Combined Semantic and Similarity Search in Medical Image Databases

Sascha Seifert[a], Marisa Thoma[b], Florian Stegmaier[c], Matthias Hammon[d], Martin Kramer[a], Martin Huber[e], Hans-Peter Kriegel[b], Alexander Cavallaro[d] and Dorin Comaniciu[f]

[a]Siemens Corporate Technology, Erlangen, Germany, [b]Database Systems Group, Ludwig-Maximilians-Universität München, Germany, [c]Distributed Information Systems, Passau University, Germany, [d]University Hospital Erlangen, Germany, [e]Siemens Healthcare, Erlangen, Germany, [f]Siemens Corporate Research, Princeton, NJ, USA.

ABSTRACT

The current diagnostic process at hospitals is mainly based on reviewing and comparing images coming from multiple time points and modalities in order to monitor disease progression over a period of time. However, for ambiguous cases the radiologist deeply relies on reference literature or second opinion. Although there is a vast amount of acquired images stored in PACS systems which could be reused for decision support, these data sets suffer from weak search capabilities. Thus, we present a search methodology which enables the physician to fulfill intelligent search scenarios on medical image databases combining ontology-based semantic and appearance-based similarity search. It enabled the elimination of 12 % of the top ten hits which would arise without taking the semantic context into account.

Keywords: content-based image retrieval, ontological modeling, semantic image annotation

1. INTRODUCTION

The objective is to develop a content-based image retrieval system that uses similarity search extended by a semantic model of lymphoma to increase the quality of an image search. Lymphoma is a cancer that originates in the lymphatic cells of the immune system and presents as a solid tumor of lymphoid cells and sometimes affects abdominal organs*. Recent work[1,2] tends towards the same direction but often lose track of the global picture. In the MEDICO project, we want to provide the user with a holistic view on the patient supporting him with a tool to search for similar-appearing lesions restricted to an individual organ, but additionally including extra-organ disease processes at the lymph nodes.

We want to provide queries such as looking for *similar* patients, