $\{U_{iq}(1)\}_{i \in \mathcal{I}}$, that is, the first-order upper and lower bounds of the set of correlations of the query signal with all signals in the database. Then we find the signal with the largest lower bound $i_{\max} \overset{\Delta}{=} \arg\max_{i \in \mathcal{I}} L_{iq}(1)$, and reduce the search space to $\mathcal{I}_1 \overset{\Delta}{=} \{i \in \mathcal{I} : U_{iq}(1) \geq L_{i_{\max}q}(1)\}$, since all other data have a smaller upper bound on their correlation with the query than the largest lower bound in the set. If $|\mathcal{I}_1| > 1$, then we set $M = 2$, compute the sets $\{L_{iq}(2)\}_{i \in \mathcal{I}_1}$ and $\{U_{iq}(2)\}_{i \in \mathcal{I}_1}$, find the index of the maximum $i_{\max} \overset{\Delta}{=} \arg\max_{i \in \mathcal{I}_1} L_{iq}(2)$, and construct the reduced set $\mathcal{I}_2 \overset{\Delta}{=} \{i \in \mathcal{I}_1 : U_{iq}(2) \geq L_{i_{\max}q}(2)\}$. In this way the set of signals most similar to the query is pared to those that match $\mathbf{x}_q$ the best.

2 Similarity Search for Sparsely Represented Audio

As such, the method of Jost et al. [11] has limited applicability to similarity search in audio signals, and image data in general. First, it requires all elements of the database and the query to be of the same dimension. It cannot be used to find a smaller-dimension object in a heterogeneous set of images, for instance. Second, the search returns different results if the query data is simply translated or rotated. These restrictions are problematic for searching audio signals where time shift is natural. We now address these restrictions so that the method is useful to similarity search in audio signals.

We want to efficiently and accurately search a heterogeneous database to find every similar instance of some query independent of its amplitude, location, or the surrounding data. This problem has of course been addressed in numerous ways for audio data (e.g., [3,5,4]); here, however, we address it in the sparse domain. As in Section 1.2, consider the query $\mathbf{x}_q \in \mathbb{R}^K$, $||\mathbf{x}_q||_2^2 = 1$, and a collection of vectors (that may not be unit-norm) in several high-dimensional spaces

$$\mathcal{Y} = \{\mathbf{y}_i \in \mathbb{R}^{N_i} : N_i > K\}_{i \in \mathcal{I}}. \tag{13}$$

As in [11] we measure similarity of two vectors by their Euclidean distance; but in this case, we must look at the minimum possible distance between $\mathbf{x}_q$ and each K-length subsequence of $\mathbf{y}_i \in \mathcal{Y}$. We define the distance by

$$\begin{aligned} D(\mathbf{P}_K(t)\mathbf{y}_i, \mathbf{x}_q) &\overset{\Delta}{=} \min_{\alpha} ||\mathbf{P}_K(t)\mathbf{y}_i - \alpha \mathbf{x}_q||_2^2 \text{ subject to } ||\mathbf{P}_K(t)\mathbf{y}_i||_2 > 0 \\ &= \min_{\alpha} ||\mathbf{P}_K(t)\mathbf{y}_i||_2^2 + |\alpha|^2 - 2\alpha \langle \mathbf{P}_K(t)\mathbf{y}_i, \mathbf{x}_q \rangle \\ &\qquad \text{subject to } ||\mathbf{P}_K(t)\mathbf{y}_i||_2 > 0 \end{aligned} \tag{14}$$

where the $K \times N_i$ matrix $\mathbf{P}_K(t) \overset{\Delta}{=} [\mathbf{0}_{K\times t}|\mathbf{I}_K|\mathbf{0}]$ is defined for $0 \leq t \in \mathbb{Z} < N_i - K$, with $\mathbf{I}_K$ the identity matrix of size K, and $\mathbf{0}$ is an appropriately-sized zero matrix, and α is some scale to be determined. We add the constraint to ensure that we are looking at a subsequence of $\mathbf{y}_i$ that has energy. Since the minimum occurs for $\alpha = \langle \mathbf{P}_K(t)\mathbf{y}_i, \mathbf{x}_q \rangle$, we can find the shift $0 \leq t < N_i - K$ at which (14) is minimized for each $\mathbf{y}_i \in \mathcal{Y}$ by the equivalent maximization problem

$$t_i = \arg \max_{0 \leq t < N_i - K} \frac{|\langle \mathbf{P}_K(t)\mathbf{y}_i, \mathbf{x}_q \rangle|}{||\mathbf{P}_K(t)\mathbf{y}_i||_2} \text{ subject to } ||\mathbf{P}_K(t)\mathbf{y}_i||_2 > 0. \tag{15}$$

Finally, analogous to (4), given the set $\mathcal{T}_\mathcal{I} = \{t_i\}_{i\in\mathcal{I}}$, i.e., the set of time translations at which each element of $\mathcal{Y}$ has a minimum in (14), the signal with content that is most similar to the query is given by solving

$$\max_{i\in\mathcal{I}} \frac{|\langle \mathbf{P}_K(t_i)\mathbf{y}_i, \mathbf{x}_q \rangle|}{||\mathbf{P}_K(t_i)\mathbf{y}_i||_2}. \tag{16}$$

There are two essential differences between (16) and (4). First, since we are considering subsequences, we must consider many translations of the query along the support of each element of $\mathcal{Y}$. Second, while (4) reduces to maximizing a single inner product, (16) entails a ratio of two values since $||\mathbf{P}_K(t)\mathbf{y}_i||_2$ is not assumed constant at every shift t of every signal $\mathbf{y}_i$. For these reasons, adapting the method of [11] to subsequence similarity search is not trivial.

We now express (15) in a sparse domain. Consider several full-rank dictionaries: $\mathbf{D}_K$ defined over $\mathbb{R}^K$, and $\mathbf{D}_{N_i}$ defined over $\mathbb{R}^{N_i}$. Using some pursuit, e.g., MP [8], we produce sparse representations of the query $\{\mathbf{H}_q(n_q), \mathbf{a}_q(n_q), \mathbf{r}_q(n_q)\} \leftarrow \mathscr{P}_{\mathbf{D}_K}\{\mathbf{x}_q\}$, as well as each element in $\mathcal{Y}$ (13):

$$\mathcal{Y}_s = \left\{ \{\mathbf{H}_i(n_i), \mathbf{a}_i(n_i), \mathbf{r}_i(n_i)\} \leftarrow \mathscr{P}_{\mathbf{D}_{N_i}}\{\mathbf{y}_i\} \right\}_{i\in\mathcal{I}}. \tag{17}$$

We express the numerator in (15) as

$$\begin{aligned} |\langle \mathbf{P}_K(t)\mathbf{H}_i(n_i)\mathbf{a}_i(n_i), \mathbf{H}_q(n_q)\mathbf{a}_q(n_q)\rangle| &= \left|\mathbf{a}_i^T(n_i)\mathbf{H}_i^T(n_i)\mathbf{P}_K^T(t)\mathbf{H}_q(n_q)\mathbf{a}_q(n_q)\right| \\ &= \left|\mathbf{a}_i'^T(n_i')\mathbf{G}_{iq}(t)\mathbf{a}_q(n_q)\right| \end{aligned} \tag{18}$$

where we have defined the subspace Gramian $\mathbf{G}_{iq}(t)$ from the projection of the n_i' non-zero K-length columns of $\mathbf{P}_K(t)\mathbf{H}_i(n_i)$ onto $\mathbf{H}_q(n_q)$, and we have defined the possibly shortened weight vector $\mathbf{a}_i'(n_i')$ as the weights associated with the atoms with support in $[t, t+K)$. Assuming, from the energy conservation of MP [8], that $||\mathbf{P}_K(t)\mathbf{y}_i||_2^2 \approx ||\mathbf{a}_i'(n_i')||_2^2$ (i.e., the energy of the segment is close to the squared ℓ_2-norm of the weights of atoms that exist in that segment, ignoring effects at the edges), and defining the outer product

$$\mathbf{A}_{iq}(t) \overset{\Delta}{=} \frac{\mathbf{a}_i'(n_i')\mathbf{a}_q^T(n_q)}{||\mathbf{a}_i'(n_i')||_2} \tag{19}$$

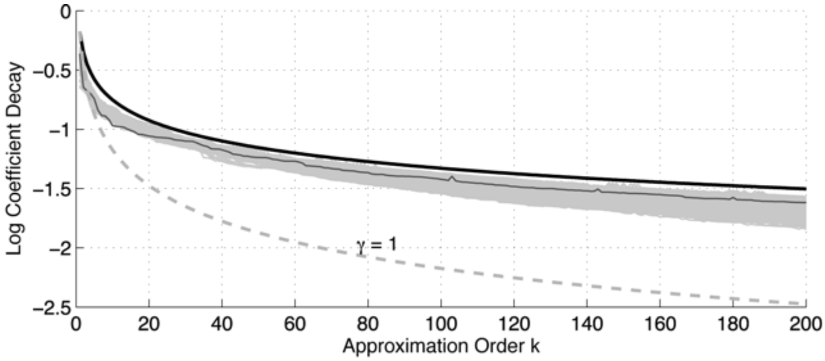


Fig. 1. Decay of elements $\log_{10}[\mathbf{a}'_i(n'_i)]_k/||\mathbf{a}'_i(n'_i)||_2$ as a function of approximation order k for several subsequences (gray), query (thin black), and the calculated bound (thick black). Dashed gray line is when $\gamma = 1$.

we can write (15) as

$$t_i = \arg \max_{0 \leq t < N_i - K} \left| \sum_{m=1}^{n'_i} \sum_{l=1}^{n_q} [\mathbf{A}_{iq}(t) \bullet \mathbf{G}_{iq}(t)]_{ml} \right| \text{ subject to } ||\mathbf{a}'_i(n'_i)||_2 > 0. \quad (20)$$

Now, for a given $\mathbf{y}_i \in \mathcal{Y}$ we want to find t_i by considering far fewer than $n'_i \times n_q$ pairs of atoms for each possible time shift.

Taking the approach by Jost et al. [11], we break apart the sum in (20) into

$$S_{iq}(t, M) \triangleq \sum_{m=1}^{M} \sum_{l=1}^{M} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}(t)]_{ml} \quad (21)$$

$$R_{iq}(t, M) \triangleq \sum_{k=1}^{n'_i} \sum_{l=M+1}^{n_q} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}(t)]_{ml} + \sum_{k=M+1}^{n'_i} \sum_{l=1}^{M} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}(t)]_{ml} \quad (22)$$

for $1 \leq M < \min(n'_i, n_q)$. We assume the weights decay exponentially as in (9). Finally, we can state as in (10) that

$$R_{iq}(t, M) \leq C^2 \sum_{m=1}^{n'_i} \sum_{l=M+1}^{n_q} (ml)^{-\gamma} + C^2 \sum_{m=M+1}^{n'_i} \sum_{l=1}^{M} (ml)^{-\gamma} \triangleq \widetilde{R}_{iq}(t, M) \quad (23)$$

which gives bounds like (11) and (12).

Before we use the approach of Jost et al. [11] to narrow down the set of similar signals based on these bounds, we must estimate C and γ. Since we assume (9) holds for both the sparse representations of the query signal and the database signals, then given $\mathbf{a}'_i(n'_i)$ from the shift t and the sparse approximation of $\mathbf{y}_i$, we

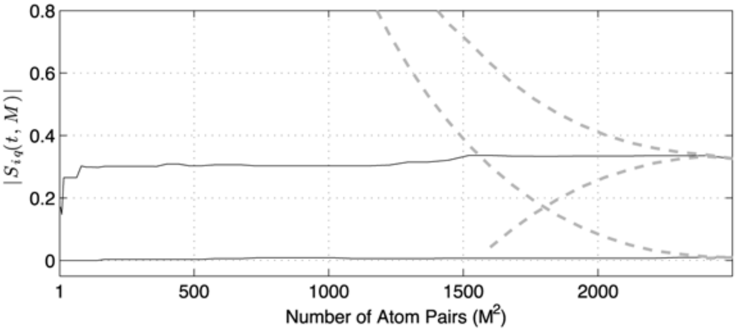


Fig. 2. $|S_{iq}(t, M)|$ with bounds $\pm\widetilde{R}_{iq}(t, M)$ (dashed gray) for two different subsequences $(n_i', n_q = 1000)$ as a function of the pairs of atoms considered M^2

know $\log_m C_i - \gamma_i \geq \log_m |[\mathbf{a}_i'(n_i')]_m|$. Setting $C' = |[\mathbf{a}_i'(n_i')]_1|$, we find an upper bound on γ_i by

$$\gamma_i \leq \frac{1}{n_i' - 1} \sum_{l=2}^{n_i'} \log_l \frac{C'}{|[\mathbf{a}_i(n_i')]_l|}. \tag{24}$$

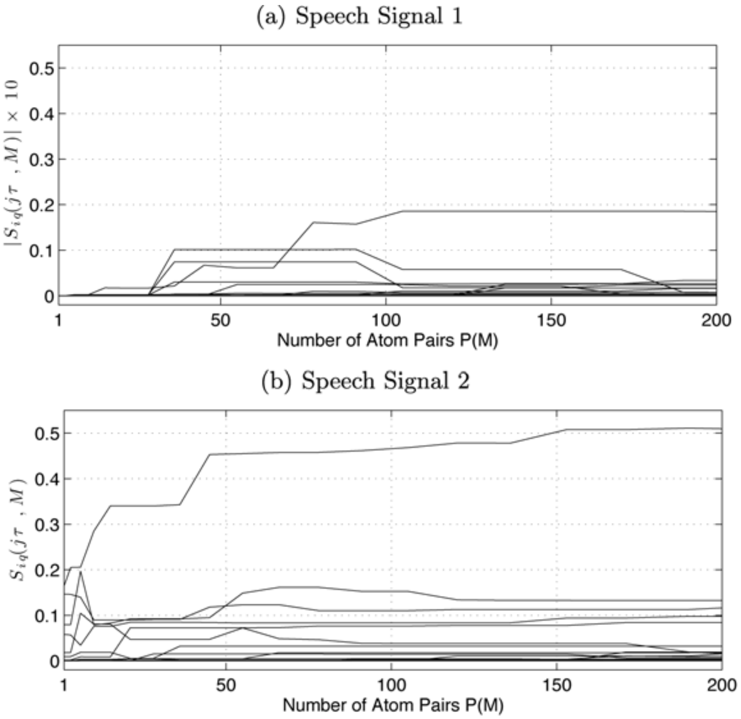
Using this bound, we set $C_i = |\,[\mathbf{a}_i'(n_i')]_2\,|2^{\gamma_i}$. (We chose this experimentally as it appears to give the tightest bounds for our data.) Using the sparse representation of the query, and all length-K subsequences of $\mathbf{y}_i$, we set $C = \max\{\{C_i\}_t, C_q\}$ and $\gamma = \min\{\{\gamma_i\}_t, \gamma_q\}$, where C_q and γ_q are found from the query signal representation using the same approach as above. Figure 1 shows the decay of the elements in $\mathbf{a}_i'(n_i')$ for several subsequences and the bound. In this case $C = 0.6702$, and $\gamma = 0.5773$.

Now we begin to see a problem. When $\gamma < 1$ (bound shown in Fig. 1), then $\widetilde{R}_{iq}(t, M)$ (23) decays slowly, implying that the bounds on $|S_{iq}(t, M)|$ are not useful until M is large. Figure 2 illustrates this problem for a query matched to two different subsequences from the same $\mathbf{y}_i$. Until $\widetilde{R}_{iq}(t, M) < 1$, we are unable to say if one subsequence is more similar to the query than another, which in this case occurs after we have looked at 1 764 pairs of atoms. We can increase γ by using a dictionary that is more "coherent" with the signals, thus giving a quicker decay; or we can take a probabilistic approach [11], determining which segments are likely to have upper bounds smaller than the largest lower bound. There may be no need to do either of these, however, as a simple threshold test may be sufficient to find a small subset of candidates in $\mathcal{Y}$.

Instead of starting with the assumption that every signal is equally similar to the query, we will assume that none are. Let us redefine $S_{iq}(t, M)$ (21) as a recursive sum along the anti-diagonals of the weight matrix:

$$S_{iq}(t, M) \stackrel{\Delta}{=} S_{iq}(t, M-1) + \sum_{m=1}^{M} \sum_{l=0}^{M-1} [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}(t)]_{m(M-l)} \tag{25}$$

for $M = 2, \ldots, \min(n_i', n_q)$, and $S_{iq}(t, 1) = [\mathbf{A}_{iq} \bullet \mathbf{G}_{iq}(t)]_{11}$. We do this instead of (21) because the decay rate is fastest in the diagonal direction of $\mathbf{A}_{iq}$. At


(a) Speech Signal 1

(b) Speech Signal 2

Fig. 3. $|S_{iq}(t, M)|$ (25) for subsequences of two speech signals with the same query as a function of (26) the number of atom pairs considered. Note change in y-axis scaling.

step M, we are considering M additional atom pairs to those considered in the previous step. Thus, the number of atom pairs considered in (25) at step M is

$$P(M) \stackrel{\Delta}{=} \sum_{m=1}^{M} m = \frac{M(M+1)}{2}. \tag{26}$$

Figure 3 shows $|S_{iq}(t, M)|$ as a function of $P(M)$ for several subsequences from two different speech signals and a query signal extracted from one of them. The similarity (25) is evaluated at integer multiples $j \in \mathcal{J}_i \stackrel{\Delta}{=} \{0, 1, \ldots, \lfloor (N_i - K)/\tau \rfloor - 1\}$, of a fundamental shift of $\tau \in \mathbb{N}$ samples. Each line in Fig. 3 represents the similarity calculated at one shift. It is obvious that we only need a small number of atom pairs to decide whether there is sufficient evidence of similarity between a subsequence and the query with respect to (15), and consequently if one signal has content more similar than another to the query. We explore this in the next section with a simple experiment.

3 Computer Simulations

To investigate the behavior of (25), we designed the following simple experiment. We concatenated six speech signals, a 3-second wide bandwidth music

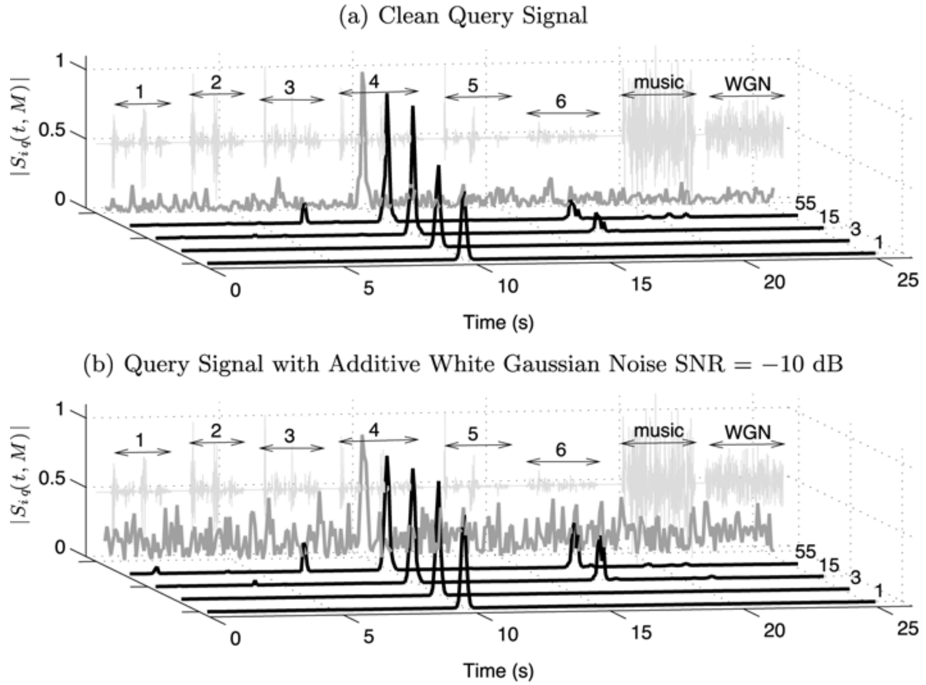
Table 1. Dictionary parameters (16 kHz sampling rate): scale s, time shifts Δ_u, and frequency resolution Δ_f

s (samples/ms)	Δ_u (samples/ms)	Δ_f (Hz)
4/0.25	1/0.06	4000
8/0.50	2/0.13	2000
16/1	4/0.25	1000
32/2	8/0.5	500
64/4	16/1	250
128/8	32/2	125
256/16	64/4	62.5
512/32	128/8	31.25
1024/64	256/16	15.63
2048/128	512/32	7.81
4096/256	1024/64	3.91
8192/512	2048/128	1.95

signal (strings and percussion), and a 3-second realization of a white Gaussian noise (WGN) process, to create a signal of dimension 411 862 samples (25.74 s) at 16 kHz sampling rate. All speech signals are from different speakers (three female, three male) saying: "Cottage cheese with chives is delicious." The query signal is the word "cheese" selected from one of the male speakers, and has a dimension of 9 347 samples (584 ms). We decomposed both signals using MP with a multiresolution time-frequency Gabor dictionary (scaled, modulated, and translated Gaussian functions) detailed in Table 1. For instance, the dictionary contains Gabor atoms of scale 16 samples (1 ms), with translations every 4 samples (0.25 ms), and spaced in frequency by 1000 Hz between 0 Hz and 8 kHz inclusive. We then computed (25) for several M, considering time segments shifted by one-eighth the query length (73 ms) — which partitions the signal into 345 segments. For these signals we estimated $\gamma = 0.1296$, so small that it diminishes the applicability of the search method using bounds of Jost et al. [11].

Figure 4 shows how $|S_{iq}(t,M)|$ changes as more atom pairs are considered for a speech query signal directly selected from the original signal that is (a) clean; and (b) corrupted by additive white Gaussian noise with -10 dB SNR. In the background of Fig. 4 we plot the signal, where each of the different signals are marked — six speech samples, one music signal, and one white Gaussian noise signal (WGN). Each black line in front of this is $|S_{iq}(t,M)|$ as a function of shift t, and each is marked on the right with the associated value of $P(M)$ (26), the number of atom pairs considered at each time shift. The thick dark gray line is the normalized magnitude correlation for every possible shift of the query.

From this data we can tell that the query signal comes from the fourth speech signal. The first black line in the foreground is $|S_{iq}(t,1)|$, which is evaluated with only one atom pair at each time shift. We clearly see a maximum occurs near the position at which the query exists. At larger $P(M)$ we see that other regions of the signal are similar to the query. These correspond to the same word, "cheese," but spoken by other subjects — including a female ("3"). We observe the same



(a) Clean Query Signal

(b) Query Signal with Additive White Gaussian Noise SNR = −10 dB

Fig. 4. $|S_{iq}(t, M)|$ (25) for subsequences of the audio signal shown (far back) for several $P(M)$ (shown at right) using (a) clean query signal, and (b) query signal with additive white Gaussian noise SNR = −10 dB. Thick gray line at back is the actual magnitude correlation. Signal contents are labeled, with speakers numbered.

behavior in Figure 4(b) even when the query is corrupted by AWGN at an SNR of −10 dB, i.e., $10 \log_{10} ||\alpha \mathbf{x}_q||_2^2/||\mathbf{n}||_2^2 = -10$ dB, where $\mathbf{x}_q$ is the original query, $\mathbf{n}$ is the noise signal, and $\mathbf{x}'_q = (\alpha \mathbf{x}_q + \mathbf{n})/||\alpha \mathbf{x}_q + \mathbf{n}||_2$ is the new query that we make unit norm.

The computational complexity of this method for finding the portion of a signal most similar to the query is $\mathcal{O}(TLP(M))$ (assuming the pursuit has already run, and the dictionary elements are tabulated), where T is the number of time segments (linearly dependent on N, the size of the signal being searched), L is the size of the segments (or maximum atom scale in dictionary for the L-point multiply of correlation), and $P(M) = M(M+1)/2$ are the number of atom pairs considered each time segment. Considering that $L \ll N$, and M and T are small, this method is much less computationally expensive than finding the direct correlation in the frequency domain, which has complexity $\mathcal{O}((N+L-1)[3 \log_2(N+L-1)+1])$ (three Fourier transforms and one $N+L-1$-point multiply), and is very sensitive to noise. In the specific example above, the direct correlation requires on the order of 23 million multiplies. Using just three atom pairs per time segment requires on the order of one-third as many. We can further reduce this if we use a closed-form expression for the inner-product of two discrete Gabor atoms, or tabulate their values before hand. In this case, the complexity reduces to $\mathcal{O}(T)$ by removing the need to perform $P(M)$, L-point correlations.

4 Conclusion

We have explored similarity search in audio signals that are simultaneously compressed and described by methods of sparse approximation. Starting from the method proposed by Jost et al. [11] for image similarity search, we have investigated how to address its restrictions (homogenous database, shift-variant) so that it is applicable to similarity search in audio signals by permitting variable-length queries and shift-invariance. We have performed a simple test as a proof of concept that within the sparse domain we can find those portions of the signal that are similar to a query with much less complexity than using direct correlation methods, and furthermore without calculating features extraneous to the data representations themselves. Future work will extend this method to searching large databases of compressed audio, and compare its precision and recall performance to other approaches to similarity search for audio signals.

Acknowledgments

B. L. Sturm is supported by the Chateaubriand Fellowship N. 634146B. L. Daudet acknowledges partial support from the French Agence Nationale de la Recherche under contract ANR-06-JCJC-0027-01 DESAM.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Proc. Int. Conf. Foundations Data Org. Algo., Chicago, IL, October 1993, pp. 69–84 (1993)
2. Chan, K.P., Fu, A.W.C.: Efficient time series matching by wavelets. In: Proc. Int. Conf. Data Eng., Sydney, Australia, March 1999, pp. 126–133 (1999)
3. Wold, E., Blum, T., Keislar, D., Wheaton, J.: Content-based classification, search, and retrieval of audio. IEEE Multimedia 3(2), 27–36 (Fall 1996)
4. Casey, M., Rhodes, C., Slaney, M.: Analysis of minimum distances in high-dimensional musical spaces. IEEE Trans. Audio, Speech, Lang. Process. 16(5), 1015–1028 (2008)
5. Foote, J.T.: Content-based retrieval of music and audio. In: Proc. SPIE Multimedia Storage Archiving Syst., Dallas, TX, November 1997, pp. 138–147 (1997)
6. Nakajima, Y., Lu, Y., Sugano, M., Yoneyama, A., Yamagihara, H., Kurematsu, A.: A fast audio classification from mpeg coded data. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., Phoenix, AZ, March 1999, vol. 6, pp. 3005–3008 (1999)
7. Ravelli, E., Richard, G., Daudet, L.: Audio signal representations for indexing in the transform domain. IEEE Trans. Acoustics, Speech, Lang. Process. (2010) (accepted for publication)

8. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. 41(12), 3397–3415 (1993)
9. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20(1), 33–61 (1998)
10. Ravelli, E., Richard, G., Daudet, L.: Union of MDCT bases for audio coding. IEEE Trans. Audio, Speech, Lang. Proc. 16(8), 1361–1372 (2008)
11. Jost, P., Vandergheynst, P.: On finding approximate nearest neighbours in a set of compressible signals. In: Proc. European Signal Process. Conf., Lausanne, Switzerland, August 2008, pp. 1–5 (2008)

Relevance Feedback for the Earth Mover's Distance

Marc Wichterich, Christian Beecks, Martin Sundermeyer, and Thomas Seidl

RWTH Aachen University, Germany
{wichterich,beecks,sundermeyer,seidl}@cs.rwth-aachen.de

Abstract. Expanding on our preliminary work [1], we present a novel method to heuristically adapt the Earth Mover's Distance to relevance feedback. Moreover, we detail an optimization-based method that takes feedback from the current and past Relevance Feedback iterations into account in order to improve the degree to which the Earth Mover's Distance reflects the preference information given by the user. As shown by our experiments, the adaptation of the Earth Mover's Distance results in a larger number of relevant objects in fewer feedback iterations compared to existing query movement techniques for the Earth Mover's Distance.

1 Introduction

Two common challenges of distance-based similarity search are the formulation of the query and the definition of a suitable distance measure. By definition of similarity search, query objects can be no more than imprecise descriptions of what users are looking for in the database. Additionally, even for high-quality similarity models, the distance measure can only be an approximation of the users' notion of similarity as said notion may be different from application to application, from user to user, and ultimately from query to query.

Relevance Feedback (RF) approaches [2,3,4,5] address these issues on the basis of relevance information gathered from the user. They aim to increasingly reflect the user's notion of similarity and return a larger amount of relevant objects from the database. Returning more relevant objects in earlier iterations turns effectiveness of the similarity model into efficiency for the user in this scenario.

Expanding on our preliminary work in [1], we present a novel statistics-based heuristic that adapts a highly flexible, high-quality similarity measure called the Earth Mover's Distance (EMD) [6] to feedback information. Moreover, we detail an extension of the traditional feedback loop that takes an adaptable similarity measure (i.e., the EMD) and significantly improves the quality of the search result returned to the user. To this end, we formulate the similarity adaptation as an optimization problem that minimizes discrepancies between relevance information and the distance measure. Experiments on real world databases show that significantly more relevant objects are returned in fewer iterations resulting in a faster exploration of the database than existing EMD-based query movement techniques allow for.

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 72–86, 2011.

2 Related Work

We first review related work for Relevance Feedback and for the EMD separately and then detail related work that combines the two.

Relevance Feedback. In distance-based Relevance Feedback on multimedia databases, similarity between the query and objects in a database is modeled via a distance function. This framework allows for the utilization of database techniques such as spatial access methods. Based on feature vector representations of objects, systems such as MindReader [3] and MARS [4] determine both a reformulated query and an adapted distance function based on the users' feedback. In the text retrieval domain, an algorithm for basing the adaptation of the query on an optimization process that takes the user's feedback into account was proposed in [7]. For their text retrieval system, a query consists of a number of terms and according weights that can be updated by the optimization process. An optimization-based RF approach with a framework similar to ours has recently been presented in [5]. It uses genetic programming to optimize the arithmetic combination of a number of general similarity functions. Inspired by the Simulated Annealing optimization technique, [23,8] propose a system where each iteration of a feedback process is a single iteration in an search process that becomes more and more focuses over time. Some image-retrieval systems based on RF break with the convention that each object in the database has to be described by a single vector. Region-based RF [9,10] exploits the position, size, shape, and/or feature distribution of connected regions within images. Recently, Li et al. [10] proposed representing objects as graphs with nodes that represent image regions and edges between nodes of neighboring regions. The nodes are annotated with feature information for the corresponding regions. Their RF process uses two optimization steps - one for matching graphs and the other one for updating the query graph. While the features stored in the nodes of the query graph are updated using relevance information, the structure of the query graph itself remains unchanged, which limits the flexibility of the approach.

The Earth Mover's Distance. The EMD is a highly flexible distance measure that has been successfully used for image comparison [6]. It is based on a ground distance in the feature space and can compare feature representations in the form of histograms with both fixed and adaptive binning. Its flexibility makes it well-suited for a multitude of application areas besides image retrieval that range from music retrieval [11] to vector field comparison [12] in physics.

The success of the EMD in the computer vision domain gave rise to recent research that allows for fast approximate and exact evaluation of EMD-based similarity queries in multimedia databases. Lower bounds of the EMD have been proposed in [13,6] and enable efficient search via filter-and-refine algorithms. Dimensionality reduction techniques [14,15] for the EMD can be utilized in a similar algorithmic framework. EMD-specific indexing techniques [13,16,14] make efficient access in large databases possible while embeddings and approximations [17,18,19] of the EMD allow for fast approximate similarity queries.

Relevance Feedback and the Earth Mover's Distance. While the EMD has been utilized in a limited number of Relevance Feedback algorithms [9,6], these techniques do not adapt the EMD itself but instead rely on its good default retrieval performance. Rubner et al. [6] propose a Relevance Feedback approach termed "Query-by-Refinement" that combines the feature representation from relevant objects to a representation of a new virtual query object. Thus, "Query-by-Refinement" performs query adaptation only but the EMD itself remains unchanged. In [9], the EMD is used as a kernel function for an SVM-based Relevance Feedback system. The query results are iteratively improved by re-weighting the query signature and training an SVM classifier according to user feedback. The feedback is not used to adapt the underlying EMD.

Adapting the EMD to user feedback remains an open research topic, which we address in this paper expanding our preliminary work in [1].

3 Formalization of the Similarity Model

Feature Representation. Multimedia databases typically describe the objects they contain via a distribution of features that the objects exhibit. A commonly employed type of discrete representation of the feature distribution is a feature histogram, where a histogram bin represents the features that belong to a certain partition of the feature space. For instance, a color histogram assigns pixels of an image to partitions of a color feature space such as HSV. Feature histograms with fixed binning partition the feature space once for the whole database. Both data-independent (e.g., regular grids) and data-dependent partitioning methods (e.g., via clustering of all features in the database) are commonly employed.

A more flexible way to represent feature distributions of multimedia objects are feature histograms with adaptive binning which are called signatures in the EMD-context [6]. For each object, the feature space is partitioned individually by grouping/clustering its features.

Definition 1. ***Feature Signature.*** *Given a feature space FS, an object o with a finite number of features ${f^o}_1, \ldots, {f^o}_k \in FS$, and a disjoint grouping ${C^o}_1, ..., {C^o}_d$ of the k features, the signature s^o of o is a finite subset $\{({p^o}_1, {w^o}_1), \ldots, ({p^o}_d, {w^o}_d)\}$ of $FS \times \mathbb{R}^+$ with representatives ${p^o}_j \in FS$ and weights ${w^o}_j \in \mathbb{R}^+$ given by*

$$ {p^o}_j = \frac{1}{|{C^o}_j|} \sum_{f \in {C^o}_j} f \ , \qquad {w^o}_j = \frac{|{C^o}_j|}{k} $$

Fig. 1 shows a visualization of signatures for two similar images where the feature space FS comprises both location and color information and each feature ${f^o}_i$ is e.g. a tuple (x,y,h,s,v) derived from a pixel of o. In both cases, the signature captures the shades of the green pastures surrounding the horses, the brown foal and the gray horse by according clusters. As a conventional histogram cannot tune its partitioning to any single image, it would require an exceptionally large number of partitions to capture the features with a comparable quality [6].

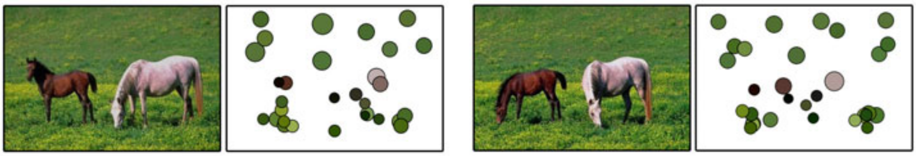


Fig. 1. Two similar images and visualizations of according signatures

While a signature representation can capture feature distributions very precisely, it also requires the similarity measure to be able to compare objects with differing partitioning. An established measure flexible enough to handle feature signatures is the EMD.

The Earth Mover's Distance. The EMD is a similarity measure that is modeled as the solution to a transportation problem based on a ground distance. Given a number of sources with goods that are to be distributed to a number of targets, EMD calculates the most cost-efficient way of distributing the goods. The cost for transporting one unit of the goods from a given source to a given target is based on a distance function that is referred to as the ground distance. For the signature feature representation, the ground distance is computed between two cluster centers and thus describes how dissimilar the features from those regions are. Note that this dissimilarity notion is restricted to the feature space FS and only extends to the dissimilarity of the objects via the EMD.

Definition 2. ***Earth Mover's Distance.*** *Given signatures s^o, s^q and a ground distance* gd*, the EMD between s^o, s^q is defined as a minimum over feasible transports $T = [t_{ij}] \in \mathbb{R}^{|s^o| \times |s^q|}$:*

$$EMD_{gd}(s^o, s^q) = \min_{T \in FEASIBLE} \left\{ \frac{1}{\widetilde{w}} \sum_i \sum_j t_{ij} \cdot \mathrm{gd}(p^o{}_i, p^q{}_j) \right\}$$

where $\widetilde{w} = \min(\sum_i w^o{}_i, \sum_j w^q{}_j)$ is the smaller of the total weights of s^o and s^q and $FEASIBLE \subset \mathbb{R}^{|s^o| \times |s^q|}$ is the set of feasible transports which is defined as $\{T' \in \mathbb{R}^{|s^o| \times |s^q|} \mid CSource \wedge CTarget \wedge CPos \wedge CTransport\}$ with

$$\begin{array}{rclrcl} CSource & \equiv & \forall i : \sum_j t_{ij} \leq w^o{}_i & CTarget & \equiv & \forall j : \sum_i t_{ij} \leq w^q{}_j \\ CPos & \equiv & \forall i,j : t_{ij} \geq 0 & CTransport & \equiv & \sum_i \sum_j t_{ij} = \widetilde{w} \end{array}$$

The EMD is the solution of a linear optimization problem and can be computed efficiently via a simplex algorithm for transportation problems. On the intuitive level, signature s^o is the collection of sources while s^q is the collection of destinations. All transports t_{ij} from sources to targets have to be positive (CPos) and can neither exceed the available goods at the source (CSource) nor the capacity at the targets (CTarget). In addition, as many goods as possible have to be transported (CTransport).

A common choice for the ground distance gd is the Euclidean distance between cluster/partition representatives. The possibility to replace gd and thus to adapt the EMD as a whole is key to our RF techniques presented in Sec. 4.

4 Relevance Feedback for the Earth Mover's Distance

With the similarity model described, we introduce our RF process for the EMD.

4.1 The Feedback Process

Figure 2 gives pseudo-code for the RF process that is the basis for the remainder of the paper. Given an initial query object q and a default distance function $dist$, an RF session starts with a k-nearest-neighbor query for q in DB. After returning k objects to the user (e.g., in a graphical user interface), the process waits for feedback from the user. Unless the user is already satisfied with the results the main feedback loop is entered. Within the feedback loop, the user is asked to let the system know which of the k returned objects are to be considered relevant. Using this information and possibly also relevancy information from past iterations, the system tries to find a new query and a new distance measure that better fit the user's requirements. Lastly, a new set of k objects similar to the new query according to the adapted distance measure is retrieved.

Algorithm `FEEDBACKLOOP` acts as a general framework where `ADAPT_QUERY` and `ADAPT_DIST` can be flexibly exchanged. For our EMD RF process, $dist$ acts as the ground distance for the EMD. Consequently, we concentrate on adapting the EMD by adapting its ground distance. For this purpose, Section 4.3 describes three instances of `ADAPT_DIST` (a baseline, a heuristic, and an optimization-based algorithm). Section 4.2 shows how the RF loop can generally be extended with an optimization-based step that improves on the result of heuristic functions `ADAPT_DIST`.

```
FEEDBACKLOOP(DB, q, k, dist, ADAPT_QUERY, ADAPT_DIST)
    iter = 0; feedback[]  = ∅;
    results = knnQuery(q, dist, DB, k);
    while (isUserSatisfied(results) == false)
        feedback[iter].relevant = getRelUserFeedback(results);
        feedback[iter].rest = results - feedback[iter].relevant;
        q    = ADAPT_QUERY(feedback, iter);
        dist = ADAPT_DIST(q, feedback, iter);
        results = knnQuery(q, dist, DB, k);
        iter = iter + 1;
END;
```

Fig. 2. Relevance Feedback algorithm

4.2 Optimization of the Similarity Measure

While a well-founded heuristic for `ADAPT_DIST` may be able to produce distance functions that enable the retrieval of more relevant database objects than is possible without adapting the distance, it does not necessarily reflect the user's feedback as well as could be wished for. Instead of directly returning the results as determined by the heuristic distance, we propose to first test if the distance can be improved upon in an extension to the traditional feedback loop. We phrase the adaptation of the distance as an optimization problem that ties the distance as the optimization variable to the consistency of the similarity model with the user feedback as the optimization criterion. Our approach is conceptually related to [7], where a greedy optimization was performed in order to find an improved query representation via changing term weights in text retrieval. Unlike [7], we optimize the similarity measure itself instead of the query representation. Optimizing the similarity measure as presented in this section is tied to the EMD in section 4.3 - Similarity Optimization. However, the idea can be transferred to other flexible distance measures such as Quadratic Form Distance for which the orientation and the extend of the ellipsoid-shaped equi-distance surface would be subject to the optimization.

Adaptation as an Optimization Problem. The main question when phrasing the adaptation of a similarity measure as an optimization problem is how to define a suitable optimization criterion. The optimization process cannot know if including an object from the database in the result set will increase the number of relevant objects returned to the user. It is not possible to adapt the similarity measure, compute the k nearest neighbors normally returned to the user, and check if the adaptation was favorable. However, it is possible to retrospectively test if the adaptation resulted in a similarity measure that is consistent with the user's feedback. To this end, the objects from all previous result sets are ranked according to the adapted similarity measure. A good similarity measure results in a ranking of the feedback where the objects already identified as relevant appear before the others. An unsuitable measure has all objects not marked relevant before the relevant ones. To automatically decide how beneficial a given ranking is, a quality measure reducing the ranking to a single figure is required.

Definition 3. ***Average Precision.*** *Given a database DB, relevant objects $\Re \subseteq DB$, and an injective ranking function* $\text{rank} : DB \rightarrow \mathbb{N}$, *the average precision is*

$$\text{avgPrecision}(\text{rank}, \Re) = \frac{1}{|\Re|} \sum_{o \in \Re} \frac{|\{\hat{o} \in \Re | \text{rank}(\hat{o}) \leq \text{rank}(o)\}|}{\text{rank}(o)}$$

Intuitively, a higher number of relevant objects toward the front of the ranking results in a higher average precision value which allows us to formulate the similarity measure adaptation as an optimization. A ranking with relevant objects at positions 1, 3, 4, and 6 has $avgPrecision = \frac{1}{4}(\frac{1}{1} + \frac{2}{3} + \frac{3}{4} + \frac{4}{6}) \approx 0.77$ while a perfect ranking (positions 1,2,3,4) has $avgPrecision = \frac{1}{4}(\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{4}{4}) = 1$.

Definition 4. ***Optimal Relevancy-Consistent Similarity Measure*** *Given query q, database DB, and relevant objects $\Re \subseteq DB$, the optimal relevancy-consistent similarity measure Sim^* according to the average precision value is*

$$Sim^* = \arg\max_{Sim}\{\text{avgPrecision}(\text{rank}^{Sim}, \Re)\}$$

where rank^{Sim} *ranks objects in DB by similarity to q according to measure Sim.*

```
OPTIMIZE(C, q, feedback)
    extern maxIter, coolFactor, $T^0$;
    optIter = 0; T = $T^0$;
    avgpNew = averagePrecision(C, q, feedback);
    avgpOld = avgpNew; avgpBest = avgpNew; Cbest = C;
    while (optIter < maxIter)
        C′ = MODIFY(C);
        avgpNew = averagePrecision(C′, q, feedback);
        if (avgpNew ≥ avgpOld)
            if (avgpNew > avgpBest)
                Cbest = C′; avgpBest = avgpNew;
            C = C′; avgpOld = avgpNew;
        else
            if (rand() < exp((avgpNew − avgpOld) / T))
                C = C′;  avgpOld = avgpNew;
        T = T * coolFactor; optIter = optIter +1;
    return Cbest;
END;
```

Fig. 3. Similarity measure optimization

During the Relevance Feedback process, we evaluate Definition 4 on the subset of the database for which the user has given relevancy information to the system either during the current or during prior feedback loop iterations. By optimizing the average precision on the objects fed back to the system, the final similarity measure better separates known relevant objects from other objects.

The Optimization Algorithm. The influence that the similarity measure parameters have on the value of the optimization criterion is inherently intricate due to the ranking computation required to calculate the value of the criterion. In the case of the EMD, where the optimization parameter is the ground distance function, an analytical optimization is also made infeasible by the algorithmic optimization of multiple transportation problems necessary to compute the EMD values for the ranking. However, there is a well-suited subclass of optimization algorithms for this situation.

The algorithm given in Figure 3 is an instance of the family of probabilistic optimization algorithms referred to as Simulated Annealing algorithms [20]. The main idea is to start the optimization at a given point in the search space (i.e., parameter collection C that defines the similarity measure) and randomly navigate through the search space (`MODIFY`) while evaluating the optimization criterion (`averagePrecision`) on its way through the search space. Unlike greedy algorithms, worse solutions can temporarily be adopted with a certain probability in order to overcome local optimums. The worse a solution is, the less likely is its adoption. In addition, the probability for adopting worse solutions also decreases over time, as the so-called temperature of the optimization process (`T`) tends toward zero. The optimization terminates after a set number (`maxIter`) of iterations. In this fashion, the algorithm first moves rather erratically while looking for regions with good solutions and avoiding local optimums. The average precision of the currently adopted solution (avgpOld) may at first decrease in order to get out of local optimums. As it progresses, it converges toward a greedy algorithm. At the end, the best solution found during the optimization process is used.

We list choices of the annealing parameters coolFactor, maxIter, and T^0 that proved suitable for the EMD Relevance Feedback process in the evaluation section. The MODIFY function depends on the variability of the similarity measure. EMD-specific definitions are given later on. However, the extended feedback loop presented in this section can readily be utilized to optimize other adaptable distance functions by defining a suitable parameter collection C and modification function MODIFY.

4.3 EMD-Based Relevance Feedback on Signatures

In this section, we exploit the flexibility of the EMD for Relevance Feedback and detail suitable instances of ADAPT_QUERY and ADAPT_DIST. The only necessary change to the general FEEDBACKLOOP framework is that knnQuery now computes nearest neighbors by treating the parameter *dist* as the EMD ground distance.

Query Adaptation. There exist several variants for combining several signatures into a new signature in the literature based on region reweighting [9] and clustering [6]. Since our focus is on adapting the similarity measure, we chose to use the approach from [6] with a simple k-means algorithm and to skip the according implementation details of ADAPT_QUERY. The main idea is to collect all signature components from all relevant signatures, recluster their representatives, and set the new weights to the median weight of the components in the clusters. Clusters with contributions from less than half of the relevant signatures are removed, and a normalization of the weights ensures that the total weight is not greater than 1. Together with a fixed Euclidean ground distance produced by ADAPT_DIST_QM in Figure 4, this reflects the "Query-by-Refinement" algorithm from [6] and serves as a baseline for our EMD-based adaptation algorithms on feature signatures in the experiments.

Heuristic for Adapting the Ground Distance. In this section, we propose to exploit the variance of the partitioning representatives from relevant signatures in order to define a ground distance that is based on Relevance Feedback information. The key to defining a suitable ground distance is the observation that we do not require a ground distance from each point in the feature space FS to each other point in FS. For each iteration, we are only interested in distances from a single query q to objects in the database. Thus, a ground distance that defines distance values from the partitioning representatives of q to arbitrary points in FS fully suffices. The main idea of our heuristic approach is to utilize the local variance information of the feedback around the query representatives. If feedback representatives are within a compact region in a dimension of the feature space around a query representative (i.e., the variance is low), the heuristic assigns a high cost to transports out of this region.

```
ADAPT_DIST_QM(q, feedback, iter)
    return new Dist(L2);
END;
ADAPT_DIST_HEUR(q, feedback, iter)
    if (feedback[iter].relevant.size() <= 1) return new Dist(L2);
    $C^{Heur}$ = COMPUTE_HEUR_MATRIX(q, feedback, iter);
    return new Dist($C^{Heur}$);
END;
ADAPT_DIST_OPT(q, feedback, iter)
    if (feedback[iter].relevant.size() <= 1) return new Dist(L2);
    $C^{Heur}$ = COMPUTE_HEUR_MATRIX(q, feedback, iter);
    $C^{Opt}$ = OPTIMIZE($C^{Heur}$, q, feedback);
    return new Dist($C^{Opt}$);
END;
COMPUTE_HEUR_MATRIX(q, feedback, iter) {
    foreach ( Signature s in feedback[iter].relevant )
        foreach ( Component (p,w) in s)
            mini = arg $\min_{i}${Dist(L2).calc(p, $p_i$) | ($p_i$,$w_i$) in q };
            Subset[mini].add(p);
    foreach ( Component ($p_i$, $w_i$) in q )
        $C^{Heur}$.row[i] = inverseVarianceVector(Subset[i]);
    foreach ( row in $C^{Heur}$)
        row = row / sum(row);
    return $C^{Heur}$;
END;
```

Fig. 4. Signature ground distance adaptation

The function ADAPT_DIST_HEUR in Figure 4 takes all signatures of objects deemed relevant by the user in the current feedback iteration and assigns each representative to its closest query representative. The array of sets Subset stores in Subset[i] the set of feedback representatives closest to the i^{th} query representative p_i. The variance information of Subset[i] is stored in the i-th row of a matrix $C^{Heur} = [c_{ij}^{Heur}]$, where c_{ij}^{Heur} is the inverted variance of Subset[i] in the j-th dimension of the feature space FS. Thus, C^{Heur} is of size $|s^q| \times \tilde{d}$ where $|s^q|$ is the number of representatives in the query signature and $\tilde{d}$ is the dimensionality of the underlying feature space FS. The resulting ground distance between a query representative p_i and point $p \in FS$ is a weighted Euclidean distance function. With the inverted variance information from the i-th row of C^{Heur} as the weights on the diagonal of matrix $diag(C_i^{Heur})$, it is computed as

$$\mathrm{gd}(p_i, p) := \sqrt{(p_i - p) \cdot diag(C_i^{Heur}) \cdot (p_i - p)^T}.$$

Fig. 5 shows an example for the resulting ground distance. For three query representatives p_1, p_2, p_3 the corresponding equi-distance lines are shown. A large variance value for the feedback representatives around a query representative results in a low weight for the Euclidean distance assigned to that query representative. While the visualization only shows the equi-distance lines for spatial

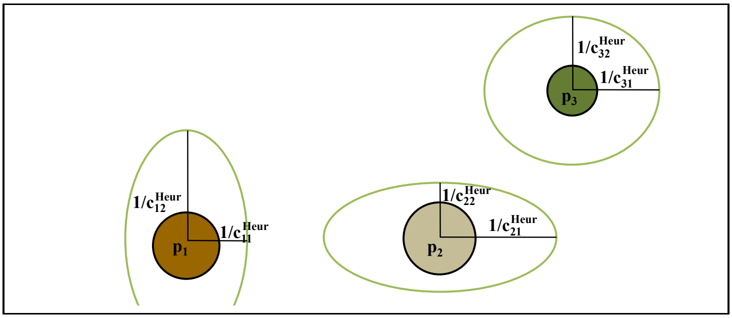


Fig. 5. An adapted EMD ground distance

```
MODIFY(C)
    extern modRowWeight;
    C′ = C;
    foreach (row in C′)
        delta = sum(row) * modRowWeight;
        (entries1, entries2)  = randomPartitioning(row);
        sum1 = sum(entries1); sum2 = sum(entries2);
        if (sum2 <= delta) delta = -delta;
        foreach (entry in entries1)
            entry = entry * (sum1 + delta) / sum1;
        foreach (entry in entries2)
            entry = entry * (sum2 - delta) / sum2;
    return C′;
END;
```

Fig. 6. Modification of the ground distance

dimensions, the same applies to the other dimensions of FS (e.g., color/texture). The resulting ground distance successfully adapts the EMD to the user's feedback as is shown in the evaluation Section 5.

Similarity Optimization. In order to apply the optimization-based similarity adaptation from Section 4.2 to the EMD on feature signatures in `ADAPT_DIST_OPT`, we need an initial parameter collection C that defines the similarity measure and a way of randomly modifying the parameter collection. The parameter that defines the EMD is its ground distance. We can make use of the heuristically determined cost matrix as an initialization for `OPTIMIZE` where the ground distance is fully determined by the variance information stored in the matrix C^{Heur}.

What remains is a function for randomly modifying the cost matrix. Algorithm `MODIFY` in Fig. 6 modifies a matrix according to the externally set parameter `modRowWeight` that determines the modification magnitude. In the cost matrix passed to the algorithm, each row reflects variance information utilized to define weights for a weighted Euclidean distance. The updating algorithm `MODIFY` thus adapts these weights for each individual query representative when it modifies a row of C^{Heur}. For each row, `modRowWeight` determines how much the row is to be changed. To update a row, its entries are randomly partitioned into two sets of equal cardinality (± 1). The entries of the first partition are increased by `modRowWeight` percent of the sum of row costs while the entries of the second partition are decreased accordingly. Should a decrease of the second partition lead to negative costs, the roles of the two partitions are switched. One of the two modifications is always possible with `modRowWeight` from $[0, 0.5)$. Lastly, the new cost matrix is returned to the optimization algorithm for evaluation via the average precision measure and possibly further iterative modification.

The modification translates to an adaptation of the dimension-wise ellipsoid extents in the example of Figure 5.

4.4 Efficient Query Processing

While the Earth Mover's Distance can be computed efficiently using a specialized transportation simplex algorithm, it is worthwhile to consider options to speed up both the k-nearest-neighbor computation that is part of the Feedback Loop algorithm (cf. Fig. 2) and the optimization algorithm (cf. Fig. 3) in order to achieve interactive query processing times. For the k-nearest-neighbor computation within a multimedia database, multi-step retrieval techniques from [13] are well-suited to cope with the changing nature of the transportation cost matrix in our feedback approach. In particular, the lower bound LB_{IM} which is based on a constraint relaxation of the EMD showed a very good selectivity and good computation times in our evaluation prototype. Due to its lower-bounding property, this filter-and-refine retrieval process returns results fast without loss of quality. A second quality-preserving improvement can be achieved for todays multi-core systems by computing mutually independent EMDs in parallel.

Further speed-up is possible for our approach by trading quality for efficiency. For the optimization process, the maximum iteration count can be decreased resulting in a less optimized similarity model. Additionally, an approximate measure such as LB_{IM} can be used to replace the more expensive EMD computations within the target function of the optimization.

Using only the quality-preserving techniques to speed up the retrieval, typical query processing times for our C++ prototype on a dual Intel E5420 computer with 2.5GHz were around 0.6 seconds (down from 4.2 seconds) for the query movement and the heuristic approach and around 1.2 seconds (down from 8.4 seconds) for the optimization-based approach on the databases described below.

5 Experiments

We performed a series of automated RF runs on two databases to evaluate the effectiveness of both the EMD adaptation via our heuristic and via our optimization approach compared with methods that rely on query movement alone.

Fig. 7. Sample query images for PHOTO (top row) and ALOI DB (bottom row)

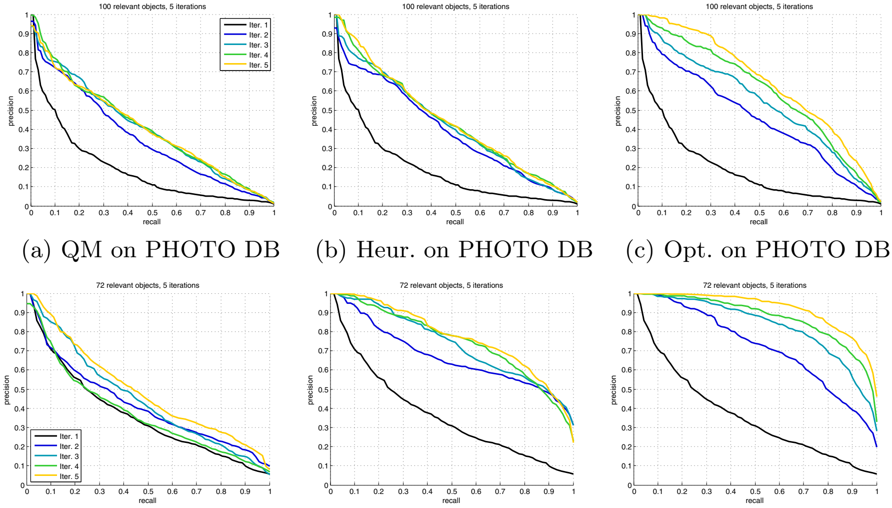
Databases and Evaluation Setup. The first database (PHOTO) includes 59,896 color images from a wide variety of themes (a.k.a. "Corel DB"). Each theme includes ~ 100 images. The second database [21] (ALOI) includes 72,000 images of 1,000 objects that were rotated around the physical y-axis. The variation per object is much smaller than the variation per theme in the PHOTO database. Signatures with up to 20 components were created via a clustering of a 7-dimensional feature space (position, color, and two texture dimensions). For both databases, 20 themes / objects were chosen as relevant images for 20 feedback runs. The choice was purely random for the ALOI database. Given the low-level feature extraction process detailed above, PHOTO themes like "recreation" and "sports" exhibit a large visual diversity within the themes and significant overlap among the themes. They are thus not suitable for our automated evaluation method. We chose twenty themes with limited overlap while excluding themes with hardly any feature diversity. Figure 7 shows a sample of the query images. The diversity per relevant PHOTO theme is still vastly greater than the diversity per ALOI object.

The number k of nearest neighbors was set to equal the number of relevant images such that the optimal result could potentially be attained. The theme/object information for the k images was used to generate feedback for the next iteration. After each iteration, the precision-recall data was collected by ranking the database according to the current EMD. The precision-recall curves are averaged over the 20 queries as described in [22].

The starting temperature T^0 of the optimization process was set to the largest possible difference in average precision ($\frac{m}{n} \cdot \sum_{i=m+1}^{m+n} \frac{1}{i}$ with n as the number of relevant objects and m as the number of objects not labeled as relevant). The number of iterations was limited at $maxIter = 500$ while the modification magnitude was set to $modRowWeight = 5\%$. A value of $coolFactor = 0.85$ showed to be suitable for the decline of the annealing temperature.

EMD Adaptation Results. Figure 8 gives information on full rankings of the databases. The baseline algorithm in (a) and (d) with a fixed, Euclidean EMD ground distance (cf. Section 4.3 - Query Adaptation) is contrasted with our heuristic adaptation in (b) and (e) and the optimization-based adaptation in (c) and (f). Figure 10 shows the improvement limited to the subset of the database that is returned to the user in the feedback iterations.

The heuristic approach gives consistently better results than the query movement approach on the PHOTO signatures ($\sim$ 7.5% according to Fig. 10 (a)). While the baseline algorithm is not able to improve its results after iteration 3, the optimization-based adaptation improves with every iteration, leading to



(a) QM on PHOTO DB (b) Heur. on PHOTO DB (c) Opt. on PHOTO DB

(d) QM on ALOI DB (e) Heur. on ALOI DB (f) Opt. on ALOI DB

Fig. 8. Precision-Recall diagrams averaged over 20 runs

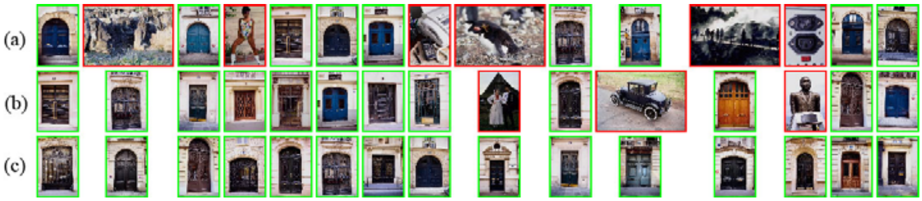
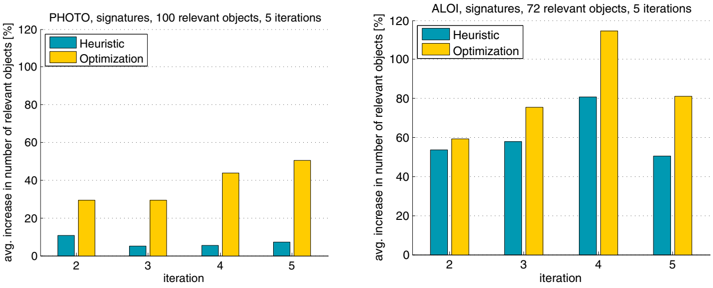


Fig. 9. The first 15 results for the 5^{th} iteration of a feedback session looking for doors in PHOTO. (a) Query Movement (b) Heuristic-based (c) Optimization-based.

$\sim 50\%$ more relevant objects in the result set of the fifths iteration when compared with the query movement approach and nearly three times as many relevant objects compared to its first iteration. Its improvement over the heuristic approach is especially pronounced in those later iterations, when it has collected more relevancy information to utilize. Figure 9 shows an example result for the fifths iteration of a feedback session looking for pictures of doors. While the query movement approach shows a first non-relevant image at position 2, the heuristic has 8 correct hits at the front and the optimization-based approach manages to return only images of doors for the top 15 positions shown. The first non-relevant image only occurs at position 21. The average precision for k=100 is 0.47, 0.67, and 0.88 respectively.

For the ALOI database, the heuristic approach shows a vastly improved performance compared to the PHOTO database. The explanation lies with the homogeneous nature of parts of the features in this database. As Figure 7 depicts, all images in the database exhibit large, black areas around the borders. The compact color and texture subclusters in this database result in signature



(a) PHOTO DB (b) ALOI DB

Fig. 10. Relevant objects among objects returned to user vs. baseline

representatives with almost random values for the non-clustered spatial dimension within the border area, which in turn dominate the standard EMD.

Both our EMD feedback approaches easily overcome this challenge. The statistics-based heuristic leads to low costs for transports in feature space dimensions with high variances. In this way, the adaptation makes the EMD largely ignore the random spatial dimensions for representatives in the border area. These properties of our heuristic for signatures lead to substantial improvements regarding the query results as depicted in Figures 8 (e) and 10 (b), where even the first iteration shows significantly better precision values than any of the baseline iterations and relative improvements of up to 80% compared with the baseline for the number of relevant objects returned per iteration. Our optimization-based RF technique still exceeds these effectiveness gains as it takes objects deemed relevant as well as objects not deemed relevant into account when minimizing the discrepancy between the feedback and the similarity measure defined via the adapted EMD. Figure 10 (b) shows that the improvements depicted in the precision-recall graph of Figure 8 (f) translate to as many as twice the number of relevant objects being returned compared with the query movement algorithm.

6 Conclusion

We proposed a statistics-based heuristic and an optimization-based extension which allow the user to retrieve significantly more relevant objects from the database in fewer iterations compared to the existing EMD-based query movement technique. We have shown how the flexibility of the EMD as a high-quality similarity measure can be exploited to explore multimedia databases via a Relevance Feedback process. Through the utilization of relevance information, the ground-distance-based EMD is adapted to effectively reflect the user's notion of similarity when searching a multimedia database.

References

1. Wichterich, M., Beecks, C., Sundermeyer, M., Seidl, T.: Exploring Multimedia Databases via Optimization-Based Relevance Feedback and the Earth Mover's Distance. In: CIKM (2009)

2. Rocchio, J.: Relevance Feedback in Information Retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing (1971)
3. Ishikawa, Y., Subramanya, R., Faloutsos, C.: MindReader: Querying Databases Through Multiple Examples. In: VLDB (1998)
4. Ortega-Binderberger, M., Mehrotra, S.: Relevance Feedback Techniques in the MARS Image Retrieval System. Multimedia Systems 9(6) (2004)
5. Ferreira, C.D., da S. Torres, R., Gonçalves, M.A., Fan, W.: Image Retrieval with Relevance Feedback Based on Genetic Programming. In: Braz. Symp. DB (2008)
6. Rubner, Y., Tomasi, C.: Perceptual Metrics for Image Database Navigation. Kluwer Academic, Dordrecht (2001)
7. Buckley, C., Salton, G.: Optimization of Relevance Feedback Weights. In: SIGIR (1995)
8. Cord, M., Philipp-Foliguet, S., Gosselin, P.H., Fournier, J.: Interactive Exploration for Image Retrieval. EURASIP J. on Applied Signal Proc. 2005 (2005)
9. Jing, F., Li, M., Zhang, L., Zhang, H.J., Zhang, B.: Learning in Region-Based Image Retrieval. In: Bakker, E.M., Lew, M., Huang, T.S., Sebe, N., Zhou, X.S. (eds.) CIVR 2003. LNCS, vol. 2728. Springer, Heidelberg (2003)
10. Li, C.Y., Hsu, C.T.: Image Retrieval With Relevance Feedback Based on Graph-Theoretic Region Correspondence Estimation. Trans. on Multimedia 10(3) (2008)
11. Typke, R., Veltkamp, R. and Wiering, F.: Searching notated polyphonic music using transportation distances. In: ACM Multimedia (2004)
12. Lavin, Y., Batra, R., Hesselink, L.: Feature Comparisons of Vector Fields Using Earth Mover's Distance. In: IEEE Visualization (1998)
13. Assent, I., Wenning, A., Seidl, T.: Approximation Techniques for Indexing the Earth Mover's Distance in Multimedia Databases. In: ICDE (2006)
14. Ljosa, V., Bhattacharya, A., Singh, A.K.: Indexing spatially sensitive distance measures using multi-resolution lower bounds. In: Ioannidis, Y., Scholl, M.H., Schmidt, J.W., Matthes, F., Hatzopoulos, M., Böhm, K., Kemper, A., Grust, T., Böhm, C. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 865–883. Springer, Heidelberg (2006)
15. Wichterich, M., Assent, I., Kranen, P., Seidl, T.: Efficient EMD-based Similarity Search in Multimedia Databases via Flexible Dimensionality Reduction. In: SIGMOD (2008)
16. Assent, I., Wichterich, M., Meisen, T., Seidl, T.: Efficient Similarity Search Using the Earth Mover's Distance for Large Multimedia Databases. In: ICDE (2008)
17. Indyk, P., Thaper, N.: Fast Image Retrieval via Embeddings. In: Workshop on Statistical and Computational Theories of Vision (2003)
18. Klein, O., Veltkamp, R.C.: Approximation Algorithms for the Earth Mover's Distance Under Transformations Using Reference Points. In: Europ. Workshop on Computational Geometry (2005)
19. Shirdhonkar, S., Jacobs, D.: Approximate Earth Mover's Distance in Linear Time. In: CVPR (2008)
20. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science 220(4598) (1983)
21. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.M.: The Amsterdam Library of Object Images. Int. J. on Computer Vision 61(1) (2005)
22. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
23. Cord, M., Fournier, J., Philipp-Foliguet, S.: Exploration and Search-by-Similarity in CBIR. In: SIBGRAPI (2003)

Prosemantic Features for Content-Based Image Retrieval

Gianluigi Ciocca[1], Claudio Cusano[1], Simone Santini[2], and Raimondo Schettini[1]

[1] Università degli Studi di Milano-Bicocca,
Dipartimento di Informatica Sistemistica e Comunicazione,
viale Sarca 336, 20131 Milano, Italy
[2] Escuela Politécnica Superior,
Universidad Autónoma de Madrid,
C/ Tomas y Valiente 11, 28049 Madrid, Spain

Abstract. We present here, an image description approach based on prosemantic features. The images are represented by a set of low-level features related to their structure and color distribution. Those descriptions are fed to a battery of image classifiers trained to evaluate the membership of the images with respect to a set of 14 overlapping classes. Prosemantic features are obtained by packing together the scores. To verify the effectiveness of the approach, we designed a target search experiment in which both low-level and prosemantic features are embedded into a content-based image retrieval system exploiting relevance feedback. The experiments show that the use of prosemantic features allows for a more successful and quick retrieval of the query images.

1 Introduction

Many content based-retrieval systems have been proposed to manage and retrieve images on the basis of their content. Among the others we can cite [1,2,3,4,5,6]. A survey of some of the most important techniques used in Content-Based Image Retrieval (CBIR) systems can be found in [7]. To overcome the necessity of manually describing the images content, many of these systems are essentially based on low-level image features that are directly and automatically computed from the images themselves. However, the use of low-level features can't overcome the gap between the content and the semantic of the images. In order to cope with this problem and to provide satisfactory retrieval performance, new techniques are introduced in the retrieval process that take into account the subjectivity of human perception. One of these techniques is *relevance feedback* [8]. Relevance feedback is based on the interaction with the user who provides the system with examples of images relevant to the query. The system then refines its result depending on the selected images. The user's feedback provides a way to learn short term and case-specific query semantics. An example of this can be found in [9] where the system learns a non-linear embedding that maps clusters of images into a hidden space of semantic attributes. Long term learning can be achieved by logging the previous user's interactions for further processing [10].

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 87–100, 2011.

Other systems explicitly extract and embed in the retrieval process semantic information about the image content by exploiting automatic classification techniques [11]. These techniques can then be employed to automatically annotate the image content by keywords, which are then used in the retrieval process. If the underlying annotation is reliable, text-based image retrieval can be semantically more meaningful than other indexing approaches [10]. Concept detection techniques categorize images into general concepts such as city, landscape, sunset, forest, sea, etc..., via supervised classification [12,13].

The annotation approaches described above can be considered as crisp annotation: if an image is annotated with a given label then the image expresses that concept or belong to that class. In [14] the authors tested two classification approaches, support vector machines (SVMs) and Bayes point machines (BPMs), to perform a soft image annotation. At the end of the annotation process, each image is annotated with a label vector, and a confidence factor is assigned to each label in the vector. These confidence factors can then be exploited in a text-based search where images are retrieved and ranked according to the confidence factors of the matching labels.

One of the first works that try to bring semantic information under the same model vector paradigm used in query-by-example systems is [15]. Semantic information is learned directly from the image content and forms a vector of semantic weights. Each weight is associated to a concept and is derived from the confidence score obtained by a support vector machine trained to recognize that concept. Retrieval in the semantic space corresponds to performing a similarity comparison between two model vectors using the L_2 measure. A similar approach is followed in [16].

With the exception of a few examples, all the above techniques tackle the problem of semantic image retrieval from the point of view of indexing, viz. they focus on the accuracy of the indexing scheme. Few have been used and evaluated in CBIR systems or tested on large image databases.

One of the first attempts to integrate and compare semantic keyword and low-level features into a single CBIR framework is the SIMPLIcity system [17]. The semantic classification is used to categorize images so that different semantically-adaptive search methods can be applied to each category. The system is also able to narrow down the subset of images to be searched by selecting those in the same category as the query. The reference categories chosen by the author are textured vs. non textured and graph-photograph. A more recent work [14] defines a new paradigm denoted as query-by-semantic-example (QBSE) that combines a query-by-example approach with semantic retrieval. Using the vector model to describe image content, the authors define a vector of semantic multinomial values, where each value is associated to a specific concept. They compared the QBSE and the query-by-visual-examples approaches in a CBIR system within a minimum probability error retrieval framework.

Following a similar paradigm, we designed an approach to CBIR based on the information provided by several image classifiers. One of the main problems in integrating automatic image classification into a content-based retrieval system is the choice of classes. It is very hard to identify a set of categories that are representative of the majority of the pictures and that can be used to reliably approximate their semantics. Moreover, state of the art image classification systems are far from perfect and, consequently, their use in image retrieval requires a high degree of tolerance with respect to misclassification errors.

To circumvent these problems, we did not exploit the classifiers to obtain a "crisp" semantic description of the images (e.g. "sunset on the beach"), but rather to provide a rich description of visual content that correlates low-level features to prototypical scenes (e.g. "image with an edge distribution that can easily be found in seaside scenes"). In our approach, this level of description is provided by a set of *prosemantic* features. These features are obtained by training several image classifiers so designed that their output can be interpreted as membership values of an image in the class that they embody. For each class, we trained multiple classifiers using different low-level features. This choice is not motivated by the need of a more robust classification (which is the most common reason for adopting a multiple classifiers strategy), but because we wanted to exploit the relationship between the classes and the individual features. We let the retrieval system, which is based on a relevance feedback algorithm, to select which features and which classes are appropriate on a case by case basis.

The proposed approach consists of three major steps: first, the images are described by a set of low-level features; then, those descriptions are fed to a battery of image classifiers trained to evaluate the membership of the images with respect to a set of 14 overlapping classes; finally, the output of the classifiers is used to index the images in an image retrieval system, using relevance feedback.

2 Image Description by Low-Level Features

Our aim is to train several classifiers for a set of classes. Therefore, we need a fairly general description of the images in terms of low-level features. We considered four features: two that convey shape information, and two that describe color distribution.

For their simplicity and satisfactory performance, bag-of-features representations have become widely used for image classification and retrieval [18,19,20]. The basic idea is to select a collection of representative patches of the image, compute a visual descriptor for each patch, and use the resulting distribution of descriptors to characterize the whole image. In our work, the patches are the areas surrounding distinctive key-points and are described using the Scale Invariant Feature Transform (SIFT) which is invariant to image scale and rotation, and has been shown to be robust across a substantial range of affine distortions, changes in 3D viewpoint, additions of noise, and changes in illumination [21]. More in detail, we adopted the implementation described in [22] for both key-points detection and description. The SIFT descriptors extracted from an image are then quantized into "visual words", which are defined by clustering a large number of descriptors extracted from a set of training images [23]. The final feature vector is the normalized histogram of the occurrences of the visual words in the image (1096 components).

Statistics about the direction of edges may greatly help in discriminating between images depicting natural and man made objects [24]. To describe the most salient edges we used a 8 bin edge direction histogram: the gradient of the luminance image is computed using Gaussian derivative filters tuned to retain only the major edges. Only the points for which the magnitude of the gradient exceeds a set threshold will contribute to the histogram. The image is subdivided into 8 × 8 blocks, and a histogram for each block is computed (for a total of 512 components).

Spatial color distribution is one of the most widely used feature in image content analysis and categorization. In fact, some classes of images may be characterized in terms of layout of color regions, such as blue sky on top or green grass on bottom. Similarly to Vailaya et al. [12], we divided each image into 9 × 9 blocks and computed the mean and standard deviation of the values of the color channels of the pixels in each block. The LUV color space is used here, since moments in this color space are more discriminant than in other spaces, at least for image retrieval [25]. This feature includes 486 components (six for each block).

Color moments are less useful when the blocks contain heterogeneous color regions. Therefore, a global color histogram has been selected as a second color feature. The RGB color space has been subdivided in 512 bins by a uniform quantization of each component in eight ranges.

3 Image Description by Prosemantic Features

In order to provide a semantically meaningful information about the content of the images, several categories in which images may be automatically classified have been proposed [12,24,26,27,28]. Based on this work, we selected a set of 14 classes: animals, city, close-up, desert, flowers, forest, indoor, mountain, night, people, rural, sea, street, and sunset. Some classes describe the image at a scene level (city, close-up, desert, forest, indoor, mountain, night, rural, sea, street, sunset) other describe the main subject of the picture (animals, flowers, people). The set of classes is not meant to be exhaustive, or to be able to characterize the content of the images with sufficient specificity for our purposes. Our intent, here, was to select a variegated set of concepts proving a wide range of low-level descriptions of typical scenes.

We queried various image search engines on the web with several keywords related to the classes, and downloaded the resulting pictures. Images have been manually inspected in order to remove those which were not relevant to the classes. Low-quality images have also been removed. The final dataset consist of 30084 pictures, divided into 14 sets of more than 2000 images each. For each class, a set of negative examples has been selected by considering pictures of the other classes. Since the classes may overlap, a manual inspection was needed to verify that all the selected images were actually negative examples. Note that this dataset is completely separated from the one we used in the retrieval experiments.

Table 1. Percentage of classification errors of the classifiers on the 14 classes, using the four low-level feature considered (Bag of features (BoF), color histogram in the RGB color space (RGB), color moments in the YUV color space (YUV), and edge direction histograms (EDH)). The errors have been estimated by a five-fold cross validation on the training sets. For each class, the best result is reported in bold.

Class	BoF	RGB	YUV	EDH
Animals	**22.5**	30.0	22.9	25.5
City	**10.1**	20.6	17.1	12.5
Closeup	17.7	27.3	17.2	**15.0**
Desert	18.7	15.7	**14.1**	22.0
Flowers	12.8	**12.0**	12.6	13.3
Forest	**7.0**	13.6	9.8	9.4
Indoor	14.7	18.5	18.3	**12.9**
Mountain	14.1	16.8	**13.7**	20.3
Night	13.5	8.3	**6.6**	27.5
People	**17.0**	23.8	20.2	20.5
Rural	18.5	15.7	**12.2**	22.6
Sea	23.1	21.9	19.4	**16.7**
Street	18.6	24.5	18.8	**17.4**
Sunset	12.5	8.4	**6.6**	16.3
Average	15.8	18.4	**15.0**	18.0

For each combination of low-level feature and class, a Support Vector Machine (SVM) has been trained using the implementation described in [29]. We chose to adopt a Gaussian kernel. There are two parameters that need to be tuned (the cost parameter C and the scale of the Gaussian kernel γ), they have been selected by maximizing the cross validation performance of the resulting classifier (see Table 1). The classification performance varies greatly depending on classes and features, ranging from 6.6% of misclassifications for the "night" class using color moments, to a 30% for the class "animals" using the color histogram. There is not a clearly superior feature and each feature obtained the lowest classification error for at least one class.

Better results can probably be obtained by combining the four scores for each class. However, our goal is not to achieve low misclassifcation rates, but rather to use the classifiers to warp the high-dimensional feature space into a low-dimensional semantic space without losing valuable information about the visual content of the images. Therefore we decided to keep the information about the individual scores obtained with the four features.

In the end, for each class c and for each low-level feature f, a SVM has been trained. Given a new image Q, represented by the feature vector $\mathbf{x}_Q^{(f)}$, the SVM provide a score $s^{(c,f)}$:

$$s^{(c,f)}(\mathbf{x}_Q^{(f)}) = b^{(c,f)} + \sum_{I \in T^{(c)}} \alpha_I^{(c,f)} y_I^{(c)} \exp\left(-\gamma^{(c,f)} \|\mathbf{x}_I^{(f)} - \mathbf{x}_Q^{(f)}\|^2\right), \qquad (1)$$

where $T^{(c)}$ is the training set for class c, $\mathbf{x}_I^{(f)}$ denotes the feature vectors computed on the image I, $y_I^{(c)}$ is the label in $\{-1, +1\}$ which indicates whether I is a positive or a negative example, $b^{(c,f)}$ and $\alpha_I^{(c,f)}$ are the parameters determined by the training procedure, and $\gamma^{(c,f)}$ is the scale parameter of the kernel. The score is expected to be positive when the image belongs to the class c, and negative otherwise. It is well known [30] that the higher the score, the more likely is that the image belongs to the class. Packing together the 56 scores we obtain a compact vector of prosemantic features.

4 The QuickLook^2 CBIR System

We choose to test the prosemantic features within the framework of the QuickLook^2 content based retrieval system [5] which easily allows the incorporation and testing of different numerical image representations. The system adopts low-level pictorial features coupled with a relevance feedback mechanism.

With QuickLook^2, an image database can be queried with the aid of sample images, or user-made sketches, and/or textual image descriptions. When a query is submitted to the system, the retrieved items are presented in decreasing order of relevance, the user is then allowed to progressively refine the system's response by indicating their relevance, or non-relevance. A query refinement mechanism and a relevance feedback algorithm are used to define the new query representing the user needs and to modify the metric used in the retrieval process respectively. For the purpose of this test we use only the low-level pictorial features retrieval capabilities of the system while discarding the textual retrieval functionalities.

Let $\mathbf{x}_I$ be the representation of the image I. Images can be described by different features so $\mathbf{x}_I$ is composed of different numerical vectors, each one representing an image characteristic (e.g. color histogram, shape, etc...). We indicate these vectors for image I as $\mathbf{x}_I^{(1)}, \mathbf{x}_I^{(2)}, \ldots, \mathbf{x}_I^{(p)}$. Given a query Q and a image I, the dissimilarity between the two representations is computed as:

$$D(Q, I) = \frac{1}{p} \sum_{f=1}^{p} D^{(f)}(\mathbf{x}_Q^{(f)}, \mathbf{x}_I^{(f)}) w^{(f)}, \tag{2}$$

where $D^{(f)}$ and $w^{(f)}$ are the dissimilarity metric and the weight associated to the feature f respectively. The weights $w^{(f)}$ allow to tune the contribution of each features in the overall similarity measure. According to the images selected by the user, the weights are determined by the relevance feedback algorithm while the query Q is computed by the query refinement algorithm. The dissimilarities are computed between the query and each image in the database. Images are sorted and presented to the user by increasing dissimilarity.

4.1 Relevance Feedback

The QuickLook^2 system uses a relevance feedback mechanism to update the weights of the dissimilarity function. The key concept of the relevance feedback mechanism, is that the statistical analysis of the image feature distributions the user has judged relevant, or not relevant, can be used to determine what features the user has taken into account (and to what extent) in formulating this judgment, and then accentuate the influence of these features in the overall evaluation of image similarity, as well as in the formulation of a new query. The structure of the relevance feedback mechanism is entirely description-independent, that is, the index can be modified, or extended to include other features without requiring any change in the algorithm as long as the features can be expressed as numerical vectors. The relevance feedback algorithm works as follows: let R_+ the set of relevant images and R_- the set of non relevant images. The feature weights are computed as:

$$w^{(f)} = \begin{cases} \frac{1}{\epsilon} & \text{if } \|R_+\| < 3 \\ \frac{1}{\epsilon+\mu_+^{(f)}} & \text{if } \|R_+\| \geq 3 \text{ and } \|R_-\| = 0 \\ \frac{1}{\epsilon+\mu_+^{(f)}} - \alpha\frac{1}{\epsilon+\mu_*^{(f)}} & \text{otherwise} \end{cases}, \tag{3}$$

where ϵ and α are positive constants, $\mu_+^{(f)}$ is the average of the dissimilarities computed on the $f-$th feature between each pair of images in R_+, and $\mu_*^{(f)}$ the average of the dissimilarities computed on the $f-$th feature between each image in R_+ and each image in R_-. If a weight is negative it is set to 0. A weight is large if the corresponding feature is present in all the relevant images while it is small or dampened if the corresponding feature is variable within the relevant image or is also present in the non relevant images respectively.

4.2 Query Refinement

In content-based retrieval images are sometimes considered relevant because they resemble the query image in just some limited low-level features. Consequently, after an initial query, a given retrieved image may be selected by the user as relevant because it has one of the characteristics of the query (e.g. the same color), and another be selected for another characteristics (e.g. the shape), although the two are actually quite different from each other. To cope with this problem, QuickLook^2 adopts a new method, called query refinement, for computing the query vector. On the basis of the images selected by the user, the system formulates a new query that better represents the images of interest to the user, taking into account the features of the relevant images, without allowing any one particular feature value to bias the query computation. Let $\mathbf{x}_I^{(f)}(k)$ be the $k-$th value of the $f-$th feature of image I. By considering only the images in the relevant set R_+, the query Q is computed as:

$$Y_k^{(f)} = \{\mathbf{x}_I^{(f)}(k)\colon \mid \mathbf{x}_I^{(f)}(k) - \mathbf{x}_{\bar{Q}}^{(f)}(k) \mid \leq 3\sigma_k^{(f)}\}, \tag{4}$$

$$\mathbf{x}_Q^{(f)}(k) = \frac{1}{\|Y_k^{(f)}\|} \sum_{\mathbf{x}_I^{(f)}(k) \in Y_k^{(f)}} \mathbf{x}_I^{(f)}(k), \tag{5}$$

where $\bar{Q}$ is the average query and $\sigma_k^{(f)}$ is the standard deviation of the $k-$th values in the $f-$th feature. The query is thus computed from the feature values that mostly agree while the outliers are removed from the computation.

5 Experiments

A user study has been conducted to evaluate the performance of our prosemantic features against the corresponding low-level ones. For our purpose, we substituted the original features in the QuickLook^2 system with ours and asked 20 subjects to perform ten target search retrieval sessions. All subjects came from the computer science department of the University of Milan - Bicocca: four of them have a background on image processing or computer vision (two Ph.D. students and two post-doctoral fellows), the other 16 are graduate (three) or undergraduate (13) students.

The subjects did the user study one by one on the same desktop with the same instructor. Each subject was constrained to retrieve the target image by selecting any number of relevant and not relevant images within the top 60 retrieved images. They were also allowed to deselect all the previously selected images. Both the search and the deselection accounted as one retrieval operation each and the subjects were instructed that they must retrieve the target image in a maximum of 20 operations without a time limit. During each session the operation performed, the images selected, and the position of the target image within the retrieved results were recorded. In order to minimize user adaptation, the retrieval sessions were conducted alternatively with the low-level features and with the prosemantic features (i.e. one query with the low-level feature and one query with the prosemantic features). For the same reason, each user searched the ten query images in a different order. The subjects were oblivious to what kind of features they were currently using.

The retrieval sessions were organized in such a way that at the end of the user study, each target image was searched half the time by using the low-level and half the time by using the prosemantic features. Before starting each session,

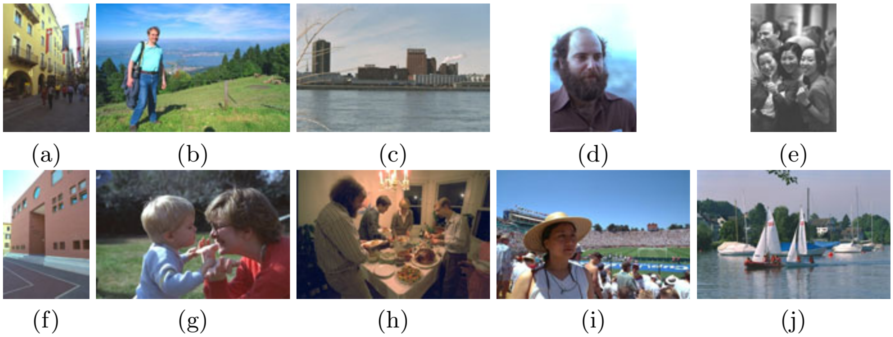


Fig. 1. The ten images used in the target search retrieval sessions

the users have been instructed in the use of the system by performing a guided retrieval test.

The dataset used consists of 1875 images taken from the Benchathlon dataset [31]. The dataset includes typical consumer photographs showing a very different distribution of concepts with respect to the dataset used to train the classifiers. For instance, very often the image would fall in the "people" class, while very few images can be considered as belonging to the "desert" or "flowers" classes. The target images have been randomly selected and are shown in Figure 1. Other 60 images have been randomly selected to compose the page from which the users started all their searches. These images are shown in Figure 2.

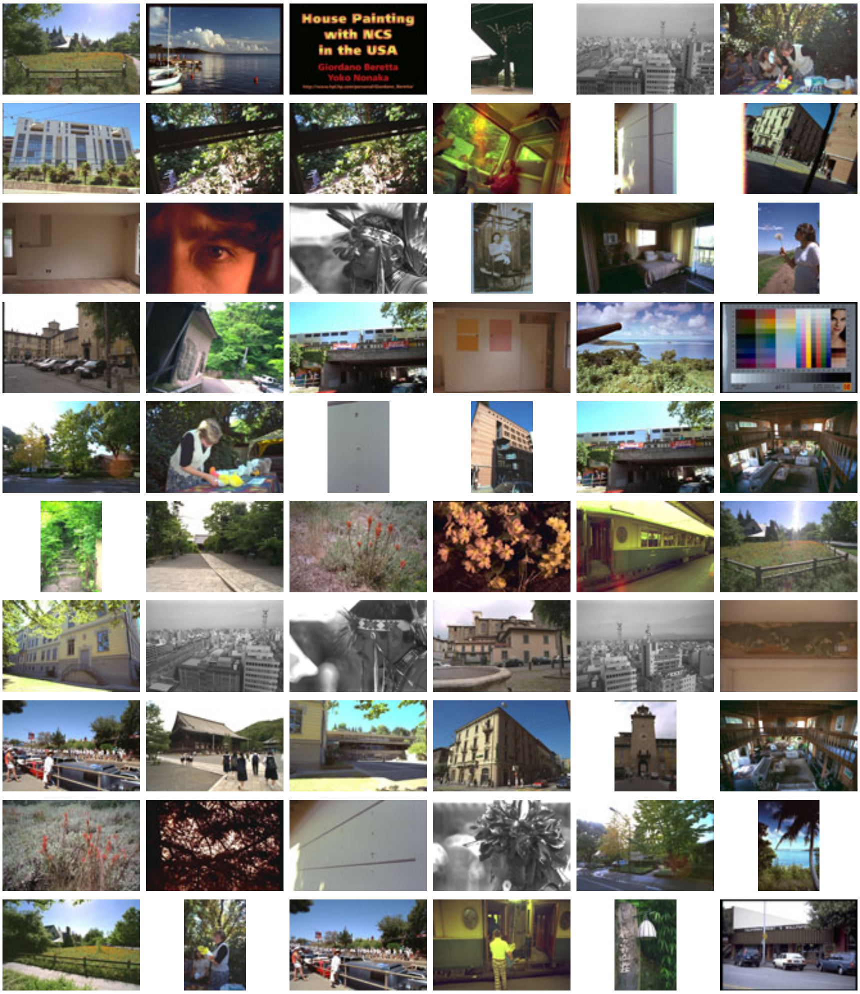


Fig. 2. The 60 images which compose the starting page of the searches

The outcome of the 200 searches clearly demonstrates the effectiveness of prosemantic features with respect to low-level features. Using the prosemantic features, only seven times were the users not able to retrieve the target images within the limit of 20 retrieval operations. By contrast the limit has been exceeded 49 times in the case of low-level features. Figure 3 shows the cumulative success rate for the two sets of features as a function of the number of iterations. The plot shows how prosemantic features allows the retrieval of more target images and with less iterations. In particular, in the case of prosemantic features in more than one third (35/100) of the cases the retrieval of the target image required only one iteration (i.e. without really exploiting the relevance feedback algorithm). Using low-level features this happened only in 11 cases.

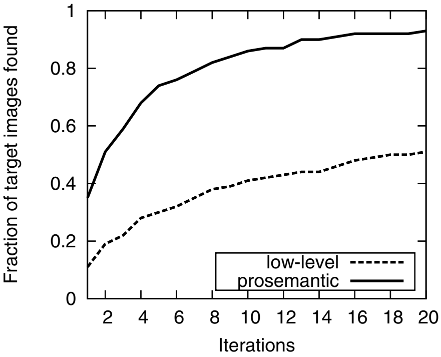


Fig. 3. Fraction of images successfully retrieved as a function of the number of iterations

Since the performance changes significantly for different target images, we reported in Table 2 the results obtained on each of the ten queries. On nine cases out of ten, the use of prosemantic features obtained a higher success rate. The only exception is query (g) which has been quite difficult to find with both the features considered. Two images have never been found using low-level features (d and f), while they have been considered among the easiest to find using prosemantic features. There are two cases (queries e and h) which present clearly distinguishable visual characteristics (one is a grayscale image, the other presents a strong color cast). This fact has been recognized by the majority of users which exploited it to quickly find the targets using low-level features; however, the few users who have not been able to master how low-level similarity works failed the retrieval task. In these two cases retrieval with prosemantic features required (on average) a higher number of iterations, but with only one failure.

Observing the users and discussing with them after the experiment, we made the hypothesis that the effectiveness of the prosemantic features derives from their capability of encoding characteristics of the images which allow a better match against users' intuition about the similarity of the images. Very often, the users started by selecting pictures with the same "general theme" of the target image (e.g. pictures of people, city shots, . . .). Conversely, reasoning about

Table 2. Detail of the results obtained on the ten query images using the two sets of features considered. For each query image are reported the number of successful searches (over 10 attempts for each feature set), the number of iterations needed to retrieve the image (averaged over the successful searches), and the corresponding standard deviation.

			Successful	Iterations	
Query Image		Features	searches	Average	Std deviation
(a)		low-level	8	9.75	5.49
		prosemantic	10	6.80	4.21
(b)		low-level	5	4.20	4.35
		prosemantic	9	4.00	2.11
(c)		low-level	6	3.67	5.09
		prosemantic	9	1.11	0.31
(d)		low-level	0	-	-
		prosemantic	9	3.33	1.70
(e)		low-level	7	1.29	0.45
		prosemantic	10	3.80	3.16
(f)		low-level	0	-	-
		prosemantic	10	1.30	0.64
(g)		low-level	9	7.78	4.39
		prosemantic	7	8.00	5.63
(h)		low-level	7	5.29	4.40
		prosemantic	9	8.11	4.56
(i)		low-level	6	7.50	5.41
		prosemantic	10	1.10	0.30
(j)		low-level	3	9.00	5.10
		prosemantic	10	1.80	1.60

low-level features would require specific training. To verify this intuition we considered the variation of the number of successfully retrieved images during the sessions. Therefore, we counted for each feature set how many images has been retrieved among the first two searches of each user. We did the same for the second two searches and so on. . . We considered pairs of searches because the two feature sets have been used alternatively by each subject. The results are shown in Figure 4. Using the prosemantic features performances are very close to the maximum attainable (i.e. 20 successes) straight from the beginning of the retrieval sessions. Therefore, it is not possible to distinguish any user

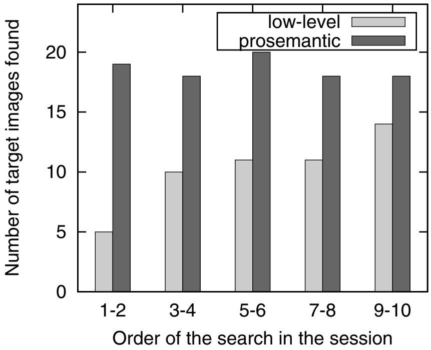


Fig. 4. Number of images successfully retrieved as a function of the order in the sequence of searches

adaption phenomenon. For what concern low-level features, instead, it seems that performance actually increased during the sessions: from only four retrieved images within the first two searches, to 14 within the last two searches. So it is possible that, provided a sufficient amount of training of the user, low-level features may reach the same retrieval performance of prosemantic features.

6 Conclusions

We have presented here, an image description approach based on prosemantic features. These features are obtained by multiple classifiers trained to identify 14 semantic concepts, on the basis of different low-level representations. To verify the effectiveness of the approach, we designed an image retrieval experiments in which low-level and prosemantic features are embedded into a content-based image retrieval system based on relevance feedback. The experiments show that the use of prosemantic features allows for a more successful and quick retrieval of the query images.

To further assess the generalization capabilities of prosemantic features, we plan to extend the experimentation by recruiting more subjects and by considering additional queries. We are also considering to test prosemantic features in other application scenarios such as automatic image annotation and classification.

References

1. Brunelli, R., Mich, O.: Image retrieval by examples. IEEE Transactions on Multimedia 2(3), 164–171 (2000)
2. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: the QBIC system. IEEE Computer 28(9), 23–32 (1995)
3. Gevers, T., Smeulders, A.: PicToSeek: combining color and shape invariant features for image retrieval. IEEE Transactions on Image Processing 9(1), 102–119 (2000)

4. Smith, J.R., Chang, S.F.: VisualSEEk: a fully automated content-based image query system. In: Proceedings of the Fourth ACM International Conference on Multimedia, pp. 87–98 (1996)
5. Ciocca, G., Gagliardi, I., Schettini, R.: Quicklook^2: An integrated multimedia system. Journal of Visual Languages & Computing 12(1), 81–103 (2001)
6. Ahmad, I., Grosky, W.I.: Indexing and retrieval of images by spatial constraints. Journal of Visual Communication and Image Representation 14(3), 291–320 (2003)
7. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
8. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: a comprehensive review. Multimedia Systems 8(6), 536–544 (2003)
9. Lee, C.S., Ma, W.Y., Zhang, H.: Information embedding based on user's relevance feedback for image retrieval. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 3846, pp. 294–304 (1999)
10. Datta, R., Li, J., Wang, J.Z.: Content-based image retrieval: approaches and trends of the new age. In: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 253–262 (2005)
11. Fan, J., Gao, Y., Luo, H., Xu, G.: Automatic image annotation by using concept-sensitive salient objects for image content representation. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 361–368 (2004)
12. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: Image classification for content-based indexing. IEEE Transactions on Image Processing 10(1), 117–130 (2001)
13. Chen, X., Wang, J.Z.: Image categorization by learning and reasoning with regions. Journal of Machine Learningn Research 5, 913–939 (2004)
14. Chang, E., Kingshy, G., Sychay, G., Gang, W.: CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology 13(1), 26–38 (2003)
15. Smith, J.R., Naphade, M., Natsev, A.: Multimedia semantic indexing using model vectors. In: Proceedings of IEEE International Conference on Multimedia and Expo., pp. 445–448 (2003)
16. Lu, J., Ma, S.P., Zhang, M.: Automatic image annotation based on model space. In: Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering., pp. 455–460 (2005)
17. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(9), 947–963 (2001)
18. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)
19. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, vol. 1, pp. 257–264 (2003)
20. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, vol. 2, pp. 1458–1465 (2005)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
22. Vedaldi, A.: Sift++ a lightweight c++ implementation of sift, http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html

23. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2161–2168 (2006)
24. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: city images vs. landscapes. Pattern Recognition 31(12), 1921–1935 (1998)
25. Furht, B.: Content-based image indexing and retrieval. In: Handbook on Multimedia Computing. CRC Press, Inc., Boca Raton (1998)
26. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision 72(2), 133–157 (2007)
27. Schettini, R., Brambilla, C., Cusano, C., Ciocca, G.: Automatic classification of digital photographs based on decision forests. International Journal of Pattern Recognition and Artificial Intelligence 18(5), 819–845 (2004)
28. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 42–51 (1998)
29. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
30. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74 (1999)
31. Gunther, N.J., Beretta, G.: A benchmark for image retrieval using distributed systems over the internet: BIRDS-I. Technical Report HPL-2000-162, HP Labs, Palo Alto (2001)

Enrichment of Named Entities for Image Photo Retrieval*

Rodrigo Agerri[1], Rubén Granados[1], and Ana García Serrano[2]

[1] Universidad Politécnica de Madrid (UPM), Spain
r.agerri@upm.es, rgranados@fi.upm.es
[2] NLP & IR Group UNED, Madrid, Spain
agarcia@lsi.uned.es

Abstract. This paper describes and discusses an approach to extract and exploit enriched Named Entities for Image Photo Retrieval. The enrichment of Named Entities is inspired by the concept of *definite description*. The approach is evaluated using the imageCLEF-08 testset for the photo retrieval task held at Cross-Language Evaluation Forum in 2008. We are particularly interested in testing and discuss whether and how linguistic techniques such as the one presented here can be of benefit for an ad-hoc photo retrieval task. More specifically, results show an improvement in precision when Named Entities are contained in the text although for an overall improvement a better integration of these techniques in a general approach to photo retrieval is needed.

1 Introduction

Named Entity Recognition (NER) is one of the most commonly used low-level linguistic annotations and analysis for Natural Language Processing (NLP) tasks such as Information Retrieval (IR), Question Answering (QA) or Machine Translation (MT). Named Entities (NEs) are viewed as *rigid designators* that uniquely denote an entity [1]. These entities usually include proper names of locations such as Spain, France, Mediterranean Sea, English Channel, Alps, etc., organizations/institutions (Microsoft, IBM, BBC, European Central Bank) and people's names (John McEnroe, King Albert II, Queen Elizabeth, Nixon).

Intuitively, NEs are seen as rigid designators because they uniquely refer to an entity in the world, as opposed to non-rigid designators (NRDE) which merely refer to an entity in a non-unique manner. Classic examples of non-rigid designators are those corresponding to descriptions [2] which are used to describe an entity in the world. For example, "the current British prime minister" is a description of a person who is (at the time that sentence was written) the prime minister of the UK. Nowadays the reference of this description is "Gordon Brown", but that changes with the time at which the description is used,

* This work has been supported by the Spanish R+D National Plan, project TIN2007-67407-C03-03, "BRAVO: Búsqueda de Respuestas Avanzada Multimodal y Multilingüe - Recuperación de Información" and by Madrid R+D Regional Plan, MAVIR Project, S-0505/TIC/000267 http://www.mavir.net

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 101–110, 2011.

namely, during the second part of the Second World War the referent would have been "Winston Churchill". Common non-rigid designators include "New York Times correspondent", "the White House speaker", "the president of the European Commission", "IBM headquarters", etc.

If we were to perform NER to some texts containing such descriptions, only those words in capitals would be extracted, which means that the topic will be taken to be about the newspaper "New York Times" instead of about a *a journalist* of the New York Times; the same applies to "White House" and the *speaker* of the White House, "European Commission" and the *president* of the European Commission and "IBM" and the *headquarters* of IBM. We believe that the inability of extracting the modifiers of NEs would probably have a negative impact in a photo retrieval task. From this point onwards, we will be talking about Named Entities plus modifiers instead of *descriptions* because, as we will see, not all Named Entities plus modifiers are descriptions, even though our approach is inspired by the idea of descriptions as non-rigid designators.

Next section motivates the treatment of NEs plus modifiers for a photo retrieval task; section 2 explains with examples the procedure to extract descriptions. In section 3 we analyze the results and discuss possible ways to improve them, and finally we offer some concluding remarks.

2 Extracting Enriched NEs

During the preparation of the MIRACLE-FI participation [3] in this year ImageCLEF Photo Retrieval task [4], a study of the testset showed great abundance of Named Entities, which it suggested that our approach may benefit from the application of a NER tagger.

The testset consists of five hundred thousand captions describing a photo, and 50 topics [4]. The NEs of the captions were tagged using a pipeline of taggers by C&C tools: tokenizer, Part Of Speech (POS) and NER taggers [5,6]. We focused on NEs referring to *Locations*, *Persons* and *Organization*, given that part of the other four categories which the C&C NER tagger annotates (numerical expressions which form Dates, Times, Percentages and Monetary expressions) are not considered during the indexation process by IDRA, the our indexation tool [7].

An examination of the NER tagger's output showed that we were leaving out the Named Entities' modifiers in expressions (some of them descriptions as characterized in the previous section) such as *Time magazine correspondent*, *prime minister Tony Blair*, *Paris-Roubaix race*, *princess Mathilde*, and *Leonardo da Vinci airport*. In other words, by extracting Named Entities from the photo captions instead of Named Entities plus their modifiers, we were leaving out crucial information to *describe* the photo. Thus, 'Time' would refer to an *Organization* whereas the description 'Time magazine correspondent' would refer to a *Person*; by including the modifier 'correspondent' we obtain a more precise reference to the event captured by a given image. In the case of 'Paris-Roubaix race', extracting the modifier of the NE helps to make explicit that we are referring to the cyclist race and not to the town itself. Furthermore, 'princess Mathilde' as oppose to 'Mathilde' would presumably help to more accurately describe the photographs to which the caption is referring to.

This issue turned out to be a symptom of a more general problem as presented by [8]. For tasks such as IR, it is not sufficient to have low level tools that produce high quality linguistic annotation and analysis, but it is also required that the results of the various levels of annotation be consistent with each other. Inconsistencies between different levels of linguistic annotations means that the information contained in them cannot be combined in an easy and effective manner. In order to solve any inconsistencies between the output of parsing and NER of a given text, [8] propose a joint model of parsing and NER to make correspond a Named Entity with a phrase in a constituent tree and to avoid cases in which a Named Entity span has crossing brackets with any spans of the parse tree. For example, the tree of the noun phrase 'the District of Columbia' may result in separating the Named Entity 'District of Columbia' in two different phrases, namely, one NP phrase containing the proper name 'District' and the prepositional phrase 'of Columbia'.

Our proposal to obtain a better bracketing of Named Entities namely, including their modifiers, consists of exploiting the interaction between the various levels of linguistic annotation provided by the C&C tools [6]. To the pipeline previously used for NER (tokenizer, POS and NER taggers) we add chunking, consisting of segmenting the text in phrases (Noun Phrase, Verbal Phrase, Adverbial Phrase, etc.).

The idea is simply to establish the Named Entities as foci. This means that whenever the C&C tagger annotates a word(s) as a Named Entity, we focus on the Noun Phrase (NP) of which forms part and on those that are contiguous to it, if any. Once this is done, we attach those surrounding terms that act as modifiers of the Named Entity according to their POS and which pertain to the same (or contiguous) NP. Currently the POS categories we are able to deal with are periods and abbreviations, prepositions, adjectives and nouns. For example, our approach correctly brackets the Named Entities in together with their modifiers in expressions such as *Paris-Roubaix race*, *princess Mathilde*, *Leonardo da Vinci international airport*, *District of Columbia* and *Royal Palace of Brussels*. It should be noted that in some cases the enriched expression of which a Named Entity is part denotes a radically different entity, as is the case of 'Leonardo da Vinci' and 'Leonardo da Vinci international airport', and in 'Time' and 'Time magazine correspondent'. Other cases the description simply helps to more accurately describe or identify a given entity, as in 'princess Mathilde' or 'Paris-Roubaix race'.

Let us illustrate our approach using the caption 1470132 of the ImageCLEF-09 dataset [4]:

> "American photojournalist James Nachtwey in a file photograph from May 18 2003 as he is awarded the Dan David prize in Tel Aviv for his out standing contribution to photography. It was announced by Time magazine on Thurs day, 11 December 2003 that Nachtwey was injured

> in Baghdad along with Time magazine senior correspondent Michael Weisskopf when a hand grenade was thrown into a Humvee they were traveling in with the US Army. Both journalists are reported in stable condition and are being evacuated to a US military hospital in Germany."

On the one hand, the Named Entities annotated by the tagger were: American James Nachtwey, Dan David, Tel Aviv, Baghdad, Michael Weisskopf, US Army, US, Germany, and Nachtwey. On the other, the Named Entities and descriptions extracted by our approach were: American photojournalist James Nachtwey, Dan David prize, Tel Aviv, Baghdad, Time magazine senior correspondent Michael Weisskopf, US Army, US military hospital, Germany and Natchtwey. It is particularly noticeable that our system is able to attach 'Time magazine senior correspondent' to 'Michael Weisskopf', that the topic of the caption is about the 'Dan David prize' and a 'US military hospital' in 'Germany', and not about 'Dan David' in the 'US'. These differences are showed in table 1.

Table 1. Comparing NER and NRDE

NER	**NRDE**
American James Nachtwey	American **photojournalist** James Nachtwey
Dan David	Dan David **prize**
Tel Aviv	Tel Aviv
Baghdad	Baghdad
Michael Weisskopf	**Time magazine senior correspondent** Michael Weisskopf
US	US **military hospital**
Germany	Germany

The words in bold font highlight those modifiers that we are able to attach to the Named Entities that act as foci. Let us take a look at the C&C output for each of the three sentences to show the re-bracketing performed to obtain the enriched NEs:

```
American|NNP|I-ORG|I-NP photojournalist|NN|O|I-NP James|NNP|I-PER|I-NP
Nachtwey|NNP|I-PER|I-NP in|IN|O|I-PP a|DT|O|I-NP file|NN|O|I-NP
photograph|NN|O|I-NP from|IN|O|I-PP May|NNP|I-DAT|I-NP 18|CD|I-DAT|I-NP
2003|CD|I-DAT|I-NP as|IN|O|I-SBAR he|PRP|O|I-NP is|VBZ|O|I-VP
awarded|VBN|O|I-VP the|DT|O|I-NP Dan|NNP|I-PER|I-NP David|NNP|I-PER|I-NP
prize|NN|O|I-NP in|IN|O|I-PP Tel|NNP|I-ORG|I-NP Aviv|NNP|I-ORG|I-NP
for|IN|O|I-PP his|PRP\$|O|I-NP out|RP|O|I-NP standing|VBG|O|I-NP
contribution|NN|O|I-NP to|TO|O|I-PP photography|NN|O|I-NP .|.|O|O
```

In this sentence, it is possible to see that the words 'American', 'photojournalist' and 'James Nachtwey' are considered to be three different spans by the tagger, 'American' being tagged as Organization type of NE and 'James Nachtwey' as Person. Our approach is able to join together all four words in the same text span. In the case of ‘Dan David’, we attach ‘prize’ to ‘Dan David’ – tagged as Person – which obviously changes the reference of the expression.

```
It|PRP|O|I-NP was|VBD|O|I-VP announced|VBN|O|I-VP by|IN|O|I-PP
Time|NNP|O|I-NP magazine|NN|O|I-NP on|IN|O|I-PP Thursday|NNP|I-DAT|I-NP
,|,|I-DAT|O 11|CD|I-DAT|I-NP December|NNP|I-DAT|I-NP 2003|CD|I-DAT|I-NP
that|IN|O|I-SBAR Nachtwey|NNP|I-PER|I-NP was|VBD|O|I-VP
injured|VBN|O|I-VP in|IN|O|I-PP Baghdad|NNP|I-LOC|I-NP along|IN|O|I-PP
with|IN|O|B-PP Time|NNP|O|I-NP magazine|NN|O|I-NP senior|JJ|O|I-NP
correspondent|NN|O|I-NP Michael|NNP|I-PER|I-NP Weisskopf|NNP|I-PER|I-NP
when|WRB|O|I-ADVP a|DT|O|I-NP hand|NN|O|I-NP grenade|NN|O|I-NP
was|VBD|O|I-VP thrown|VBN|O|I-VP into|IN|O|I-PP a|DT|O|I-NP
Humvee|NN|O|I-NP they|PRP|O|B-NP were|VBD|O|I-VP traveling|VBG|O|I-VP
in|IN|O|I-PRT with|IN|O|I-PP the|DT|O|I-NP US|NNP|I-ORG|I-NP
Army|NNP|I-ORG|I-NP .|.|O|O
```

The modifiers of the Named Entity ‘Michael Weisskopf’ are detected and attached to it so that we extract the whole expression in the same text span.

```
Both|DT|O|I-NP journalists|NNS|O|I-NP are|VBP|O|I-VP
reported|VBN|O|I-VP in|IN|O|I-PP stable|JJ|O|I-NP condition|NN|O|I-NP
and|CC|O|O are|VBP|O|I-VP being|VBG|O|I-VP evacuated|VBN|O|I-VP
to|TO|O|I-PP a|DT|O|I-NP US|NNP|I-ORG|I-NP military|JJ|O|I-NP
hospital|NN|O|I-NP in|IN|O|I-PP Germany|NNP|I-LOC|I-NP .|.|O|O
```

In this case our re-bracketing is able to capture more accurately the reference from *US* to *US military hospital*.

3 Results and Discussion

The approach describe in the previous section is used to extract both the Named Entities and their modifiers from the captions of the photo retrieval task at the imageCLEF-09 [4]. The 2009 collection (from the Belga News Agency) was organized following the IAPR TC-12 collection used for the ImageCLEF-08 Photo Retrieval task. Thus, an XML document for each of the captions was created. We automatically annotated the collection extracting both Named Entities and Named Entities plus modifiers (NRDEs) into two XML elements with the aim of doing structured indexations and retrieval.

The topics were treated by converting all the NEs and NRDEs to lowercase and use that list to annotated the topics 1 to 25. The same method was also applied to the captions corresponding to the images that formed the clusters in topics 26-50.

Given that at the moment of writing we did not have available yet the imageCLEF-09 query relevance assessments we instead use the IAPR TC-12 collection of the ImageCLEF 08 Photo Retrieval task to evaluate our system. The collection was NER and NRDE annotated as explained above and in the process of doing that we realized that last year’s testset contained little Named

Entity-related information, so that the positive or otherwise impact of our approach was going to be minimized. Therefore, instead of comparing results on the overall set of topics, we will be comparing the performance for those topics that do not contain any Named-Entity information and for those that do.

3.1 IDRA

IDRA is the indexation and retrieval tool used in our experiments [7]. It is being developed by the ISYS-GSI research group at the Universidad Politécnica de Madrid, Spain, in order to have a system capable of supporting experimentation related to information retrieval. The main goal is to have an open-source tool ready for improvement by adding new functionalities to support indexing, retrieval and management of different types of data, according to the task that needs to be carried out. In its current form, IDRA allows indexing a wide range of text and image annotation files, launching queries to the indexed data and retrieving the most relevant results.

IDRA is able to index whole collections of files in a variety of formats. For this paper, we will be indexing the IAPR TC-12 collection consisting of 8 XML elements plus the 2 we add containing the Named Entities and NRDEs extracted using the approach previously described.The indexation and retrieval algorithm is based on the classical Vector Space Model (VSM) approach [9] using TF-IDF weights, complemented with stopwords detection and filtering of punctuation marks, accents, and some special characters. The results retrieved by IDRA for a concrete query will be ranked from high to low depending on the relevance value assigned to each retrieved document. The relevance value is calculated using the cosine similarity measure. IDRA includes functionalities to fuse results lists obtained using different methodologies. For example, IDRA was used for the ImageCLEF-08 and ImageCLEF-09 Photo Retrieval tasks to merge results using text- and content-based results [10,3], obtaining higher Mean Average Precision scores (MAP) the average. It has also been compared to Lucene [7], with similar results for MAP, precision and recall.

3.2 Evaluation

For the experiments, we first parsed the 20 thousand XML documents of the IAPR TC-12 collection (used at the ImageCLEF-08 Photo Retrieval task); we modified the IAPR TC-12 collection by adding two new elements containing the Named Entities and Named Entities plus modifiers (NRDE) extracted using the procedure sketched in section 2. We then used IDRA to do two different indexations:

1. One consisting of the text in the XML elements Title, Description Notes and Location.
2. Another one consisting of the same elements plus the two new elem elements added to the collection containing the Named Entities and NRDEs.

With respect to the topics, we parsed and extract any Named Entities and NRDEs detected and create three different query files:

1. Title, Topic and Narration for each topic.
2. Title, Topic, Narration and any Named Entities extracted.
3. Title, Topic, Narration and any NRDEs extracted (enriched Named Entities extracted).

Three experiments were run. The first run consisted of the indexation 1 and query file 1. This run corresponds to the text baseline run as presented at the ImageCLEF-08 [10]. The other two experiments were devised to test whether the extra information provided by the Named Entities and modifiers extracted had any beneficial consequences for the retrieval task using IDRA. Furthermore, we were interested in testing if there are any gains in extracting the enriched Named Entities (with respect to standard NER).

Given that IDRA's performance at ImageCLEF-08 was only evaluated in terms of precision, and that our primary interest is to discuss and (if possible) determine whether IDRA can be improved by employing enriched Named Entities, we will also restrict the comparison to precision scores. Moreover, as it was previously mentioned, a substantial number of the captions and topics of the ImageCLEF-08 Photo Retrieval collection do not contain any Named Entity expressions. This being the case, the only clear result of an overall evaluation would be that NER is not suitable for such task. However, we can compare the performance for each of the 39 topics distinguishing between those that contain any Named Entities and those that do not. It should be said, however, than the overall scores of the NER/NRDE runs were similar to the text baseline published in [10].

Table 2 shows the results of the three experiments for the 39 topics of the imageCLEF-08 photo retrieval task. Specifically, for each topic, its precision at rank 5 (P5) and at rank 20 (P20) (these, together with Recall 20 and F1 measure was the metric used for the ImageCLEF-08 organizers), the mean average precision (MAP) the number of relevant/retrieved documents and finally whether any Named Entities (standard and enriched) were extracted from the topic. The results fall under three different types:

1. Those cases in which the NER + NRDE treatment improves the text baseline: Topics 6, 10, 11, 23 and 49 (in separated cells).
2. Those that do not change the results: 16, 18, 24, 28, 37, 41, 44, 48 and 54.
3. Those that score worse than the text baseline: 29, 31, 34 and 54 (marked by !)

The first case is easy to explain. Given that we were able to extract enriched Named Entities from these topics, extra weight was given to those terms and therefore precision scored higher than the baseline. For example topic 24 was about showing "sport activities in the US state of California" and we were able to extract 'California' and 'US state of California'.

Table 2. Impact of NER/NRDE on individual queries for ImageCLEF-08 Photo Collection

ImageCLEF-08				ImageCLEF-08 + NER			ImageCLEF-08 + NRDE			
Query	**P5**	**P20**	**map**	**P5**	**P20**	**map**	**P5**	**P20**	**map**	**NER/NRDE**
2	0.0000	0.0000	0.0280				0.0000	0.0000	0.0291	-
3	0.2000	0.2500	0.2368				0.0000	0.3000	0.2161	-
5	0.0000	0.1000	0.0940				0.2000	0.2500	0.1277	-
6	0.2000	0.2500	0.2388	0.6000	0.6500	0.3983	0.6000	0.6500	0.3983	YES
10	0.6000	0.6000	0.5034	0.6000	0.6500	0.4751	0.6000	0.6500	0.4760	YES
11	0.6000	0.3000	0.3021	0.6000	0.4500	0.4200	0.6000	0.4500	0.4200	YES
12	0.4000	0.2000	0.1534				0.2000	0.3500	0.1562	-
13	0.6000	0.3500	0.2251				0.4000	0.3500	0.1831	-
15	0.0000	0.0000	0.0443				0.0000	0.0500	0.0441	-
16	0.0000	0.1500	0.1735	0.0000	0.0500	0.2446	0.0000	0.0500	0.2446	YES
17	0.8000	0.6000	0.5593				1.0000	0.6000	0.6371	-
18	0.0000	0.3000	0.1587	0.0000	0.0000	0.0267	0.0000	0.0000	0.0267	YES
19	0.2000	0.1500	0.0613				0.0000	0.1500	0.0538	-
20	0.2000	0.0500	0.0079				0.2000	0.0500	0.0098	-
21	0.0000	0.0500	0.2511				0.2000	0.2000	0.2788	-
23	0.0000	0.5000	0.1343	1.000	0.6500	0.2194	1.0000	0.6500	0.2194	YES
24	0.0000	0.0500	0.0197	0.0000	0.1000	0.0190	0.0000	0.1000	0.0190	YES
28	0.2000	0.2000	0.0899	0.2000	0.1000	0.0479	0.2000	0.1000	0.0479	YES
! 29	0.8000	0.9000	0.7720	0.6000	0.4000	0.4067	0.6000	0.4500	0.5146	YES
! 31	0.8000	0.5000	0.3272	0.4000	0.2000	0.1438	0.4000	0.2000	0.1438	YES
! 34	0.8000	0.6000	0.3077	0.4000	0.2500	0.1154	0.2000	0.1000	0.1395	YES
35	0.8000	0.8500	0.6108				0.8000	0.9000	0.6166	-
37	0.2000	0.1500	0.0796	0.0000	0.0000	0.0691	0.0000	0.0000	0.0691	YES
39	0.0000	0.2667	0.2308				0.0000	0.3000	0.2401	-
40	0.0000	0.1500	0.0779				0.2000	0.1000	0.0748	-
41	0.0000	0.0000	0.0204	0.0000	0.0500	0.0200	0.0000	0.0500	0.0200	YES
43	0.0000	0.2500	0.1797				0.2000	0.2000	0.1824	-
44	0.0000	0.1000	0.0437	0.0000	0.3000	0.0442	0.0000	0.1000	0.0480	YES
48	0.0000	0.2000	0.1627	0.0000	0.0000	0.0933	0.0000	0.0000	0.0933	YES
49	0.8000	0.5000	0.1505	0.8000	0.6000	0.1609	0.8000	0.6000	0.1609	YES
50	0.6000	0.2500	0.1497				0.4000	0.2000	0.1471	-
52	0.6000	0.3000	0.2459				0.6000	0.3500	0.2617	-
53	0.8000	0.5500	0.3945				0.6000	0.4500	0.3349	-
! 54	0.8000	0.5500	0.6464	0.8000	0.4500	0.5349	1.0000	0.5000	0.5968	YES
55	0.2000	0.0500	0.0126	0.2000	0.0500	0.0126	0.2000	0.0500	0.0126	YES
56	0.0000	0.2000	0.2240				0.0000	0.1500	0.2053	-
58	0.8000	0.4500	0.3599				0.8000	0.4000	0.3242	-
59	0.2000	0.1500	0.1324	0.0000	0.0500	0.0873	0.0000	0.0500	0.0873	YES
60	0.8000	0.7500	0.6268				1.0000	0.7500	0.7166	-

The second case, in which the results of the NER/NRDE runs are similar to the baseline, indicate that a finer treatment of these topics is needed. For example, topic 16 is about showing *images of San Francisco with at least one person*; also images of San Francisco without people are not relevant. As our system extracted just the Named Entity 'San Francisco', that obviously did not have any positive impact on the precision score (as bad as the baseline). The same phenomenon goes on in query 18, in which sport stadia outside Australia were relevant, but for which the NER tagger just picked 'Australia'. It looks like as if our system would benefit from a deeper parsing of the topics (e.g., targeting negation).

The third case is similar to the previous one. However, the results are worse than the baseline because the great number of Named Entities wrongly detected. In topic 31, for example, was about retrieving images of "volcanoes around Quito", providing a rather long list of volcano names. Our system detected these names as Named Entities, but did not pick the term 'volcano'. As those names also correspond to counties in Ecuador, it is reasonable to assume that it created a lot of noise and retrieved many non-relevant images, as the poor results show. Topics 29 and 34 are similar cases.

Finally, even though not many enriched NEs were present in the topics and captions, the NRDE run outperformed the NER run, which suggest that the re-bracketing procedure of NEs has a positive impact.

These results point out to a need for a better integration of linguistic analysis, one in which NER and NRDE treatments are conditional to other techniques. For example, if a Named Entity occurs inside the scope of a negation then that Named Entity should not be included in the query. Nevertheless, it should be noted that for those topics for which Named Entities and NRDEs were extracted, the results in terms of P20 were quite close to the best runs at the ImageCLEF-08 Photo Retrieval task [11].

4 Concluding Remarks

In this paper we have offered a first approach and discussion of the benefits and problems of extracting enriched Named Entities for a NLP task such as Photo Retrieval. It is our belief that linguistic analysis such as NER can be highly beneficial for NLP tasks. The results show, however, that such linguistic techniques cannot be applied in isolation, and that a better analysis of the topics should be performed. We believe that further improvement of the approach presented in this paper would positively impact on the overall task.

In this sense, ongoing work includes a comparison our approach (with and without NEs and NRDEs) using the imageCLEF-09 testset; preliminary results seem promising so far. As it has already been mentioned, the Belga collection and topics are much more appropriate to test our approach that last year's testset.

Finally, our approach to re-bracketing Named Entities when they are part of descriptions will in itself be improve and evaluated following a similar procedure to the one described in [8].

References

1. Kripke, S.: Naming and Necessity. Harvard University Press, Cambridge (1980)
2. Russell, B.: On denoting. Mind 14, 479–493 (1905)
3. Granados, R., Benavent, X., Agerri, R., García-Serrano, A., Goñi, J.M., Gomar, J., de Ves, E., Domingo, J., Ayala, G.: MIRACLE-FI at ImageCLEFphoto 2009. In: CLEF Working Notes 2009 (2009)

4. Lestari Paramita, M., Sanderson, M., Clough, P.: Diversity in photo retrieval: Overview of the imageCLEFPhoto task 2009. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsikrika, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 45–59. Springer, Heidelberg (2010)
5. Clark, S., Curran, J.: Language Independent NER using a Maximum Entropy Tagger. In: Proceedings of the Seventh Conference on Natural Language Learning (CoNLL 2003), Edmonton, Canada, pp. 164–167 (2003)
6. Clark, S., Curran, J.: Wide-coverage efficient statistical parsing with CCG and Log-Linear Models. Computational Linguistics 33 (4), 493–553 (2007)
7. Granados, R., García-Serrano, A., Goñi, J.: La herramienta IDRA (Indexing and Retrieving Automatically). In: Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN 2009 (2009)
8. Finkel, J.R., Manning, C.D.: Joint parsing and named entity recognition. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, Colorado, pp. 326–334. Association for Computational Linguistics (2009)
9. Salton, G.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
10. Granados, R., Benavent, X., García-Serrano, A., Goñi, J.M.: MIRACLE-FI at ImageCLEFphoto 2008: Experiences in merging text-based and content-based retrievals. In: Proceedings of CLEF 2008 (2008), http://www.clef-campaign.org
11. Arni, T., Clough, P., Sanderson, M., Grubinger, M.: Overview of the imageCLEFphoto 2008 photographic retrieval task. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 500–511. Springer, Heidelberg (2009)

A Continuum between Browsing and Query-Based Search for User-Centered Multimedia Information Access

Julien Ah-Pine[1], Jean-Michel Renders[1], and Marie-Luce Viaud[2]

[1] Xerox Research Centre Europe
6 chemin de Maupertuis, 38240 Meylan, France
firstname.lastname@xrce.xerox.com
[2] Institut National de l'Audiovisuel
4 avenue de l'Europe, 94366 Bry-sur-Marne Cedex, France
mlviaud@ina.fr

Abstract. Information seeking in a multimedia database very often implies a search process that is complex, dynamic and multi-faceted. Moreover the information need with respect to a topic is likely to evolve during the same search session, going from a simple lookup search to a thorough discovery of connected subtopics. We propose a system that aims at addressing these challenges. It couples serendipitous browsing and query-based search in a smooth manner. The proposed system offers two levels, one global and one local, of visualizing the context of the information seeking task and it also allows to view and search the data using either monomodal or cross-modal similarities. Furthermore, the system integrates a new relevance feedback model that takes into account the multimodal nature of the data in a flexible way and a combination of two parameters, the locality and forgetting factors, that allows the user to design adaptive metrics in the interactive search process. The paper also presents a preliminary user-centered evaluation of our system and concludes with an analysis of the evaluation results.

1 Introduction

In the Information Seeking field [1], we can distinguish different strategies to access and explore multimedia databases. One strategy is *serendipitous browsing:* the aim is to navigate in a digital library in order to get an overview of its different themes and its underlying structure. In that case, a tool that groups together similar objects and allows the user to visualize the similarity relations between them is required. Another strategy is *query-based search:* here, the aim is to find quickly relevant objects with respect to a given query using a tool that takes into account the user feedback in order to iteratively improve the search results. In this use case, the key features rely on avoiding redundancy and visualizing the similarity relations between the retrieved objects so that the user can have a better understanding of the different subtopics. But a more general scenario happens when the user wants to have a mix between serendipitous search and

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 111–123, 2011.

query-based search. Indeed, it is often very hard for the user to formulate an unambiguous query, which is the direct translation of her information needs. It also happens that the user does not know exactly what she is looking for: she has a general question in mind, but she does not know in which direction she needs to start searching. In that perspective, the ideal search process is a discovery process, where the user could incrementally precise her requirements depending on what the system proposes; understand the direction she is currently investigating with respect to the global picture; and go back to explore new directions, being aware of the boundaries of this discovery process.

The system we propose here aims at addressing these complex needs, with a "mixed-strategy" approach. It offers some continuum between browsing-based search and query-based search. In this paper, we focus on digital libraries whose objects are constituted of texts and/or images but the proposed methods and tools could be extended to other modes (speech, music, . . .). This paper is also related to information fusion and, especially, to cross-media techniques that can combine visual and textual aspects efficiently in order to bridge the gap between these two modes when exploring, exploiting and searching databases of multimedia objects.

The rest of this paper is organized as follows. In section 2, we give an overview of the global architecture and the main novelties of our system, while detailing each component. In section 3, we present some results of a preliminary user-centered evaluation based on the Cognitive Walkthrough method. Then, in section 4, we analyze some related works before concluding in section 5.

2 Description of the System

2.1 Global Architecture and Main Functionalities of the System

The architecture of the system is depicted in Fig. 1. It consists of several interlinked components: the Graphical User Interface, the monomodal Search Engines, the Ranker/Scorer and the Graph Layout Map Builders. The designed system aims at achieving the following functionalities:

• **interlinked multi-scale visualization and navigation:** the system offers (at least) two levels of visualizing the context of the information seeking task. One visualization is the 2D *global map* of the whole multimedia corpus, emphasizing the underlying structure of the digital library. The structure is typically characterized by different clusters and sub-clusters, with mutual positions indicating how these clusters relate to each other. The second map, called the *local map*, synthesises the history of the current session, by representing all objects the user has to interact with, on a single 2D map. These two maps result from the Graph-Layout Map Builders component that we detail in subsection 2.4. Objects of the local map are linked to their counterpart in the global map. Having two maps allows the user to be aware of the boundaries of her search, to understand the different landscapes at different scales, and to better control the exploration (global visualization) and exploitation (local development) phases.

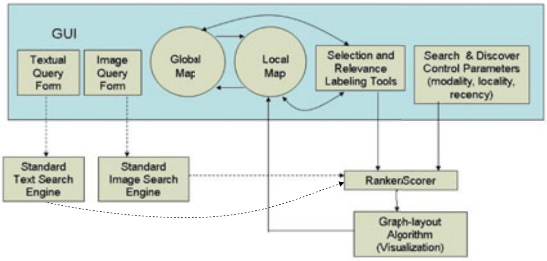

GUI
Textual Query Form
Image Query Form
Global Map
Local Map
Selection and Relevance Labeling Tools
Search & Discover Control Parameters (modality, locality, recency)
Standard Text Search Engine
Standard Image Search Engine
Ranker/Scorer
Graph-layout Algorithm (Visualization)

Fig. 1. Architecture of the proposed system

• **multimodal views of the data:** on the global map, the user can have different views of the data. She can switch to textual, visual or hybrid similarities, so that there are actually 3 static maps that co-exist. This allows the user to change the global map according to the modality she is interested in at each step of the session.

• **flexible multimodal relevance feedback:** on the local map, the user can label the text and the image parts of a same multimedia object differently. In that way, she can associate relevant texts with relevant images that best correspond to her current multimodal information need.

• **adaptable search/development metrics:** our relevance feedback technique implemented in the Ranker/Scorer component described in subsection 2.3 contains several parameters that allow the user to design metrics that adapt to her information need at each iteration. First, after giving feedback, the user will typically promote different kinds of similarity for the next step: her search can rely on textual, visual or hybrid similarities. Next, the system allows the user to tune a suitable combination of a locality factor and a forgetting factor, that will weight accordingly all the accumulated information (the initial query and the objects with relevance feedback) in the current session. More particularly, locality allows some selected objects to have more weight than others, in order to "develop" the similarity graphs locally and to give a new direction to the search. The forgetting factor assumes that the user is naturally more prone to give more importance to what she interacts with recently, an assumption that was already introduced in the "Ostensive Model" [2].

Before detailing each component, let us introduce the following definition: a *session* is a sequence of interactions between a user and the system, that corresponds to the same information need or task. These interactions include visualizations and proposals from the system side, query formulation and/or object selection and/or relevance feedback from the user side.

2.2 The Graphical User Interfaces

The GUIs performing the global and local maps contain 3 parts (see Fig. 2 a and b). The central part represents the maps, the left part shows some detailed view of the data that corresponds to what is selected by the mouse in the central part. The right part of the interface is different for both maps:

- For the global map, the right part includes two standard query interfaces (textual query and image query) that are typically used at the beginning of the session, in order to generate an interesting subset of objects for further developments as the "page zero" of the local map. The search engines are standard ones, typically returning the k nearest neighbors of a given query (the latter objects are then highlighted on the global map). Note that the use of query forms is optional as the user can simply select one or more objects of the global map to "develop" them in the local map.
- For the local map, the right part is dedicated to the parameter settings of the adaptable search/development metrics (feedback, modality, locality and forgetting factors) as it is shown in Fig. 2 b). Labeling of the retrieved elements (text and/or image) is done by selecting the corresponding items with the mouse and by clicking on the "+" (relevant) or "-" (non-relevant) button of the right panel. The items which are neither annotated "+" nor "-" are considered as neutral. They have a null weight and they remain displayed on the local map with a grey color; even if of not immediate interest, these neutral points could still be labeled later in the session. Finally, the type of search modality could be chosen among textual, visual or hybrid modes; the locality and forgetting factors can be tuned in order to better reflect the current needs.

The user can zoom in/out or move the map panels with the mouse roll. To launch a local map, the user selects one or several objects of the global map by clicking on them and activates the adequate menu item. A new window appears and the local discovery/search process can start. On the local map, chosen positive items are put in red whereas negative items are first put in green but finally disappeared at the next iteration. Once the items have been labeled, clicking on the "submit" button launches the retrieval process with feedback. Results appear instantaneously on the local map and the corresponding items are highlighted on the global map.

2.3 The Ranker/Scorer Component

The Ranker/Scorer component is the core of the system: it generates at each iteration a ranked list of objects, that are considered to have the largest probability of being relevant, given all the cumulative information (relevance feedback labels and initial query) and the different search/discovery parameters of the current iteration (selected search modality, forgetting and locality factors). This component also deals with the issue of merging the textual and visual modes, when needed; in what we propose, this could be partly realized by defining a cross-media similarity measure based on a mix of real and pseudo-relevance feedback[1]. We propose the formula given in eq. 1 for computing a new relevance score for each (unlabeled) object, x, of the database based on the accumulated feedback

[1] Our proposal is an interactive extension of the trans-media pseudo-relevance feedback introduced in [3,4,5] for the non-interactive case.

information and the control parameters chosen at the current iteration t. It can be seen as a non-trivial extension of Rocchio's method [6] to the more general case of interactive multimedia information seeking.

$$
\begin{aligned}
f^{t+1}(x) = \\
&\gamma_T^t \left[\sum_{y \in \mathcal{T}_+^t} \frac{\alpha_T^t(y)}{\sum_{y' \in \mathcal{T}_+^t} \alpha_T^t(y')} \left(S_T(y,x) + \lambda_T \frac{\sum_{z \in \mathcal{B}_T^t(y)} S_T(y,z) S_I(z,x)}{\sum_{z' \in \mathcal{B}_T^t(y)} S_T(y,z')} \right) \right. \\
&\left. - \sum_{y \in \mathcal{T}_-^t} \frac{\beta_T^t(y)}{\sum_{y' \in \mathcal{T}_-^t} \beta_T^t(y')} \left(S_T(y,x) + \delta_T \frac{\sum_{z \in \mathcal{B}_T^t(y)} S_T(y,z) S_I(z,x)}{\sum_{z' \in \mathcal{B}_T^t(y)} S_T(y,z')} \right) \right] \\
&+ \gamma_I^t \left[\sum_{y \in \mathcal{I}_+^t} \frac{\alpha_I^t(y)}{\sum_{y' \in \mathcal{I}_+^t} \alpha_I^t(y')} \left(S_I(y,x) + \lambda_I \frac{\sum_{z \in \mathcal{B}_I^t(y)} S_I(y,z) S_T(z,x)}{\sum_{z' \in \mathcal{B}_I^t(y)} S_I(y,z')} \right) \right. \\
&\left. - \sum_{y \in \mathcal{I}_-^t} \frac{\beta_I^t(y)}{\sum_{y' \in \mathcal{I}_-^t} \beta_I^t(y')} \left(S_I(y,x) + \delta_I \frac{\sum_{z \in \mathcal{B}_I^t(y)} S_I(y,z) S_T(z,x)}{\sum_{z' \in \mathcal{B}_I^t(y)} S_I(y,z')} \right) \right]
\end{aligned}
\tag{1}
$$

In eq. (1), $f^{t+1}(x)$ is the new relevance score of the (unlabeled) multimedia object x provided at iteration $t+1$. The subscripts T and I respectively correspond to text and image modality. S_T and S_I are then the textual and visual similarity matrices. Let denote mod^t the search modality(ies) chosen by the user at iteration t. mod^t can take the value T or I or H (hybrid: both T and I). In the sequel, we introduce the notations with respect to the text modality only. However, since text and image play symmetric role, one can deduce the corresponding definition for the image part by simply replacing the subscript T with I, the set notation $\mathcal{T}$ with $\mathcal{I}$ and "text" with "image" in the text (and vice-versa). γ_T^t reflects the weight given by the user to the text modality at iteration t. More precisely, we have γ_T^t is null if $mod^t = I$ and set to a positive constant c_T otherwise. $\mathcal{T}_+^t$ is the set of objects whose textual part was labeled as relevant by the user up to step t. On the contrary, $\mathcal{T}_-^t$ is the set of texts that were labeled as irrelevant up to iteration t. α_T^t and β_T^t are weights[2] that give the importance of texts within $\mathcal{T}_+^t$ and $\mathcal{T}_-^t$ in order to compute the new relevance scores vector f^{t+1}. These weights take into account different parameters. First, the user can select a special subset of the items annotated at the current step t. These selected texts, $\mathcal{S}_T^t$, correspond to the text part of the nodes of the graph around

[2] Note that by default, we take $\beta_T^t(y) = \alpha_T^t(y), \forall y \in \mathcal{T}^t = \mathcal{T}_+^t \cup \mathcal{T}_-^t$, since this setting works better according to some preliminary experiments.

which the system should develop new elements. In comparison to other labeled items, the selected objects are given an extra weight $loc_T^t \in [0,1[$ specified by the user. The greater the locality value, the more the user wants to focus on the newly selected objects. Second, the user can also explicitly mention to the system what is the importance to be given to previously annotated items. This is the role of the forgetting factor $forg_T^t \in [0,1]$. With such a factor, the weight of an annotated item will decrease with time: the older the labeling of an object, the lower its weight. This effect is even stricter as the forgetting factor increases. The "recentness" of the labeling is something not so trivial. Let assume that, at the current iteration t, the user decides to go back to the results provided at iteration $t' < t$ and select some of the items "issued" at t'. This might mean that the user wants to pursue another direction in her information seeking. Therefore, we assume that the objects that were annotated from step $t'+1$ up to $t-1$ are not important anymore. Hence, we give a null weight to these items[3]. More formally, we compute the weight vectors for the annotated objects as follows: $\forall y \in \mathcal{T}_+^t$, we have:

$$\alpha_T^t(y) = \begin{cases} \frac{1}{1-loc^t} & \text{if } y \in \mathcal{S}_T^t \\ (1-forg_T^t)^{m_T(y)} & \text{if } y \notin \mathcal{S}_T^t \text{ and } m_T(y) \geq 0 \\ 0 & \text{if } y \notin \mathcal{S}_T^t \text{ and } m_T(y) = -1 \end{cases} \tag{2}$$

where:

$$m_T(y) = \begin{cases} t - date_T(y) & \text{if } \mathcal{S}_T^t = \emptyset \\ \min\limits_{z \in \mathcal{D}_T^t(y)} (date_T(z) - date_T(y)) & \text{if } \mathcal{S}_T^t \neq \emptyset \text{ and } \mathcal{D}_T^t(y) \neq \emptyset \\ -1 & \text{if } \mathcal{S}_T^t \neq \emptyset \text{ and } \mathcal{D}_T^t(y) = \emptyset \end{cases} \tag{3}$$

where $date_T(y)$ is the iteration number when the text of object y was annotated and $\mathcal{D}_T^t(y) = \{z \in \mathcal{S}_T^t : date_T(z) \geq date_T(y)\}$. In other words, given $y \in \mathcal{T}^t$, $\mathcal{D}_T^t(y)$ is the set of selected texts z that were annotated after y.
Notice that a locality factor loc_T^t equal to 0 amounts to give no extra weight to selected objects[4]. On the contrary, a locality factor very close to 1 will result in discarding non-selected items. Indeed, when this factor tends to 1, the non-null contributions in the different terms of eq. (1) come only from the selected objects, due to the weighted average effect.

With respect to eq. (1), positive and negative text pseudo-relevance feedbacks are respectively introduced through the terms weighted by λ_T and δ_T. To be more precise, it is a trans-media pseudo-relevance feedback which considers as relevant the visual part of texts that are very similar to the texts fed back as relevant by the user; but this feedback mechanism discounts a pseudo-relevant object by the factor λ_T and by the specific (normalized) textual similarity between the pseudo-relevant text and the corresponding labeled texts whose it is the neighbor.

[3] This case corresponds to the third case in eq. (3).

[4] In this case, it does not make sense to select any object.

Accordingly, we denote $\mathcal{B}_T^t(y)$, the set of texts that haven't been annotated yet and which are the nearest neighbors of (the text part of) y. Similarly, the system considers as irrelevant the visual part of texts that are very similar to the texts fed back as non-relevant by the user. Likewise, this dual negative view of the pseudo-feedback mechanism consists of the terms weighted by the discount factor δ_T in eq. (1). To be consistent, neighbors of labeled objects that are themselves labeled are never considered as pseudo-relevant objects.

2.4 The Graph-Layout Map Builders

This component is the one that produces as outputs the different maps for visualizing globally or locally the objects of interest. Global maps are computed off-line. We first apply a sequence of several force directed layout algorithms to generate the maps, we then use the LinLog energy model [7] as the final stage. The basic material consists of thresholded similarity matrices. A standard agglomerative hierarchical clustering algorithm is then applied to identify clusters in the 2D space. Cluster naming techniques allow then to extract the most representative keywords of each cluster. The local map layout is a dynamic process: results are appended to the map at each "interactive query" performed by the user. Regarding dynamic representations, one additional constraint has been established by the visualization community: the problem of preserving the user's mental map [8]. The objective is not to loose the user by constantly changing the map layout from one iteration to the next one: new objects are added by slightly perturbing the previous layout and using the similarity metrics promoted by the user at the current iteration, while already present objects keep their mutual similarity relations, as a result of all previous interactions. This is realized by increasing the inertia of existing nodes and by using the Fruchterman-Rheingold layout algorithm [9], that appears to be the most adequate for this kind of task. Optionally, a clustering algorithm could be applied as well in the 2D local map, in order to avoid redundancy in the results given at the next iteration and to favor quick local exploration: only the most relevant objects of each cluster will be displayed on the map (see for example [4]). This could be considered as an indirect way of realizing diversity-based re-ranking and can be particularly valuable during the early stages of the search process.

3 Preliminary Evaluation of the System

3.1 User Evaluation Methodology

Evaluations involving users become essential to validate interactive methods for information retrieval [10,11]. Our goal is to have a preliminary user-based evaluation to validate our contributions namely, the interlinked multiscale and multimodal visualization and navigation and the flexible multimodal searching and relevance feedback. To this end, we chose the Cognitive Walkthrough Inspection methodology [12,13] to perform the evaluation. This method involves an experimented user and an evaluator. The evaluator pre-specified scenarios and tasks to be realized, while the user[5] is already familiar with similar tasks and interfaces. The procedure is the following one: the evaluator explains the goals of the task and the functionalities associated to the different sub-components of the GUI. During the test, the evaluator manipulates himself the tool and asks the user for the functional actions to be executed at each step. In fact, our goal is not to consider specific ergonomic aspects, but rather to validate the relevance of new functionalities to achieve the task. It focuses on the following points: achievability (is the set of elementary functions sufficient to solve the task?), efficiency (does the system promote the most efficient paths?), predictability (is the user able to predict the effect of the launched action?), obviousness (how intuitive is it?), proactivity (after the action, is the feedback good enough to encourage her to continue?) and, finally, confidence (is the user more confident about the obtained results?). What is eventually measured is (i) that the user is able to understand the link between the sequence of actions and the final goal of the task, and (ii) that she is able to memorize the corresponding actions and settings. At the end of the evaluation, the user is asked extra questions, related to the comparison with her existing tools, in terms of complementary or new possibilities.

3.2 Design of the Evaluation Scenarios

The corpus used for the experimentation consists of text/image objects extracted from the French Wikipedia around the theme “Tourism in France”. From each page, we extracted several multimodal objects, namely the images present in the page with their associated texts (image caption, text of the surrounding paragraph and sequence of titles and subtitles leading to this paragraph). Note that, due to our construction mechanism, the relationship between an image and its associated text could be noisy or very vague. This collection[6] is made of more than 50,000 text/image objects. A task consists in solving two subtasks related to a same topic: a specific search (closed problem) and a discovery analysis (open problem). The tasks were designed in such a way that it is very unlikely to obtain the information directly by a single textual query and that combining both modalities in a flexible way is essential. Five topics presented to the user were related to: Eiffel Tower, surfing, old stamps, Charles de Gaulle and Nantes.

3.3 Description of a Particular Evaluation Scenario

In this subsection, we report some retrieval results for the Eiffel Tower scenario. For this topic, the user had to retrieve old pictures representing the Eiffel Tower

[5] In our case, she is an expert archivist who is used to seek information in a multimedia database with classical systems.

[6] This collection was constructed for the purpose of the *Infom@gic* project. See the acknowledgments section.

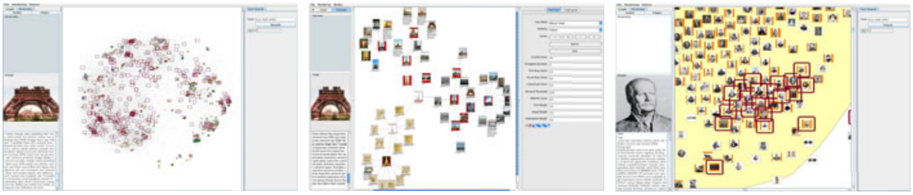


Fig. 2. Screenshots of the system for the Eiffel Tower scenario: a) Global map on the left (with relevant objects highlighted) b) local map in the middle (with different developped search directions) and c) an example of a subtopic viewed from the global map on the right.

and dating from the beginning of the 20th century (closed subtask). She also had to explore the collection in order to gather different multimedia objects that cover all potential subtopics related to the Eiffel Tower, as if she wanted to find as much material as possible to make a multimedia presentation on that topic (open subtask). The evaluator let the user free to solve these subtasks sequentially or in parallel, but the user actually found it more efficient to solve them simultaneously. The user started with a general textual query "Eiffel Tower" using the basic text search engine, whose results were highlighted on the global map (see Fig. 2 a)). She observed that a lot of items were surrounded and their distribution spread all over the global map. After a quick observation, the user mentioned that many of the highlighted objects were not relevant for the specific task. The reason is that the Eiffel Tower is often used as a generic French emblem. After zooming in some particular areas presenting a high density of highlighted results, the user picked an object whose image represents the Eiffel Tower. The latter is the black and white drawing of the Eiffel Tower in the center of Fig. 2 b). From this chosen element, the user started a local deployment with the "hybrid modality". The user asked to set the forgetting factor to 0.2. After 8 iterations during which 12 objects were labeled relevant and 20 irrelevant, the user obtained the results presented in Fig. 2 b). Three orientations have been developed using locality feature to allow a deeper focus on selected objects: the bottom left area gathers technical drawings of the tower; the right branch shows contemporary landscape pictures of Paris and the upper left one contains postcards from the 1900 universal exhibition, portrait of an engineer and so on. Only the upper left branch is relevant to the first task, but all deployment are useful for the secong one. The user noticed here that the effect of the forgetting factor was effective since the strong visual contribution of the first drawing was progressively lowered iteration after iteration. To conclude the second subtask, the user further analyzed the global map and particularly areas presenting a strong density of results. Three more interesting clusters relevant to the task were discovered. Fig. 2 c) shows one of them related to portraits of famous scientists whose names are written on the Eiffel Tower.

3.4 First Conclusions Drawn Out of the User Evaluation

Based on the user's reactions collected during the evaluation (including global comments at the end of the evaluation), we address the different points raised in subsection 3.1:

- Achievability and predictability: in general, the user succeeded in finding satisfying results for the search and discovery parts of each task and had no trouble to perform the list of actions needed to obtain the results. However, some local deployments were very noisy, most often due to the wrong association between text and image in the multimedia collection[7]. In those cases, the user would have liked to mark a positive feedback only on a selected portion of the text. Besides, the control on the local deployment is rich, innovative and interesting but more training is needed to really understand all its possibilities.
- Efficiency, obviousness and proactivity: the use of different modalities appeared very useful for the search. The use of textual queries for generating a "page zero" is particularly valuable. The user exploited both maps for all tasks, their connection turned out to be very intuitive.
- Usability: the user was globally comfortable with using the maps. Navigating in the global map results appeared easy, especially with the display of the text and image in the left part of the panel just by moving the mouse on the items. Selection and launch of the local map was also easy. But, the local map parameters' setting was not obvious. The selection and labeling of the texts and/or images were all right. The use of the focus mode was intuitive.

We can formulate the following preliminary conclusions from this evaluation:

- Using different modalities and particularly cross-media techniques allows to provide faster ways to achieve relevant results particularly when the information need is difficult to express in terms of queries and when the different modalities of the same object do not match from a semantic viewpoint[8].
- Using one global map and one local map jointly allows the user to better control the exploitation/exploration trade-off. The local map allows the user to express her information need more precisely while the global map allows her to better understand the different boundaries of her search and discover non-expected subtopics.
- While using the local map, the user can progressively express her information need by selecting relevant texts and/or images and discarding negative examples, in a flexible manner. This flexibility provides an efficient way to define iteratively complex queries to achieve relevant results.
- The use of forgetting and locality factors are encouraged though we should not loose the user by asking him to tune a lot of parameters. These factors clearly offer a continuum between browse-based and query-based search: the locality

[7] This is a side effect of the way we pre-processed the French Wikipedia corpus. In the case of the Eiffel Tower topic for instance, objects corresponding to other monuments appeared because there is a Wikipedia page that lists the most visited monuments in Europe, so that the same text is associated to very different images.

[8] For example, images that have poor or noisy textual descriptions.

factor allows one to focus on some topic at any time while the forgetting factor helps in discovering new topics.

4 Related Work

The literature covering interactive multimedia retrieval is very vast. In the particular context of keyword-tagged images, the paper [14] presents a relevance feedback approach which integrates semantic (keywords) and low-level features for image retrieval. Their method is an extension of the Rocchio technique [6] but their system only targets basic query-based search and there is no possibility for the user to judge independently texts and images. Focusing on text/image collections, there have been many works in the context of ImageCLEF[9] Photo evaluation campaigns. The paper [15] is a good example of image/text interactive search that shares goals similar to the ones presented here. Their combination method is based on a hierarchical late fusion approach which is different from our technique [3,4]. Systems addressing video retrieval are also related to our proposal. Particularly, the work presented in [16] shows several common aspects with our work. The authors use a multimodal similarity space for representing the multimedia objects; they then apply a one-class SVM in order to learn a classifier that separates relevant from non-relevant examples. Concerning the visualization part of the system, a state of the art of visualization methods and tools developed for multimedia information is given in [17]. Some systems propose a multi-scale view of objects for browsing and interactively searching within a multimedia corpus. The most closely related work to our proposal is [18]. However, in [18], the authors use non-linear embedding algorithms whereas we rather use graph-layout methods. Then, the main difference between the two systems is that we propose not only a multi-scale but also a multimodal view of the data. The Ostensive Model introduced initially in [2] considers the information retrieval process as dynamic and proposes a relevance feedback model that integrates a temporal notion to relevance, very close to our forgetting factor concept. More specifically, our system shares many common points with the work described in [10], which addresses content-based image retrieval from a multimodal perspective, based the Ostensive Model. Still, there are important differences: first, we use two interlinked multi-scale maps whereas in [10] only one map is employed; second, the multimodal nature of the data is not emphasized in the feedback and the search processes in [10]; third, our system explicitly allows the user to annotate many candidates at each step which is not the case in [10]; lastly, the combination of textual and visual information is different in both systems.

5 Conclusion

In this paper, we have introduced the architecture and the key components of a user-centered system for accessing information in a multimedia digital library.

[9] See http://www.imageclef.org/

The novelty of our proposal is to offer some continuum between serendipitous browsing and query-based search by developping the following features: multiscale, multimodal navigation and adaptive multimodal relevance feedback technique. The next steps of this work will consist in evaluating the performances of our system at a larger scale from the user viewpoint. One objective of the work presented here was to introduce sufficient flexibility in the system in order to adapt to evolving user needs. But, from an ergonomic viewpoint, there still remains an important issue in better controlling the trade-off between new degrees of freedom and the mental load of mastering all these degrees of freedom to be more efficient. This can be done by clustering the different usages into several usage families and by further constraining the different parameters values to optimal default value (by family) without requiring any tuning from the user.

Acknowledgments. This work was funded by the *Infom@gic* project, part of the "Pôle de compétitivité Cap Digital de Paris".

References

1. Marchionini, G.: Exploring search: from finding to understanding. Communications of the ACM 49, 41–46 (2006)
2. Campbell, I., Van Rijsbergen, C.: The ostensive model of developing information needs. In: Proc. of CoLIS 2, pp. 251–268 (1996)
3. Clinchant, S., Renders, J.-M., Csurka, G.: XRCE's participation to ImageCLEF 2007. In: Working Notes of CLEF 2007 Workshop (2007)
4. Ah-Pine, J., Cifarelli, C., Clinchant, S., Csurka, G., Renders, J.M.: XRCE's participation to ImageCLEF 2008. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) CLEF 2008. LNCS, vol. 5706, Springer, Heidelberg (2009)
5. Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.M.: Crossing textual and visual content in different application scenarios. Multimedia Tools Appl. 42(1), 31–56 (2009)
6. Rocchio, J.: Relevance feedback in information retrieval. In: The SMART Retrieval System, pp. 313–323 (1971)
7. Noack, A.: Visual clustering of graphs with nonuniform degrees. In: Healy, P., Nikolov, N.S. (eds.) GD 2005. LNCS, vol. 3843, pp. 309–320. Springer, Heidelberg (2006)
8. Misue, K., Eades, P., Lai, W., Sugiyama, K.: Layout adjustment and the mental map. Journal of Visual Languages & Computing 6, 183–210 (1995)
9. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Softw., Pract. Exper. 21, 1129–1164 (1991)
10. Urban, J., Joemon, M., van Rijsbergen, C.: An adaptive technique for content-based image retrieval. Multimedia Tools Appl. 31(1), 1–28 (2006)
11. Diou, C., et al.: VITALAS at TRECVID 2009. In: TRECVID (2009)
12. Huart, J., Kolski, C., Sagar, M.: Evaluation of multimedia applications using inspection methods: the cognitive walkthrough case. Interacting with Computers 16(2) (2004)
13. Polson, P.G., Lewis, C., Rieman, J., Wharton, C.: Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. Int. J. Man-Mach. Stud. 36(5), 741–773 (1992)

14. Lu, Y., Zhang, H., Wenyin, L.: Joint semantics and feature based image retrieval using relevance feedback. IEEE Transactions on Multimedia 5, 339–347 (2003)
15. Rahman, M.M., Desai, B.C., Bhattacharya, P.: Multi-modal interactive approach to imageCLEF 2007 photographic and medical retrieval tasks by CINDI. In: Working Notes of CLEF 2007 Workshop (2007)
16. Bruno, E., Moenne-Loccoz, N., Marchand-Maillet, S.: Design of multimodal dissimilarity spaces for retrieval of video documents. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1520–1533 (2008)
17. Goëau, H., Thièvre, J., Verroust-Blondet, A., Viaud, M.L.: State of the art on advanced visualisation methods. Report D7.2 of the Vitalas EC project FP6 - 045389 (2007)
18. Nguyen, G.P., Worring, M.: Interactive access to large image collections using similarity-based visualization. J. Vis. Lang. Comput. 19(2), 203–224 (2008)

Security-Relevant Challenges of Selected Systems for Multi-user Interaction

Marcus Nitsche[1,3], Jana Dittmann[2,3], Andreas Nürnberger[1,3], Claus Vielhauer[2,3], and Robert Buchholz[2,3]

[1] Data and Knowledge Engineering Group
[2] Research Group Multimedia and Security
[3] Faculty of Computer Science
Otto-von-Guericke-University Magdeburg, Germany
P.O. Box 4120, D-39016 Magdeburg
{marcus.nitsche,jana.dittmann,andreas.nuernberger,
claus.vielhauer,robert.buchholz}@iti.cs.uni-magdeburg.de

Abstract. One important goal in the field of multi-user interaction is to support collaborative work of several users as ergonomic as possible. Unfortunately, security-relevant aspects were neglected in the past. Therefore, we study in this contribution the risks and challenges for security of such collaborative working environments on the basis of five selected pen and gesture-based input techniques. We show that the underlying technologies (Anoto pens, Wii Remotes, DiamondTouch, FTIR Table tops, Microsoft Surface) do have deficits, in particular regarding the insurance of user authenticity and data integrity, and that collaborative working brings new challenges for formal security models. We discuss some of the major challenges on situation and context recognition for dynamic role assignment based on a scenario from the field of energy engineering and point out that several of the underlying problems are of special importance for the development of reliable collaborative multimedia applications for object organization and exchange.

1 Introduction

Multi-user interaction is a rising field of research in which it is studied how several users can work collaboratively and efficient in groups operating only one device. One main focus of this research field is the development of easy to learn or efficiently executable interaction techniques that can be used context dependent by different persons. However, the straight use of new input devices as well as collaborative working itself creates new challenges for IT-security. Currently, these aspects are considered only insufficiently in the design of such interaction systems, even though they are of high relevance for, e.g., plants in terms of safety [1] or military applications [2]. First studies for the dynamic recognition of users in multimedia applications were investigated in the project „CoMET" (Collaborative Media Exchange Terminal)[1].

[1] For further information see, e.g., http://www.dfki.de/iui/advanti/lab/index_de.html

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 124–134, 2011.

In the following, the problems of applying formal security models to multi-user interaction and the weak points produced by the technical restrictions of typical systems for multi-user interaction are examined and required changes to solve some of these issues are discussed. In addition, possible more general improvements of the security of multi-user systems are discussed. Based on typical scenarios from the energy industry and multimedia terminals, interaction techniques and associated security aspects are briefly analyzed. The main goal of this work is to analyse security related problems of multi-user systems in order to sensitize the reader for security related questions that otherwise, if neglected, might strongly affect the reliability of multi-user applications.

The remainder of this contribution is arranged as follows: Sect. 2 gives a short introduction to the formal terms to security. In Sect. 3 general requirements for security of multi-user systems are described and challenges for situation and context recognition resulting from dynamically changing user roles and user rights are discussed. In Sect. 4 we analyze current systems for multi-user interaction, describing the effects of their operational principles on the security of the overall system and make suggestions, how some of the security problems might be solved. Finally, Sect. 5 summarizes the main results of this contribution.

2 The Term Security

Typically, in the English language the complementary terms “security” and “safety” are commonly used. Here, the term ”security” refers to aspects that are related to the defence of intentional attacks, while “safety” refers to aspects that are related to danger prevention, e.g. insurance of the system/working reliability despite of unforeseen events. However, since the causes and effects of safety risks are independent of the number of users working simultaneously at a single system, we consider in this contribution only ”security“ related aspects.

More formally, ”security” deals with assurance of the following five properties [3]:

- Authenticity
- Integrity
- Privacy/Confidentiality
- Availability
- Commitment / Not-Repudiation

Authenticity defines that the origin of a message or the identity of a person can be determined free of doubts, while *integrity* describes soundness, e.g., that a message was not modified during the transmission between transmitter and receiver. *Privacy* is the ability to ensure the secrecy of contents of a message towards unauthorized third persons. *Availability* is the characteristic of a system or resources to be usable within fixed time intervals (or even permanently). Finally, *commitment* or *non-repudiation* is the possibility of third parties to verify authenticity of data or a specific event performed by a specific entity.

Since the risks for the availability of a system do not differ for multi-user interaction or single user desktop systems, in this contribution availability related issues are not discussed explicitly. Additionally, the non-repudiation (traceability) of user actions presupposes compellingly an authentication of users. As shown later, none of the systems discussed in this article offers the possibility to identify an individual user and therefore non-repudiation cannot be ensured at all.

3 Formal Security Challenges

The underlying operating systems of the interaction systems are usually not adapted to the requirements of multi-user interaction and thus work with the assumption that after a user registration, the graphical user interface is exclusively used by this one user. Therefore, during the execution of applications and opening files, only the authorizations of only one user are considered – even if several users collaboratively work on specific documents subsequently. Also, when saving a document the reading and write authorizations are assigned only with the consideration of the user registered at the system and do not dependent on the authorizations of the collaborating users.

Even under the assumption that a collaborative time sharing system could determine reliably the identity of all users acting at the same time, it is so far unsettled, how the authorizations of the group can be derived from the safety authorizations of the individual users. This influences, in addition to the above mentioned security aspects, data protection, in particular if several users are working collaboratively on personal data, where parts should be hidden from some participants of the collaboration (e.g. treatment planning in hospitals, where nurses might not be allowed to see data of an operation).

Here, probably only a *least common privilege* approach would be formally secure, by which the group receives exactly those authorizations, which each member of the group possesses. Thus a group would never have more rights than the group member with the most limited rights and for groups of users with disjunctive quantities of authorizations the group would have no rights at all. Another approach would be to consider the user with the highest security level as the ”protector” of the data integrity and privacy, relying that her/his “leadership qualities” prevents the other group members from passing of confidential information. On these idealized assumptions for a group, all the authorizations could be awarded, which possesses at least one of its members. Thus it would be guaranteed that in a collaborative working environment nobody need to work with fewer rights, than he or she would have in a single user environment. Probable violation of the basic assumptions of this model would endanger both, the integrity and the privacy of the seen and worked on data. More flexible models would be possible, if position, identity and range of vision of all collaborating participants could be determined surely. In this case for the presentation of a document within the interface only those users need to be considered who can see the respective range.

If we regard, for example, the control room of a superregional power supplier, which is a highly sensitive system since the operation lead by not authorized persons can result in spurious action and thus to a substantial legal and in the long run financial damage for the operator. Besides the maintenance of normal operation, a high interest in public security exists, as – among other things – also public traffic systems, road lightings and hospitals depend on reliable, non-interrupted current supply. Now, if a critical system state has been entered and the fast and safe decision of authorized persons is necessary, then it can be meaningful to transfer temporally the user rights of an authorized person to another user. However, an authorized person could be unavailable at the time the system enters a critical system state (e.g. due to sudden health caused absence). Of course, this should happen exclusively in sensitive exceptional cases and secured with the system over a suitable situation and context recognition by means of appropriate parameter monitoring. Time-critical interactions in security sensitive environments, as the presented, can be supported in this way.

Another scenario is a public service portal for exchange of multimedia objects, like the collaborative media exchange terminal mentioned in the introduction. If several users could up- and download media objects at the same time while their specific authorizations for objects are obeyed, this would strongly improve the collaborative experience. The authentification might be ensured as long the users are using their personal mobile device for interaction. However, how is ensured that a specific device on the table belongs to a specific user? This becomes even more critical also from the perspective to ensure intellectual property rights assigned to a specific user – and challenging to handle – if several users are interacting with the terminal at the same time and would like to provide objects only to some of the collaborating users.

3.1 An Environmental Model of Authentication Related Dependencies

Figure 1 shows the dependencies between contextual information (environmental model), the user profiles and the user authorization level of users working on an interactive system. Thus the situation and context recognition presuppose a suitable environmental model, in which the permanent evaluation of various sensor measurements flows, in order to assign user to a certain user profile with appropriate rights and working roles. In the long run so the authentication of a user in a certain situation is clarified.

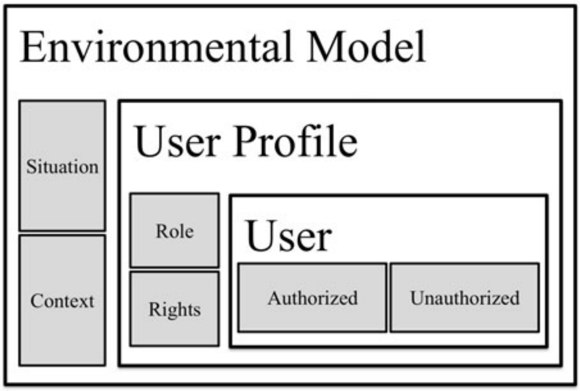


Fig. 1. User authentication based on a specific environmental model

4 Practical Security Challenges of Multi-user Systems

In this section, five common technologies for multi-user interaction are presented and their security related aspects are evaluated. Furthermore, the underlying operational principle of the respective systems are briefly discussed and consequences for IT-security are pointed out.

4.1 Interaction with Anoto-Patterns

A common technology for the multi-user interaction is the use of the Anoto-Pattern and the associated digital pens (e.g. Maxell PenIt, Logitech IO Pen, Nokia SU-1B and SU-27W). The Anoto-Pattern is a proprietary, highly soluble, particularly structured point pattern, which can be applied with infrared-reflecting ink or toner on nearly arbitrary surfaces [4]. The Anoto pens are pen-formed input devices with an integrated, downward arranged infrared camera near the pen tip, which can recognize the section of the pattern under the pen and by this, the position of the pen on the writing surface can be calculated. The current position of the pen can be transferred together with a value, which represents the pressure applied at present on the pen, in real time by wireless communication (Bluetooth) to a computer. Additionally, some smart pen vendors and licensees of the Anoto technology have developed devices with audio recoding technology. An example for such devices is the "Pulse Smartpen"[2], which enables the pen recording as described above, with the additional possibility to record speech of the user.

If several Anoto pens are used on a Anoto-pattern surface then this system can be used to work collaboratively in a multi-user interaction system.

Integrity and confidentiality. A primary concern regarding the security arises from potential vulnerabilities of the Bluetooth protocol. In order for an attacker to listen to a coded Bluetooth communication it is usually necessary to know the common Bluetooth ID used by both devices (in this case a computer and an Anoto pen). Additionally a user PIN needs to be known to derive the used keys for encrypted communication. In practise very often Bluetooth devices (e.g. Bluetooth headsets) are initially pre-configured with 0000 and do not provide on-device interfaces for users to change the PIN. Therefore, it is for an attacker in these cases only necessary to guess or determine the Bluetooth ID. But even with a free selected four-digit pin guessing is easy by exhaustive search. Thus, further communication between pen and computer can be logged, which injures the aspect of confidentiality. With pen-similar input devices the injury of confidentiality (and also privacy) is particularly more problematic, since very often also handwritten notes and even signatures are entered. The digital version of user signatures opens various possibilities of abuse.

Authenticity. The inherent disadvantage of Anoto pens concerning the authenticity is the absence of any user authentication. To identify the individual user, a biometric identification or verification could be introduced to permanently observe human individual handwriting behaviour and to verify this behaviour with a pre-registered identity and behaviour profile (e.g. by handwriting recognition [5]). To detect a change of

[2] See, e.g., http://www.livescribe.com/smartpen/index.html

a pen device between users, this identification or verification needs be done frequently to ensure that a pen is always associated with the correct user identity. With the help of the knowledge about the pin of an Anoto pen as well as its Bluetooth MAC it is also possible, with a Bluetooth equipped PC in place of the system admitted Anoto pen, to spoof the pen device and to send arbitrary data. This would clearly injure the authenticity of the pen user and the integrity of the data.

Possible improvements. A software-based possibility of continuos user authentication could improve the authenticity shortcomings. This could be done, for example, by permanent analysis of writing characteristics of all users, in particular the handwriting within text inputs. Upon significant changes of these characteristics a new authentication could be required.

Further improvements of security would require a change of the firmware and hardware of Anoto pens. By suitable encryption and authentication protocols, the privacy of communication between pen and computer could be guaranteed, for example, by layer 5 security protocols such as TLS/SSL. In addition to this, a fingerprint sensor on the pen, or a biometric speaker recognition function could examine permanently the authenticity of the user. Beyond that, the pen could compute a Biometric Hash [6] on the basis of the fingerprint or voice data and derive from this a secret private key. By use of these cryptographic procedures the authenticity of the user, the not-repudiation of the user interaction and also the trust of communication could be guaranteed. Because of security reasons the pen need to do this examination and it should not be allowed to delegate it to the computer. Therefore, each pen need to store suitable fingerprint templates of all users involved.

4.2 Wii Remote Interaction

A further input device frequently used for prototypes of multi-user interaction is the Nintendo Wii Remote [7, 8]. The advantages of this input device actually developed for Nintendo Wii game console are its low price and its support of different interaction methods. The Wii Remote is a wireless input device, which is connected by Bluetooth to the game console. The communication protocol on basis of the Bluetooth HID profile is proprietary and is also not licensed to external prospective customers, however the most important parts became public by reverse engineering [9]. Thus it became possible to use the Wii Remote directly with computer systems and to interact in such way.

The Wii Remote supports several input modes: It possesses several buttons and a control cross for simple inputs. Further it possesses an integrated accelerometer, which consider accelerations towards the three axes of coordinates and which can seize – considering the acceleration due to gravity – orientation of the Wii Remote in space (not the position). Additionally the Wii Remote possesses an infrared camera with digitally signal processor which computes and transfers the position of the four brightest infrared points in the range of camera vision (IR LED or heat sources). The pointing direction of the Wii Remote to this surface can be determined and the Wii Remote is able to fix points in the proximity of an interaction interface. If several Wii Remotes are used as input devices for the same interaction interface (e.g. projection of a user interface) several users can interact together.

Authenticity. Similar to Anoto pens also with the use of the Wii Remote for multi-user interaction the users usually cannot be authenticated. The authentication could be also made by biometric characteristics while showing towards the interaction surface or by taking up user-specific movement patterns by the accelerometer, e.g. via handwriting characteristics. However intentional or coincidental exchanging of two Wii Remotes could be detected relatively easy, also without modification of the hardware, just by accomplishing the additional interaction possibilities of the Wii Remote for re-authentication.

Integrity and confidentiality. Since the Wii Remote transfers like Anoto pens its data via Bluetooth technology it is susceptible to the same attacks. Therefore, the authenticity of the user and/or the Wii Remote as well as integrity and confidentiality of the transferred data are potentially endangered.

Possible improvements. The Wii Remote revealed by Bluetooth communication the same weaknesses as the Anoto pens and has still the problem of a fixed and well-known Bluetooth PIN. In principle the same improvements of security are possible as with Anoto pens, even if a permanent scan of fingerprints would limit the usability of a Wii Remote far more strongly than this is the case for Anoto pens. Additionally the sensors provides information about location and also explicit user inputs by the integrated control elements, which could be used for as additional source for an individual user analysis and user identification.

Furthermore the ignition button at the lower surface of the Wii Remote (button “B”) could be used as a “dead-man’s button”, which the user needs to keep pressed after the authentication and whose releasing leads to a new authentication. Also a placing of the Wii Remote could be detected by the accelerometer and with a renewed waiving a renewed authentication could be started. By placing a Wii Remote the accelerometer measures only a constant acceleration downwards, which corresponds to acceleration due to gravity.

It should be noted that the Wii Remote as input device is often used for research prototypes, because of its low costs. However, due to the difficult legal situation by use of Nintendo technologies it can be assumed that for commercial products other multi-user input devices will be preferred.

4.3 MERL Diamond Touch Tabletop

Mitsubishi Electric Research Laboratories (MERL) developed the Diamond Touch as a multi-user tabletop system with radio-based input [10]. It consists of an embedded LCD with a field of transmitting antennas near and parallel to the table surface. Each user of the system needs to sit on a chair attached to the tabletop or stand on an attached mat. If the user moves his finger over the surface of the Diamond Touch she or he acts as a receiver for the antennas. These signals are taken up also by the chair/the mat and passed on to the system. The individual antennas send their signals deferred in time, so that the receiving of a signal corresponds to affecting a certain area of the tabletop by a user at a certain time.

Authenticity. Since the recognition of the interaction of each user needs to be made compellable by its chair and/or its mat, the authenticity of the users is guaranteed in principle, provided the identity can be linked to the mat or chair. An identification of the user could be made by gestures or handwritten inputs on the Diamond Touch table, and leaving the chair/the mat could be detected by weight sensors and would require in this case a new authentication.

If two users of the Diamond Touch touch each other and one of the users contact the table surface, then this will be detected by both chairs/mats and recognized by the system like if both users would have clicked in the same area. For most systems this behaviour might be only an unimportant annoyance, in security sensitive applications this becomes important to determine individual user action. For example, in the case of four-eye principal, where two users need to confirm one action separately. In this case touching another user creates a possible threat by assuming recognition of two user actions although only one user performed the action.

Integrity. Since the system requires a cable connection between chair/mat and computer, the system is not susceptible to attacks on radio protocols contrary to the Wii Remotes and Anoto pens. An injury of the privacy/confidentiality, data integrity or authenticity must take place either physically, or directly at the computer - a security risk, which usual PCs are suspended as well.

4.4 FTIR Tabletops

The effect of the Frustrated Total Internal Reflection (FTIR) and its usefulness for the user interaction are well known for a long time, but only a few years ago, it has been taken up by Jeff Han [11] as a user interface concept. FTIR tabletops are based on the effect of the total reflection: Along a glass plate irradiated light is reflected at the upper and lower surface and does not penetrate from there outwards. This effect depends on the boundary surface between two materials – in this case of the glass surface and the surrounding air. If the glass area is affected for interaction, then no total reflection takes place at the point of contact (the FTIR effect). Instead the light will be reflected vaguely – the point of contact seems to shine regarded by the lower surface of the glass plate. During application infrared light is irradiated at the contact points into the glass plate. Under the glass area a infrared-sensitive camera is installed. It records these lighting points and by techniques of automatic image analysis the points of contact are determined. So an interaction with the system is possible. With suitable materials additionally a diffuse-reflecting surface can be applied on the glass plate and by top-down or bottom-up projection a user interface can be projected on the interaction surface.

Authenticity. A FTIR table generally possesses no possibility for user authentication since every contact appears only as non-distinguishable light point. Thus all users are equal from the system's point of view. To identify individual users biometric gesture recognition could be introduced similar to the handwriting recognition and handwriting verification or identification, see section interaction with anoto-patterns. Due to visual pattern recognition with variable light conditions in the process of recognition, verification and identification errors are expected.

Integrity. The visual and optical information path without explicit integrity measure connects the communication between the human gesture as communication source and the table as communication destination. **Confidenciality.** Since all components of the system (projector, FTIR surface with IR LEDs, camera) either built in a common case or at least connected by cables, a certain degree of confidentiality of the interaction data concerning unauthorized third parties is guaranteed in principle. However attacks are possible, but would require physical access to these cables or a logical attack on the computer system.

Possible improvements. One possibility to guarantee authenticity, integrity and not-repudiation of interaction would be to use a high-resolution camera, which could dissolve the light points to identify the fingerprints of the users. This would not only guarantee the authenticity of interaction, but also - if by the system only interactions of recognized fingerprints will be accepted – reduce strongly the number of wrong recognized artefacts and thus the integrity of interaction could be guarantee. This would be probably only possible in a sufficient quality on an interaction surface without diffuser, so that in this case the user interface could not be projected on the interaction surface. A further practical restriction is the creation of cloudy stains when direct working with the fingers on a glass plate, which can prevent from recognizing the fingerprints even with using of high solution cameras. Furthermore with the storage and processing of fingerprints aspects of data protection need to be considered.

4.5 Microsoft Surface

Also Microsoft has developed a tabletop solution for multi-user interaction, Microsoft Surface [12]. The Surface consists of a horizontally installed LC-display, which is framed on the top with LEDs and cameras for emission and recognition of light closed to infrared range. Users can interact directly on this surface with their fingers or objects.

Integrity. The concept is only based on the optical recognition of user interaction. From the data fusion of five cameras the position of the interacting fingers or objects is computed. This object recognition is based only on algorithms of computer vision and the fusion of information of these five cameras. So it is error-prone and in some cases also ambiguous, which undermines the integrity of the data. Therefore data integrity is influenced by the optically and visually introduced error rates.

Additionally to the problem of the ambiguity, the Surface suffers also from the problem of covering, which impairs the integrity of the interaction: If a finger or an object from view of the camera is covered by another object then it can not be seen by this camera and thus not recognized. The Surface possesses five cameras at different positions of the framework, but this does not guarantee a complete detection. A trivial example would be to place a hollow cylinder on the interaction surface. Contacts of the Surface through the hollow cylinder are invisible for all cameras and can not be recognized therefore.

Authenticity. The actual system setup does not offer any possibility of differentiating between parallel working users whereby no individual user can be authenticated and therefore also non-repudiation cannot be ensured.

Confidentiality. Since Microsoft Surface is an integrated overall system without external interwiring between the components its communication can be regarded as confidential.

5 Conclusion

The systems for multi-user interaction discussed here allow for interesting new concepts for collaborative work. However, all of them exhibit severe problems regarding important security aspects. Figure 2 shows in summary the security problems of the presented systems discussed in this work. It can be seen that all technologies shown have already security problems in the fundamental aspects of authenticity, integrity or confidentiality and non-repudiation.

Furthermore, Sect. 3 points out problems that still need to be solved within the scope of formal security models for multi-user interaction. Future collaborative working environments need a support of individual user authentication with the security classification depending on the individual rights and the analysis of the actual roles of all users working together to derive the actual required security settings and mechanisms. Base on the examples from the energy industry and multimedia exchange terminals these special challenges were briefly motivated with a focus on situation and context recognition as well as the rights/roles belonging to it.

While the security problems discussed in this paper do not have to be a disadvantage for research prototypes, it is urgently recommends to examine the consequences of these weaknesses for products; especially in security sensitive areas of application.

	User authenticity	**Data integrity**	**Confidentiality**
Anoto-Patterns	No. But differentiation of users / pens	Endangered by Bluetooth transfer	
Wii Remote	No. But differentiation of users / Wii Remotes	Endangered by Bluetooth transfer	
DiamondTouch	Endangered by touching of other users	Endangered by simultaneous touching of multiple picture elements	Yes
FTIR Tabletop	No	Endangered by errors of picture recognition	Yes
Microsoft Surface	No	Endangered by errors of picture recognition	Yes

Fig. 2. Conclusion of compliance of the most important security aspects by the shown systems

Acknowledgement

Marcus Nitsche is funded by the German Ministry of Education and Science (BMBF) within the ViERforES project (no. 01IM08003C).

References

[1] Rogner, H., Langlois, L.M., McDonald, A., Weisser, D., Howells, M.: The Costs of Energy Supply Security. In: 20th World Energy Congress, Rome, Italy (November 2007)

[2] Szymanski, R., Goldin, M., Palmer, N., Beckinger, R., Gilday, J., Chase, T.: Command And Control in a Multitouch Environment. In: 26th Army Science Conference Proceedings, Orlando, Florida (2008)

[3] Dittmann, J., Wohlmacher, P., Nahrstedt, K.: Multimedia and Security – Using Cryptographic and Watermarking Algorithms. IEEE MultiMedia 8(4), 54–65 (2001)

[4] Technology-Website of Anoto AB, `http://www.anoto.com/digital-pen-paper.aspx`

[5] Scheidat, T., Vielhauer, C., Dittmann, J.: Handwriting verification - comparison of a multi-algorithmic and a multi-semantic approach. In: Image and Vision Computing, Bd. 27, March 2009, pp. 269–278. Elsevier, Amsterdam (2009)

[6] Scheidat, T., Vielhauer, C., Dittmann, J.: Advanced studies on reproducibility of biometric hashes. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BIOID 2008. LNCS, vol. 5372, pp. 150–159. Springer, Heidelberg (2008)

[7] Low-Cost Multi-point Interactive Whiteboards Using the Wii Remote, `http://johnnylee.net/projects/wii/`

[8] Schlömer, T., Poppinga, B., Henze, N., Boll, S.: Gesture recognition with a Wii controller. In: Proceedings of the 2nd International Conference on Tangible and Embedded Interaction, New York, pp. 11–14 (2008)

[9] `http://en.wikipedia.org/wiki/Wii_Remote`

[10] Dietz, P., Leigh, D.: DiamondTouch: A Multi-User Touch Technology. In: Proceedings of the 14th annual ACM Symposium on User Interface Software and Technology, Orlando, Florida, pp. 219–226 (2001)

[11] Han, J.Y.: Low-Cost Multi-Touch Sensing through Frustrated Total Internal Reflection. In: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology, Seattle, Washington, pp. 115–118 (2005)

[12] Microsoft Surface Product website, `http://www.microsoft.com/SURFACE/`

Cognitive-Linguistics-Based Request Answer System

Wolf Fischer[1] and Bernhard Bauer[2]

University of Augsburg
wolf.fischer@informatik.uni-augsburg.de,
bauer@informatik.uni-augsburg.de

Abstract. Closed domains pose very specific problems to applications of all kinds. While e.g. search applications in open domains can access a huge pool of diverse information (e.g. in the internet), this is not possible for closed domains. An example application for this problem are customer support systems. These systems normally administrated by humans have to cope with different types of requests, e.g. a user asking not only a single question, but also giving a description of an experienced situation. Analyzing those complex requests is very difficult in general. Therefore it is nearly impossible to handle for common search, question-answer or customer support systems. In this paper we propose a system which should be capable of (semi-) automatically analyzing and reacting to different types of requests in closed domains based on a cognitive linguistics approach.

1 Introduction

The age of information technology has brought humanity a lot of advantages, especially regarding the accessibility of information. Besides normal key-word based search engines question-answer systems are capable of analyzing simple questions in open domains (e.g. ask.com[1], Answers.com[2], Yahoo! Answers[3] etc.). However these systems are not well suited for closed domains, because the amount of information / requests is smaller than in open domains. Following this fact there are much less syntactic variations available for similar or equivalent information. Due to the syntactic nature of normal search engines this leads to problems in answering user questions in a way which satisfies the person asking. It is therefore either up to the user to alternate his / her search query by using synonyms or an expert of the domain to answer the request.

To deal with these problems there exist different approaches. Most of todays systems focus on indexing documents with different degrees of effort. Some of them cluster documents based on syntactic information like word distribution or frequency, whereas others make use of more advanced natural language processing technologies (e.g. Answers.com) as well as certain semantic measurements

[1] http://www.ask.com

[2] http://www.answers.com

[3] http://answers.yahoo.com/

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 135–146, 2011.

(e.g. with the help of WordNet). Many problems like word sense ambiguities, coreferences etc. are mainly handled by stochastic or statistical methods within generative grammar approaches.

We think that a clearly knowledge-driven approach ([7], [17]) to request-answer systems incorporating the concepts of cognitive linguistics will ease many of todays problems and deliver more satisfying results.

The novelty of our approach lies in the combination of the following points:

1. An easier way to create construction grammars for real world scenarios
2. The use of contextual data within the request analysis process
3. A self-organizing system for the analysis of text based on construction grammars, simultaneously combining syntax and common-sense as well as context semantics

The focus of this paper lays on giving an overview on our approach. It shows the first concepts and ideas behind the framework. Its implementation has started recently.

The rest of the paper is organized as follows: (2) presents some related work. Sections (3), (4) and (5) describe the ongoing project. Finally, (6) concludes the paper.

2 Related Work

Systems which can be compared to our approach are any kind of question-answer systems like the previously mentioned ask.com, Yahoo Answers etc. Most of these open domain question answer systems rely on NLP as well as document clustering technologies.

There are also many systems which rely on additional knowledge, e.g. [16], [15] or [6]. More recent systems are Wolfram Alpha[4] as well as The True Knowledge Answer Engine[5]. The latter is a new kind of search machine which seems to use a similar way to our approach. They maintain a fact driven knowledge base consisting of thousands of facts, which is altered by humans. There are however some differences between their approach and ours: First questions are being converted to a query which returns results from the knowledge base, therefore there is a direct separation between knowledge and text. Secondly the system is only capable of parsing simple questions whereas our aim is to also use the context given by a prior request.

In the field of computer science there have been different approaches to language processing using the foundations of CxG, e.g. Embodied Construction Grammars ([1]) and Fluid Construction Grammars ([13], [14], [12]). The goal of FCG is to create a linguistic formalism which can be used to evaluate how well a construction grammar approach can handle open-ended dialogues ([13]). To evaluate the approach it has been implemented within autonomous, embodied

[4] http://www.wolframalpha.com/

[5] http://www.trueknowledge.com/

agents. Some of the key assumptions of FCG are that it is usage-based (inventories are highly specialised), the constructions are bi-directional (i.e. FCG can handle parsing as well as production of language), it uses feature structures (which are directly incorporated within the constructions) and there is also a continuum between grammar and lexicon ([14]). The production as well as parsing process is handled by a unify and merge algorithm, which allows for an emergent creation of either the semantics or the syntax using best-match-probabilities in the unification of the constructions. This leads to a high robustness of the algorithm and more natural creation of language.

FrameNet is a project which creates a lexical resource for English based on frame semantics ([11]). Recent research tries to combine FrameNet with constructions ([5]), as certain linguistic structures can not be detected by the current mechanisms. In comparison to FrameNet, our approach a) yields for specific domains and b) tries to analyze a text with the goal to identify the users intention based on what is already known about possible requests (see 4.2 and 5.1). Still, FrameNets knowledge could serve as a good common sense basis.

Texai[6] is an Open Source approach to an artificial intelligence system. Its target is "to acquire linguistic and common sense skills that improves its own performance". Therefore it also uses a common sense knowledge base (currently based on OpenCyc) as well as FCG to handle the interpretation of natural language. The project is currently in development and therefore still lacks some knowledge, especially for the FCG component.

FCG and its concepts will serve as a foundation for our ongoing project. The project is going to be used in different types of domains therefore adjustments will have to be made, some of which will be introduced in the following sections.

3 Problem

Our main challenge is the analysis of textual requests of any kind. The system must therefore be able to identify the knowledge as well as the intention of the user as best as possible and 'act' accordingly. One of our main tasks is therefore a consequent way to combine semantic knowledge and language in order to gather the knowledge of a domain. Due to the amount of knowledge needed to make such a system reliable this approach has rarely been tried in the past. In cases this way has been tried there has been a clear separation between language (e.g. a normal glossary) and knowledge itself (e.g. a domain model) without direct combination of these two worlds. However it is clear that knowledge is needed in order to fully understand language (e.g. to disambiguate language [7], [17]). Another problem that we will try to tackle is the inclusion of additional data in the analysis process like earlier requests, context- and / or profile-information ([2],[10]). Basically each piece of information which could be relevant will be usable within the analysis process. As the system should not only be a one-way experience, feedback of as well as interaction with the user is important ([8]) and will also help to improve the accuracy of the system.

[6] http://www.texai.org

4 Knowledge Structure

This section describes the reference meta model (4.1) which acts as the central model to the different knowledge parts. Each part is accompanied by a simple example for demonstration. The examples notation is leaned to UML.

As seen in figure 1, the meta model is separated in different **Scope**s. Scopes are there to model different aspects of a **Domain**. A Domain would be e.g. a car manufacturer or a computer seller. Therefore a Domain contains everything that is needed in order to answer requests which are specific to that certain Domain.

As can be seen in figure 1, there are different specializations of scopes. The three main ones are the **SemanticScope**, the **SyntacticScope** and the **ConstructionScope**. The SemanticScope is further endetailed by a **SemanticProfileScope** as well as a **RequestAnswerScope**.

Scopes can be seen as some kind of view onto a database (in this case the domain) and not a clear separation. This way the syntax-lexicon continuum stays intact, as all the data is stored in the domain itself (e.g. [3], [9]).

Another argument for this separation is that we think it will be easier to let experts handle their specific fields without interfering in other experts fields. This way a classic 'domain' expert can care about modeling and populating the semantics of the domain, another one (e.g. a linguist) handles the syntactic scope and a last one finally brings both these worlds together by creating the construction scope. Especially in contrast to the way constructions are handled in FCG this should improve the time which is needed to create constructions.

The five different scopes will be described in the following paragraphs.

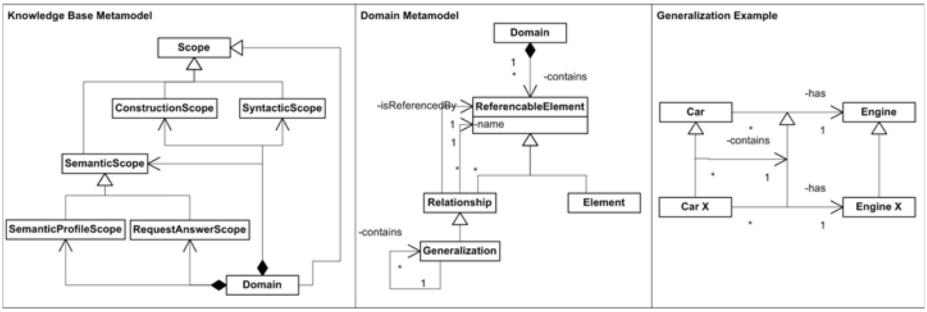


Fig. 1. Overview of the knowledge base structure (left), domain metamodel (middle) and example (right)

4.1 Domain

The Domain is the scope, which contains all data of the domain to be modeled. Therefore it acts as a) a container for all the data within the actual domain as well as b) a reference model for all the data which can exist within the domain. This allows the direct combination of different parts of the model, as described later in this paper.

The structure of the domain is seen in figure 1. Central to the Domain is the **ReferencableElement** (RE). Everything which should be referenced within the KB is of type ReferencableElement, starting with the **Relationship** and the **Element**. The former is used to relate REs with each other. As a Relationship is an RE itself, it can reference other REs as well as Relationships. An Element is a more concrete specialization, used to express different kinds of concepts in later phases. Central to our KB is the **Generalization**, which will be used in every Scope. It allows the creation of taxonomies in the KB. Like in modern programming languages there is be the possibility to overwrite certain parts of the hierarchy. An example is shown in figure 1. In there the SemanticElement 'Car X' specifies that it has exactly an engine 'Engine X' and not 'Engine'.

In the following sections all elements of the Domain, which will be reused are pictured in the figures using a gray background.

4.2 SemanticScope

The SemanticScope is about modeling all the factual knowledge which is inherent to the actual domain, e.g. specifying that a house consists of walls or a car has an engine. To model this kind of facts the model sticks to a generic approach. Central to the SemanticScope are the **SemanticElement** (SE), the **Generalization** as well as the **Association**. A SemanticElement is a specialization of the Domain::Element, as can be seen in figure 2. It is used to represent the concepts (in the former example 'Car' and 'Engine') which are relevant to a specific domain. An Association is used as a generic mechanism to relate SemanticElements between each other. The type of the Association is, opposite to other ontological standards, defined by a SE that the Association can reference via the 'isOfType'. This will help us later in combining syntax and semantics, using constructions.

As already mentioned there are two further specializations of the SemanticScope: The SemanticProfileScope and the RequestAnswerScope. The former will hold additional semantic information about the current user, whereas the latter contains knowledge about what is needed in order to match a request to a specific answer.

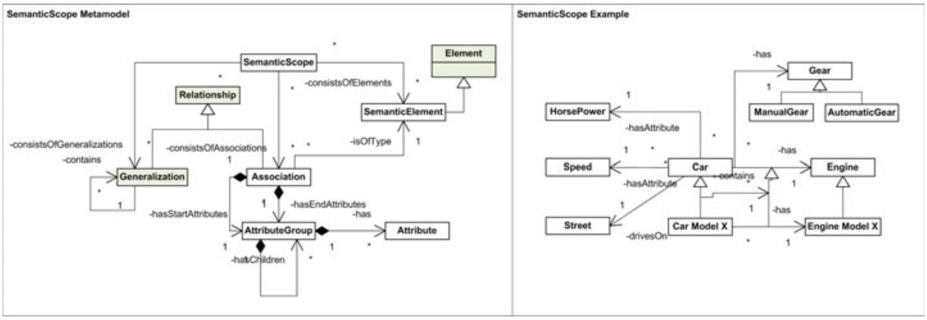


Fig. 2. Overview of the SemanticScope of the metamodel and an example

Figure 2 shows an excerpt of a SemanticScope, which contains information about cars (note that instead of an actual link to another concept, as described in the knowledge structure, the type of an association is given as the textual description of an association, this way keeping a better overview).

4.3 SyntacticScope

The SyntacticScope references all the forms which will later be used to represent the actual concepts, e.g. the SemanticElement 'Car Model X' could be referenced by the actual forms 'Car' or 'Model X'. The SyntacticScope (figure 3) contains the **SyntacticElement** at the top. This is inherited to the **Form** as well as the **SyntacticCategory**. A Form will actually represent the different strings (in the example above 'House' or 'Building'). Each Form can further be associated with a SyntacticCategory (e.g. 'Substantive', 'Verb') etc.. This is in alignment with the fact that different cultures use different syntactic categories, therefore we can include an arbitrary amount of syntactic categories. However it is a very time consuming task to add every possible form of a word (just think of verbs and their different forms throughout different times like go, went, gone etc.). Therefore there is a more generic way to represent forms, i.e. the root of a word (**FormRoot**). This will serve as the basis for a collection of different forms of the same word. A FormRoot can further directly associate its more specific full forms.

An example of a SyntacticScope is seen in figure 3. The FormRoot "Driv" is referencing three more specific occurences: "Drive", "Drives", "Driven". The first one can either be a verb (thus referencing the SyntacticCategoryGroup SynCatGroup 3) or a noun.

4.4 ConstructionScope

The ConstructionScope contains all the information necessary for combining the SemanticScope with the SyntacticScope. This is done by creating **Construction**s. A Construction contains **Symbol**s, which can be linked by **SymbolLink**s. A Symbol can either be a **SemanticSymbol** (referencing a SemanticScope::SemanticElement), a **SyntacticSymbol** (referencing a SyntacticScope::SyntacticElement) or a **Unit** (referencing several other Symbols at once).

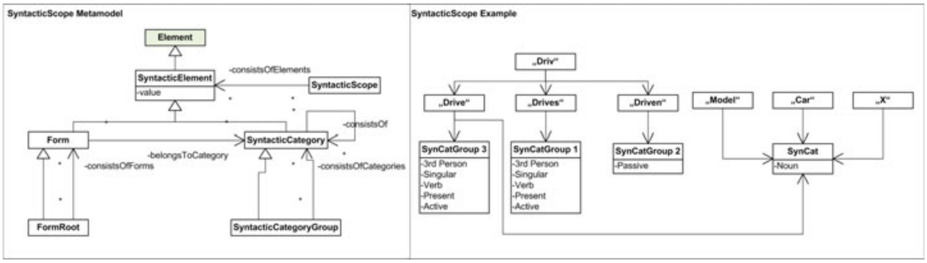


Fig. 3. Overview of the SyntacticScope of the metamodel and an example

Table 1. Construction for mapping a SemanticElement to its forms

Attribute	Content
Name:	'Car Model X' representation
Syntactic Symbols:	$Model(m), Car(c), X(x)$
Semantic Symbols:	$CarModelX(cmx)$
Relations:	$Aggregates(u, \{m, c, x\})$
Mapping:	$u \Leftrightarrow cmx$

To create a mapping between a SemanticSymbol and a SyntacticSymbol, the SymbolMapping is used.

In case of abstract constructions (i.e. constructions which mainly reference syntactic categories), there can be further specifications about the referenced (syntactic or semantic) elements. For example, there could be an abstract construction C, which contains two semantic symbols s_1, s_2. In case of both symbols referencing the same semantic element e, a SemanticRelation could state, that the textual occurrences of s_1 and s_2 must be different subtypes of e: $subtype(e, s_1) \neq subtype(e, s_2)$.

The following lines will give examples on how Constructions can associate the SyntacticScope with the SemanticScope. The construction in 1 shows how the SE 'Car Model X' is associated with its corresponding words. First, the different syntactic elements are referenced, namely the Form "Model", "Car" and "X". These three elements are the same as in figure 3. Each element is therefore associated with a variable name (standing in braces behind the corresponding element). Next, the SEs are defined, in this case the 'Car Model X' only. The construction has to specify, that all three syntactic symbols should be treated as one, therefore 'aggregating' these three into one. The new element, which will represent the three single ones, is called u and corresponds to a 'Unit' within the ConstructionScope (4.4). Finally the mapping between the SyntacticScope and the SemanticScope is specified by specifying $u \Leftrightarrow cmx$. Another more abstract

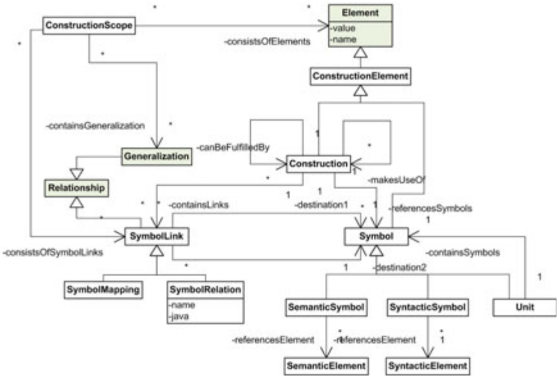


Fig. 4. Overview of the ConstructionScope of the metamodel

Table 2. Construction for a noun phrase

Attribute	Content
Name:	NounPhrase
Syntactic Symbols:	$Noun(n), Adjective(a), Determiner(d)$
Semantic Symbols:	$Element(e1), Element(e2)$
Relations:	$relation(e1, e2), inOrder(d, a, n)$
Mapping:	$n \Leftrightarrow e1, a \Leftrightarrow e2$

example is given in 2. It specifies what a noun phrase looks like. Therefore, the SyntacticElements Noun, Adjective and Determiner a referenced. Further two abstract Elements $e1$ and $e2$ are specified. Next it is defined that both SemanticElements must be related somehow (specified by $relation(e1, e2)$) and that the SyntacticElements must exist in a specific order, i.e. determiner → adjective → noun. Finally, the mappings between the $e1$, $e2$, n and a are specified. A final example will use the previously defined construction 2 for specifying a complete sentence structure. It consists of a noun phrase → verb phrase (specified accordingly to the NounPhrase construction) → noun phrase. In order to reference other constructions, a new field 'Other Constructions' is introduced (which is equivalent to the 'makesUseOf' association in figure 4). As a Construction consists of a semantic and a syntactic part we can use these parts in further specifying the relations in the new construction. There it specifies that the semantic part of noun phrase 1 and the verb phrase must be related (the same accounts for the verb phrase and noun phrase 2). Further the syntactic parts of noun phrase 1, the verb phrase and noun phrase 2 must be in order.

5 Request Analysis

To accomplish our goals we will use an iterative approach which tries to 'shape' the result until it matches users intention. Each iteration is comprised of different steps. First the user enters her request as she would in any generic online support

Table 3. Construction for a simple sentence structure

Attribute	Content
Name:	NP-VP-NP
Other Constructions:	$NounPhrase(np1), VerbPhrase(vp),$ $NounPhrase(np2)$
Syntactic Symbols:	
Semantic Symbols:	
Relations:	$relation(np1.semantic, vp.semantic),$ $relation(vp.semantic, np2.semantic),$ $inOrder(np1.syntax, vp.syntax, np2.syntax)$
Mapping:	

site. Next the system analyzes the input by simultaneously analyzing the syntax as well as the semantics, which is in line with Texai, i.e. newly created knowledge is checked based on a common sense knowledge base. However as there are different foci between our project and Texai or FCG on the goals to be achieved, there are some things left which must be adapted in order to make the system usable for our intentions.

One of these things is incorporating knowledge adapted in prior contacts with the user. The system will therefore store prior requests as well as specific user information within a corresponding semantic profile. Further a semantic 'guessing' based on the RequestAnswerScope information will reduce the overall processing complexity of the system. Additionally, if not pleased with the result the user can alter his request and let the system process it again. The next paragraphs will give an outlook on the concept behind the analysis steps.

5.1 Textual Analysis

The brain is often referred to as an emergent system as it is not (yet) deducible for humans, how a large collection of neurons can yield such complex and still reasonable behavior. The emergence attribute also accounts for language parsing and production, which is sometimes also referred to as the most complex task humans are capable of. Therefore the system is going to be designed as a self-organizing system. We will first have a look at the different components of the system. On the micro level the system consists of (Construction-) Cells (CC). If being created a cell is omni-potent, i.e. it can differentiate into any known construction (it's the cells 'DNA'). Differentiation means that a cell looks for a construction which seems promising to its current local context and interprets this construction with all local information available to the cell. This allows a cell to perfectly fit into a specific context. A cell is able to fission. Depending on the fission direction (down, up or aside) the new cell has prior information helping it differentiate (only if fission is going downwards). No cell is able to control other cells (which is one aspect of a self-organizing systems ([4])). Each cell is attached to a specific level (see 5): Level 1 contains the cells which are directly attached to the text. Level S contains the cells holding construction knowledge about complete sentences. In between are cells which span a web between Level S and 1. The Interpretation Organism (IO) is the macro level of the system and therefore represents the macro behavior which is a direct implication of the micro level. The macro behavior is seen in the semantic as well as syntactic interpretation which are being created by the single cells.

The following lines will show a short scenario of how the cells can adapt to a sentence in a decentral and self-organizing manner. In order to analyze text an IO is initialized with the text (e.g. 'She drives'...) as well as a single cell (CC1). As text is expected to consist of sentences the first cell differentiates into a generic sentence cell (i.e. a sentence is just a sequence of words) and is therefore located on Level S. Based on the premise that the SyntacticCategory 'Sentence' can consist of 'Word's (e.g. in English or German), the cell fissions downwards, trying to fill the wholes in its sentence construction. The new cell

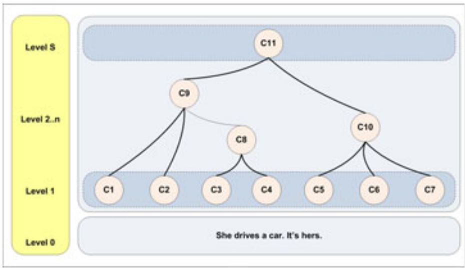


Fig. 5. Illustration of the different cell levels

(CC2) has the information about being a word from its parent and therefore differentiates based on a construction which can be attached to the first position of the text (i.e. ‘She’, possibly a ‘Subject’). Attachment to the text means that this construction will be located on Level 1. As CC2 ‘sees’ that there is more text left it fissions into a new cell (CC3), i.e. it fissions aside. The new cell is located on level 1 and therefore attaches to the next word (i.e. ‘drives’, potentially a ‘Predicate’). At the current state there are now 3 cells: One on L-S and two on L-1. The new cell on L-1 knows that there is a direct neighbor but there is no direct connection to it yet i.e. there is a cell missing which connects both. CC3 is alone and therefore fissions into a level above its parent (it can and will also fission aside). The new cell (CC4) tries to attach to CC2 and CC3, as both are near to CC4. Based on the information given by these cells it searches for a construction matching on CC2 and CC3. In our example fitting constructions could be any starting with a ‘Subject’ + ‘Predicate’ description which leads to a simple ‘Subject - Predicate - Object’-Sentence. Further the construction has the information that the subject should be related to the predicate. Based on the available domain knowledge the cell can confirm that a person (‘She’) can drive. This makes CC4 a better match for this context than CC1. Therefore CC2 dismisses its connection to CC1 making it possible for CC4 to dock to CC2 and CC3. CC1 however ‘dies’, therefore making room for CC4 on L-S.

The prior example should illustrate the way the algorithm in a self-organizing manner manages the creation of a construction network. As this only illustrated the basic process behind the self-organizing process, there are more mechanisms needed in order to give feedback to this dynamic aspect of the system. Most of these incorporate feedback mechanisms, especially between the micro- and the macro- levels. These will be explained in the following lines.

If a CC attaches to another CC, it will trigger an increase of importance of the referenced semantic elements (SE). This increase works in two phases:

1. Phase 1 is the direct attachment of a new CC, in turn increasing the corresponding SEs.
2. Phase 2 is the attachment of a new CC to another CC, yielding another one-time increase of the SEs importance of the ‘old’ CC.

Phase 2 can therefore be seen as a heightening of the context relevance of the SEs. However there can also be an increase of importance in the other direction, i.e. from the SEs to the CCs, again working in two phases:

1. Phase 1 is in case of the initial increase of importance of an SE leading to an importance increase of all CCs attached to this SE.
2. Phase 2 is in case of the SEs being part of a possible semantic request. If this request is referenced by a certain amount of SEs, it will trigger an increase of importance of all attached CCs.

This way of strengthening (especially phase 2) accomplishes our goal of 'guessing' the correct request from the start and thus also helping reducing overall processing complexity. The importance of the cells and elements directly affects the cells 'willingness' to dock to other elements.

The strengthening of importance of different elements is also the key component to easily incorporate existing knowledge of a user into the analysis process. Therefore, prior knowledge will be marked as having a high importance at the beginning.

Following these description the cells participate in the selection of an answer which should match users input best.

6 Conclusion

In this paper we have given an overview of a system which should be capable of handling all sorts of user requests in case the corresponding information is available to the system. Therefore it heavily relies on a domain dependent ontology containing information about factual as well as language related knowledge. By the strong connection of these both worlds we hope to achieve better results in the analysis of textual customer requests.

This promising approach is currently in development and will yield first results soon. Existing corpora (e.g. from the Delph-In[7] project) and databases will be reused especially for languagerelevant information.

References

1. Bergen, B., Chang, N.: Embodied construction grammar in simulation-based language understanding. In: Construction grammars: Cognitive grounding and theoretical extensions, pp. 147–190 (2005)
2. Chai, J., Jin, R.: Discourse structure for context question answering. In: Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004, pp. 23–30 (2004)
3. Croft, W.: Radical construction grammar: Syntactic theory in typological perspective. Oxford University Press, USA (2001)
4. De Wolf, T., Holvoet, T.: Emergence versus self-organisation: Different concepts but promising when combined (2005)

[7] http://www.delph-in.net/

5. Fillmore, C.: Border conflicts: Framenet meets construction grammar (2008)
6. Fu, J., Xu, J., Jia, K.: Domain ontology based automatic question answering. In: International Conference on Computer Engineering and Technology, ICCET 2008, vol. 2, pp. 346–349 (2009)
7. Isahara, H.: Resource-based natural language processing. In: International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2007, 30 August - September 1, pp. 11–12 (2007)
8. Kaufmann, E., Bernstein, A.: How useful are natural language interfaces to the semantic web for casual end-users? In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 281–294. Springer, Heidelberg (2007)
9. Langacker, R.W.: An introduction to cognitive grammar. Cognitive Science: A Multidisciplinary Journal 10(1), 1–40 (1986)
10. Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., Karger, D.: The role of context in question answering systems. In: Conference on Human Factors in Computing Systems, pp. 1006–1007. ACM, New York (2003)
11. Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., Scheffczyk, J.: Framenet ii: Extended theory and practice (2006) (unpublished manuscript)
12. Steels, L.: Fluid Construction Grammar Tutorial. Tutorial (2004)
13. Steels, L., De Beule, J.: A (very) brief introduction to fluid construction grammar. In: Proceedings of the Third Workshop on Scalable Natural Language Understanding, pp. 73–80. ACL
14. Steels, L., De Beule, J.: Unify and merge in fluid construction grammar. In: Vogt, P., Sugita, Y., Tuci, E., Nehaniv, C.L. (eds.) EELC 2006. LNCS (LNAI), vol. 4211, pp. 197–223. Springer, Heidelberg (2006)
15. Tapeh, A., Rahgozar, M.: A knowledge-based question answering system for b2c ecommerce. In: Fifth International Conference on Information Technology: New Generations, ITNG 2008, pp. 321–326 (2008)
16. Hermjakob, U., Hovy, E.H., Lin, C.-Y.: Knowledge-based question answering (2002)
17. Wang, Y.: A tri-level knowledge representation model for nlp. In: IMACS Multiconference on Computational Engineering in Systems Applications, vol. 1, pp. 547–553 (October 2006)

Using Search Logs to Recommend Images to New Users

Yan Xu and Michael Oakes

Department of Computing, Engineering and Technology,
University of Sunderland, Sunderland SR6 0DD, UK
{Yan.Xu-1,Michael.Oakes}@Sunderland.ac.uk

Abstract. We make use of search logs provided by the Belga News Agency to recommend images downloaded by previous users to new users. Each search session in the logs consists of a session ID number, the ID of the images which were downloaded at the conclusion of that session, and the various search terms which were input leading up to the selection and downloading of those images. In our approach, we match the queries of future users against the search terms in each session of the logs, and return the images selected in the best matching search sessions. In this way images considered relevant by previous users are recommended to future users with similar queries. An evaluation using P@50 for ten common queries produced encouraging results. This work describes a variation on the traditional Information Retrieval paradigm, where instead of text documents or images being indexed according to their content, they are indexed according to the search terms previous users have used in finding them.

Keywords: Search Logs, Relevance Feedback, Search Engines, Recommendation, TF.IDF, Cosine Similarity Coefficient, Still Images.

1 Introduction

As part of the VITALAS [1] project, we are looking at search log data usage in search engines, especially for relevance feedback. The search log data is provided by a multimedia professional archive, Belga News Agency of Belgium, which holds 1.5 million still images in a multi-media search engine.

In Multi-Media Information Retrieval, search log data have the potential to be used in search engines, either for relevance feedback or to enhance other search tasks in Information Retrieval. The Belga News Agency (referred to as Belga) is an online news press agency which covers all national and international news, e.g. politics and economics, finance and social affairs, sports, culture and personalities, and supplies news content in text, pictures, audio and video formats [2]. Our work is focused on past users' browsing and searching behaviour on Belga pictures [3], which have been recorded by Belga search log data of their picture website.

Our aim is to use search log data to list relevant still images, downloaded by previous users with the same or similar queries, as searching assistance to future users. The motivation of our work is that this method can be useful to Multi-Media search engines to provide a form of relevance feedback service to users. It is of interest to

M. Detyniecki, A. García-Serrano, A. Nürnberger (Eds.): AMR 2009, LNCS 6535, pp. 147–158, 2011.

investigate whether previous users' choices could be of interest to new users when they input the same or similar queries.

In conventional relevance feedback, the users are asked to evaluate an initial set of retrieved documents. Those judged most relevant become the seeds for a second pass search, where new documents are found which are most similar to those judged relevant at the first pass. In this paper, the relevance judgments have been made by previous users. We know that they found certain documents relevant to them, since they took the decision to download those images and pay for them. Our research question is that based on search log data, the previous users' "download" actions can form the seeds for retrieving relevant documents for future users with the same or similar queries. Thus previous users provide the relevance feedback, while future users potentially benefit from this information.

The rest of this paper is structured as follows. In Section 2, we describe previous work where people have used the information in previous search logs to help current users find documents. In Section 3, we describe the search logs we use, which are provided by the Belga news agency from the existing image search engine on the Belga website. In Section 4 we compare three automatic query session identification techniques: time-based, session-ID based and query-content based. In Section 5 we describe how previously downloaded images can be indexed using search log data, and in Section 6 we compare our search-log based engine with the existing Belga search engine. Our conclusions are in Section 7.

2 Background

Other researchers have been interested in the use of past user search logs, particularly those which record relevance feedback dialogues, to help query formulation for future users. A number of authors have employed support vector machine (SVM) learning to improve information retrieval with relevance feedback logs. Hoi and Lyu [4] proposed a modified SVM technique in log-based relevance feedback. Another study investigated the search logs from an intranet environment and a query lattice was created with an SVM-Light model and Formal Concept Analysis (FCA) [5]. One of the limitations of the SVM approach is that it relies on users to provide sufficient and correct relevance judgments. On the other hand, there are large amounts of search logs which record user's interactions with most Web search engines. Cui, Wen and Ma [6] extract correlations between query terms and document terms by analysing user logs with a probabilistic model. Researchers also studied the importance of query chain identification, generally called session identification of search logs [7], [8]. Searchers often perform a sequence of queries with a single information need. If a user retrieves a document with a later query, it is logical to assume that the user would have preferred to have seen the relevant document from similar queries by previous users. If a technique could associate the relevant documents with all attempted queries by all previous users, paying especial attention with the query terms repeatedly entered, then it would possible to retrieve relevant documents to the future user with only one or two attempted queries. This is the goal of our work.

A characteristic of still image archive website search logs is that a sequence of queries is often followed by picture downloading actions. The sessions of such search logs combine a sequence of queries for a single information need and retrieved documents (pictures). By holding still image archive website search logs, a search session is determined only when a user downloaded a picture(s). By their downloading activities we assume their search tasks were fulfilled and the documents retrieved were relevant to their queries. Each user could try out several queries before they hit the real ‘answer’. The ‘unsuccessful’ queries nevertheless provide valuable information and we use them as potential entry points in a search for the image which was eventually downloaded. We first used time boundaries and then exploited Belga session IDs to identify our searching sessions; finally we have adapted a query context based algorithm [9] to improve our session identifications for Belga search logs.

3 Belga Search Log Datasets

Since the Belga website was upgraded during late 2007, the search log datasets are in two different forms. The first set contains their recorded search logs from the 22nd of June to the 12th of October 2007. This Belga search log set is simply called the SL1 set. SL1 contains 404621 interactions from 358 users. Examples of the search log entries are as follows:

```
2007/22/06 12:02:32: [7970] NIEUWS SEARCH_PICTURES
id=NULL|image_id=|max_result=2000|type=1|period=ALWAYS|
eq1=|text1=voting|fuzzy1=|eq2=|text2=|fuzzy2=|eq3=|text
3=|fuzzy3=|credits=5+39|nr_results=2000
2007/22/06 12:02:49: [8868] VRT1 DOWNLOAD_PICTURE
id=1361693
```

Each line/entry records a user’s action. It can be seen that the logged data includes date, time, etc. The fourth token is the user’s ID, e.g. NIEUWS, and thus we can easily identify which entry belongs to which user. The fifth token records the user’s action, such as browsing, searching, showing or downloading. There are seven actions altogether, but the two actions “SEARCH_PICTURES” and “DOWNLOAD_PICTURE”, are what interest us. “SEARCH_PICTURES” contains the user’s query terms and “DOWNLOAD_PICTURE” means that after searching and browsing, the user finally decided to download the picture with payment. We assume that if the user downloaded a certain picture(s), then his/her information need was fulfilled by the downloaded documents. The sixth token contains the parameters of the “SEARCH_PICTURES” action which correspond to entries in Belga’s advanced search interface prior to October 2007, where “text1 =”, “text2 =” and “text3 =” are followed by the user’s query text inputs. The user ID can help us to segment this huge search log set into smaller search log data for each individual user. The “DOWNLOAD_PICTURE” action provides the image ID the user downloaded.

The Belga picture website then had its search log system updated, and the provided search log data from the 7th of December 2007 to the 4th of September from this new version was simply called the SL2 set. SL2 contains 2586244 interactions from 440 logged in users and anonymous users. The major difference related to our work was that in SL2, Belga added a 5-digit number to identify a user’s session, referred to as “Belga session ID” for logged-in users where this information did not exist in SL1. There were also two other search activities in SL2, namely, "SEARCH_COVERAGES" and "SEARCH_GALLERIES", which mean that the user is searching within a smaller dataset, a certain photographer's works or a certain news topic or event. Both activities also require a text query, so they are treated the same way as the "SEARCH_PICTURES" activity. Apart from these differences, the search log entries in SL2 still provide the same information that we want to process, only in a different format. Since the format was different, SL1 and SL2 were processed with different JAVA programs to produce query sessions and to process query terms. A few lines from SL2 are shown below:

```
2007/18/12 01:12:43: [14236] TRENDS1 LOGIN_SUCCESS
2007/18/12 01:13:07: [14236] TRENDS1 SEARCH_PICTURES
||||||100|Jo and cornu|TODAY|||
2007/18/12 01:13:52: [14236] TRENDS1
PUT_PICTURE_IN_BASKET 6830191
2007/18/12 01:14:06: [14236] TRENDS1 SEARCH_PICTURES
||||||100|karel and boone||||
```

4 Session Identification

Since we are interested only in query-download sequenced actions, session identification is needed to associate queries and downloaded documents. A session is generally defined as *a series of queries submitted by a user during one episode of interaction between the user and the Web search engine*. All the terms of the series of queries ultimately leading to the retrieval of a document can be associated with that retrieved document. Three session identification methods have been put into practice. Each method is relatively simple and easily implementable without involving probabilistic methods. Therefore, the computational costs are low. Unlike general web search logs, the Belga search logs provide user ID information, so they can be separated for each user. Thus our methods do not focus on identifying user IDs but on identifying session boundaries. We constrain the entries of each session to be all on the same day to simplify the procedure for all session identification methods. Our first method for identifying a session is a time based method. Having manually inspected some of the search log data, we chose thirty-minute temporal boundaries. Catledge and Pitkow [10] argued that it is quite reasonable to use thirty-minutes as the temporal boundary, including time for browsing activities. The idea is that the first line of each user's search log data is the beginning of the first session. Once the user has performed a download action, the program saves all the preceding query entries from within the previous half hour. If the user then makes a new query entry, then it is the beginning of the next session, and any download actions before this new query are saved to the first session. The same searches/download sequence identification routine is followed to segment further sessions. This method is straightforward but could include irrelevant query terms, or miss relevant query terms if they were entered more than thirty minutes before the eventual download action.

For Belga search log SL2, it is possible to use their login session IDs to identify whether the query terms and following download actions belonged to a single Belga session, so it was not necessary to rely on temporal boundaries. However, if the user made more than one search/download sequence in a single Belga session, then it would be saved as several query sessions. This website session ID method detects query sessions using search logs metadata but can lead to inaccuracy and tends to include more irrelevant queries because it only indicates log in and log out information about users which does not always correspond to our search sessions.

The third method, a query content-based method, was adapted from Jansen et al. [9] who proposed this algorithm to detect query reformulation patterns within a session by a searcher. Their results showed that defining sessions by a query reformulation algorithm provided the best performance. Queries are classified into six categories: *Assistance*, *Content Change*, *Generalization*, *New*, *Reformulation*, and *Specialization*. Although they are all interesting categories, we mainly used the 'New' category. The 'New' category is defined as the query being on a new topic. In [9], the initial query (Qi) is identified from a unique IP address and cookie. When a subsequent query (Qi+1) contains no terms in common with the previous query (Qi), the start of a new session is identified. This idea is adapted in our work since we define search-retrieve pair activities for each session. The last query before the 'download' activities, called the 'final query', can identify the end of each session, so we compare the 'final query' backward with all the previous queries before the previous download activities to identify the start of each session.

In our method, we first identify fuzzy session boundaries by segmenting search log data with 'DOWNLOAD_PICTURE' activities. If there are consecutive download activities with no queries in between, then they belong to the same session. If there are queries in between download activities, the queries are all considered to be in the same session as well as the later download activity, and we want to eliminate irrelevant queries. Unlike Jansen et al.'s method, the last query before the 'download' activities, called the 'final query' is first identified for each session. In order to find the first query, called the 'initial query', we compare the 'final query' consecutively backward with the queries above to find the number of matching terms. If this is one or more, the pattern is not categorized as 'New', then the procedure is repeated until we do find a 'New' pattern. If no 'New' pattern is found, then all the queries are included in the same session. If we do find a 'New' pattern, the queries between the 'New' query and the final query are considered to be on the same topic, and thus these queries along with the final query are included in the session, so the start and the end of the session are identified.

The rationale is that the final query is assumed to consist of the most effective query terms for the user to retrieve the relevant document, so when we compare the final query with previous queries one by one in a backwards direction, the queries which are not in the same topic as the final query ('New') are all excluded.

Our sessions are given IDs with user name and index number of the session in that user's search log data. For example, `session_ACKROYD_19` is the nineteenth session in user ACKROYD's search log data.

According to Jansen et al.'s analysis, the accuracy of classifications for both time-based and query content-based methods is quite high, but the query content-based method addresses the contextual aspects that the time-based method does not and it appears to provide the most detailed method for session identification. Therefore, this method of session identification was applied for search log indexing in this paper.

A brief investigation of the effectivenesses of the three session identification techniques was carried out by comparing thirty automatically-derived sessions with manually identified search sessions. The results are shown in Table 1. A true positive (TP) is a search statement which appeared in both the automatic and human-assigned sessions; a false positive (FP) is a search statement found in the automatic session, but not in the human-assigned session; a false negative (FN) is a search statement assigned by the human judge, but not by the automatic technique; Precision = (TP / TP + FP); Recall = (TP / TP + FN). Overall, the time-based and session-ID based methods had slightly better recall, while Jansen et al.'s query-content based method produced much better precision.

Table 1. The precision and recall of proposed session identification approaches

	TP	FP	FN	Precision	Recall
time-based	53	18	1	0.75	0.98
Belga session ID-based	53	28	1	0.65	0.98
query content-based	50	0	4	1	0.93

5 Search Log Sessions Indexing

Search log set two (SL2) was chosen to be identified into sessions and then indexed. In SL2, 47829 sessions are identified with 286 users. It should be noted that the total number of users of sessions is smaller than the total number of users in the search log data (SL2). This is because some users never downloaded any pictures during their interactions so no sessions were identified for them. The query terms in sessions are first tokenized and lower cased. Highly frequent words, such as *the, a,* and *of*, called stop words, are much less useful and are removed using the *SMART* stop words list [11]. Although the queries were occasionally in other European languages, such as French or Dutch, only English stop words are removed since combining stop words lists in many languages could result in the loss of meaningful terms. No stemming rules were applied. The TF.IDF measure was used to index the search sessions as it takes into account not only the raw frequency of the term in a particular document (*term frequency*, or TF), but also the inverse of the number of documents in the collection in which the word appears (*inverse document frequency*, or IDF). In our work, the sequence of queries in a session is considered as a document, whereas the whole set of sessions is seen as the collection. In each session, TF.IDF weights of each query term are calculated with the following formula [12]:

$$w_{kd} = f_{kd} \cdot \log\left(\frac{NDoc}{D_k}\right) \quad , \tag{1}$$

where w_{kd} is the weight reflecting the typicality of term k with respect to session d, f_{kd} is the raw frequency of term k in session d, $NDoc$ is the total number of sessions in the collection, and D_k is the number of sessions which contain term k at least once. The highest TF.IDF scores are given to those terms which are common in the session we are looking at, but do not occur in many other sessions.

Instead of investigating the vocabulary characteristic of a document, such as a Web page, we are interested in the vocabulary characteristic of a session made up of a sequence of queries and retrieved documents. This will enable us to match our stored index terms session by session against a future user's query to enable the sessions to be ranked. The TF.IDF weight of each query term then forms a feature vector for each downloaded still image in each session. The image downloaded in the best-matching session would then hopefully be the most relevant previously-downloaded image to the future user's query. An excerpt of TF.IDF session term indexing vector is shown in the following example: the first parameter is the session ID, consisting of the user ID and session index number; the second parameter is the ID of the image the previous user downloaded, and the third parameters are the query terms and their TF.IDF weights in each session. For each session, the past user could have downloaded one or more pictures; they could also have downloaded the same picture more than once. Thus it is possible to have repeated image IDs within the same session.

```
session_ALM_461    9728608  hamilton:9.66|
session_ALM_461    9610643  hamilton:9.66|
session_ALM_462    711048
evans:9.55|la:10.04|cancellara:42.24|trial:46.42|
time:48.95|contre:13.75|montre:13.97| millar:14.24|
session_ALM_463    1981238
de:5.65|france:7.77|tour:8.33|public:12.86|
session_ALM_464    9719861  italy:9.12|cheese:13.97|
session_ALM_465    9717601  bierset:11.70|
...
session_ALM_474    168619
de:11.30|la:20.07|students:23.95|numerus:12.17|clausus:
12.24|fuente:27.12|medecine:13.75|
```

The work described in this paper is implemented in Java SE JDK 1.6. We first separate the search log set according to the user IDs, then produces search log sessions of each user, then calculate the TF.IDF weights of each query term in each session of the search log set. This approach whereby search logs are searched to find previously downloaded images most relevant to a current query is illustrated in Figure 1:

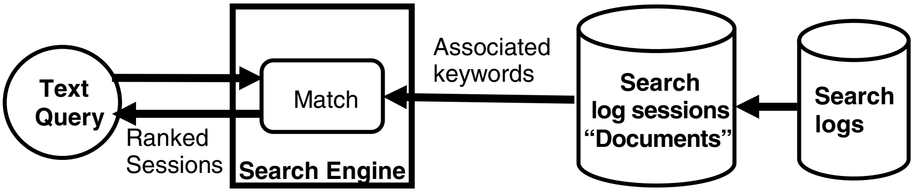


Fig. 1. A search engine using search logs for relevance feedback

6 Evaluation

To evaluate our work, we needed a set of queries to test the system and to measure the results. Due to the lack of existing gold standard data, we produced a frequency list of all queries (from SL1 without using the session identification procedure) and the twenty most frequent queries were sent to a Belga professional who selected ten queries from them, as shown in Table 2.

Table 2. The list of ten queries used to evaluate this image recommender approach

INDEX	QUERY	INDEX	QUERY
1	Cycling	6	Tour de France
2	Anderlecht	7	Sarkozy
3	Henin	8	King Albert
4	Standard	9	Reynders
5	Clijsters	10	Angelica

We chose a popular formula, the *Cosine Similarity Coefficient*, to match each set of query terms shown in Table 2. against each set of session index term TF.IDF weights derived from the SL2 sessions. The formula of the Cosine Similarity Coefficient is as follows:

$$Sim(D_i, Q_j) = \frac{\sum_{k=1}^{t} (DTerm_{ik} \cdot QTerm_{jk})}{\sqrt{\sum_{k=1}^{t} (DTerm_{ik})^2 \cdot \sum_{k=1}^{t} (QTerm_{jk})^2}} . \qquad (2)$$

In this formula, $DTerm_{ik}$ is the TF.IDF weight of index term k in document i (in the traditional use of the formula), or session i in the way we have used the formula. $QTerm_{jk}$ is the number of times term k appears in query j, t is the total number of terms in the vocabulary of the system, and $Sim(D_i,Q_j)$ is the similarity between the document and the query in the range 0—no overlap at all between the query and document terms—and 1—the query terms and the document terms are identical. The algorithms of indexing and evaluation take linear time with respect to the number of search log sessions.

In our experiments, the cosine similarity measure was used to determine which of the indexed sessions were most relevant to the query out of all of the sessions in the search logs. The queries were tokenised and stop-listed, and became the query terms. No stemming rules were applied as the queries could be one of several European languages. The sets of words determined previously by the TF.IDF method as being most typical of each session were regarded as "document" terms. The sessions were ranked with respect to each query, according to how well the query terms matched the session terms in each case by the cosine similarity coefficient. Each session is associated with downloaded image IDs, so the images themselves can be used to measure the results. The 50 most highly ranked pictures for each query were sent to a Belga professional to judge their relevance. The precision@50 for each of these ten queries is given in Table 3:

Table 3. The Precision@50 of the ten queries using indexed SL2 sessions

INDEX	QUERY	PRECISION @50	INDEX	QUERY	PRECISION @50
1	Cycling	0.98	6	Tour de France	0.98
2	Anderlecht	0.78	7	Sarkozy	0.48
3	Henin	0.8	8	King Albert	0.86
4	Standard	0.7	9	Reynders	0.84
5	Clijsters	0.76	10	Angelica (P@9)	1

The pictures judged relevant (top block) and non-relevant (lower block) from the first 50 images downloaded from the top-ranked sessions for the query 'Anderlecht' (a well-known Belgian football team) are illustrated in Figure 2.

The document terms associated with some of the ranked top fifty images when matching with the query 'Anderlecht' are listed below with their TF.IDF weights:

```
session_RTLTVI_1088      1198404  anderlecht:13.72|
session_VRT6_26     8091332 anderlecht:6.86|
session_JCDL_22     8091476 anderlecht:6.86|
session_IVDB_7      8097648 anderlecht:6.86|
```

For the query 'Angelica', only nine pictures were returned, so the precision among just those 9 images was calculated. For other queries, such as 'Cycling' and 'Sarkozy', more than fifty previously downloaded images matched with a cosine similarity measure equal to 1, so the equal top-ranked images were ranked by their image IDs, leaving some pictures outside the top 50 even though they are probably also relevant.

We compared this search log based approach with the conventional search engine approach used by the existing Belga website search engine. The queries listed in Table 3 were also input to the Belga website and the first returned fifty results were judged by the same Belga professional for their relevance. The precision@50 for the Belga website search engine is illustrated in Table 4:

Table 4. The Precision@50 of the ten queries with Belga website search engine

INDEX	QUERY	PRECISION @50	INDEX	QUERY	PRECISION @50
1	Cycling	0.46	6	Tour de France	0.06
2	Anderlecht	0	7	Sarkozy	0.58
3	Henin	0.04	8	King Albert	0.28
4	Standard	0.58	9	Reynders	0.82
5	Clijsters	0.86	10	Angelica (P@9)	0

One reason for the relatively poor performance of the Belga website search engine is that it sorts the results by date rather than similarity to the query, meaning that the most recent pictures with the query term appearing in their captions would appear first. For example, Henin was a famous tennis player in Belgium, but she quit her

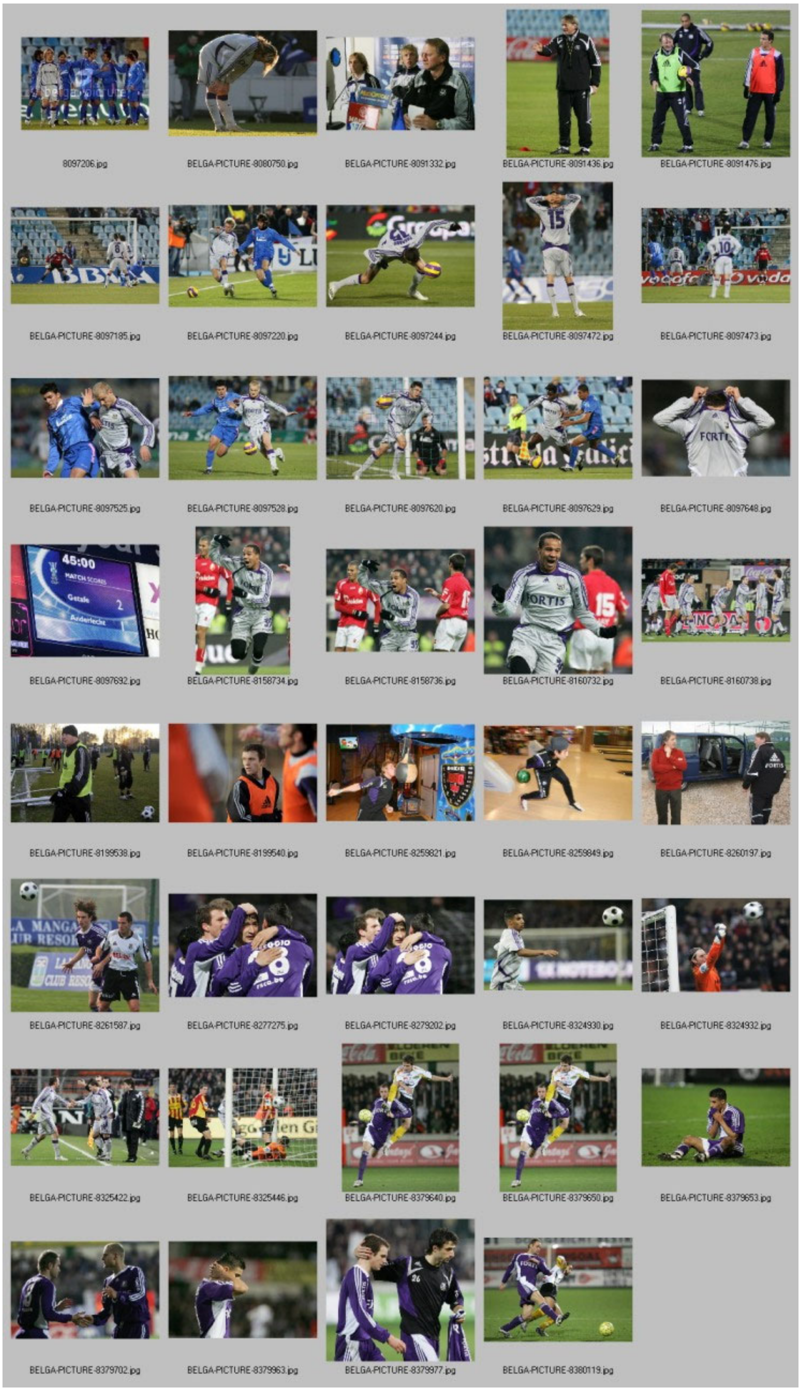

8097206.jpg
BELGA-PICTURE-8080750.jpg
BELGA-PICTURE-8091332.jpg
BELGA-PICTURE-8091436.jpg
BELGA-PICTURE-8091476.jpg
BELGA-PICTURE-8097185.jpg
BELGA-PICTURE-8097220.jpg
BELGA-PICTURE-8097244.jpg
BELGA-PICTURE-8097472.jpg
BELGA-PICTURE-8097473.jpg
BELGA-PICTURE-8097525.jpg
BELGA-PICTURE-8097528.jpg
BELGA-PICTURE-8097620.jpg
BELGA-PICTURE-8097629.jpg
BELGA-PICTURE-8097648.jpg
BELGA-PICTURE-8097692.jpg
BELGA-PICTURE-8158734.jpg
BELGA-PICTURE-8158736.jpg
BELGA-PICTURE-8160732.jpg
BELGA-PICTURE-8160738.jpg
BELGA-PICTURE-8199538.jpg
BELGA-PICTURE-8199540.jpg
BELGA-PICTURE-8259821.jpg
BELGA-PICTURE-8259849.jpg
BELGA-PICTURE-8260197.jpg
BELGA-PICTURE-8261587.jpg
BELGA-PICTURE-8277275.jpg
BELGA-PICTURE-8279202.jpg
BELGA-PICTURE-8324930.jpg
BELGA-PICTURE-8324932.jpg
BELGA-PICTURE-8325422.jpg
BELGA-PICTURE-8325446.jpg
BELGA-PICTURE-8379640.jpg
BELGA-PICTURE-8379650.jpg
BELGA-PICTURE-8379653.jpg
BELGA-PICTURE-8379702.jpg
BELGA-PICTURE-8379963.jpg
BELGA-PICTURE-8379977.jpg
BELGA-PICTURE-8380119.jpg

Fig. 2. The thumbnails of the relevant (top block) and non-relevant (lower block) still images from the first 50 images downloaded from the top-ranked sessions for the query ‘Anderlecht’

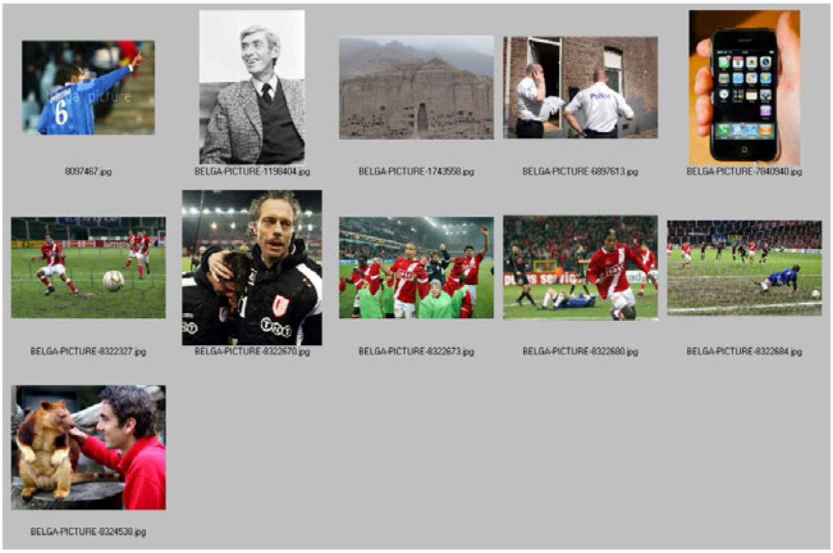

8097467.jpg
BELGA-PICTURE-1198404.jpg
BELGA-PICTURE-1743558.jpg
BELGA-PICTURE-6897613.jpg
BELGA-PICTURE-7840940.jpg
BELGA-PICTURE-8322327.jpg
BELGA-PICTURE-8322670.jpg
BELGA-PICTURE-8322673.jpg
BELGA-PICTURE-8322680.jpg
BELGA-PICTURE-8322684.jpg
BELGA-PICTURE-8324538.jpg

Fig. 2. (*continued*)

career about one year ago. So the first fifty results (the recent pictures with 'Henin' in the caption) had little relevance to the tennis player Henin. Still, the results show that our work, which indexes the search log sessions by linking previous query terms with previously retrieved documents, performs well when compared with the existing conventional search engine approach which indexes the captions of pictures.

7 Discussion and Conclusion

The limitation of this query log approach is that if previous users have never downloaded a particular image, then that image can never be retrieved by this technique. This is a practical limitation of the training data, not a theoretical limitation, but shows the need for large amounts of search log training data.

In the work described here, the search unit is the session from the user logs. In future, we plan to modify and reevaluate this approach, by making the previously downloaded image the search unit. Collated under each downloaded document ID will be the terms of every query ever submitted in a session leading up to the downloading of that document. Thus query terms from separate sessions leading to the retrieval of the same image will all become index terms for this image. Once again we will use the TF.IDF measure. For each image, TF will be the frequency with which each query term was submitted leading up to the downloaded image, and IDF will be the number of downloaded images in the search logs for which this query term was submitted. This approach would overcome the need for removing duplicates from the list of matching images retrieved in response to new user queries.

Acknowledgments

This work was supported by the EU-Funded VITALAS project (project number FP6-045389). We would like to thank Tom Wuytack of Belga for making available the Belga search logs and for judging the performance of the search engine described in this paper.

References

1. VITALAS Project, `http://vitalas.ercim.org`
2. Belga, Belga News Agency, `http://www.belga.be`
3. Belga Picture, `http://picture.belga.be/picture-home/index.html`
4. Hoi, C.-H., Lyu, M.R.: A Novel Log-Based Relevance Feedback Technique in Content-Based Image Retrieval. ACM Multimedia, 24–31 (2004)
5. Lungley, D.: Automatically Adapting the Context of an Intranet Query. In: 2nd BCS IRSG Symposium on Future Directions in Information Access, London (2008)
6. Cui, H., Wen, J.-R., Nie, J.-Y., Ma, W.-Y.: Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering 15(4), 829–839 (2003)
7. Radlinski, F., Joachims, T.: Query Chains: Learning to Rank from Implicit Feedback. In: Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago (2005)
8. Jones, R., Klinkner, K.: Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In: Proceedings of ACM 17th Conference on Information and Knowledge Management, Napa Valley, CA, USA (2008)
9. Jansen, B.J., Spink, A., Blakely, C., Koshman, S.: Defining a Session on Web Search Engines. Journal of the American Society for Information Science and Technology 58(6), 862–871 (2007)
10. Catledge, L.D., Pitkow, J.E.: Characterizing Browsing Strategies in the World-Wide Web. Computer Networks and ISDN Systems 27(1), 1065–1073 (1995)
11. Buckley, C.: Implementation of the SMART Information Retrieval System. Technical report TR85-686, Computer Science Department, Cornell University (1985)
12. Belew, R.K.: Finding Out About. Cambridge University Press, Cambridge (2000)

Author Index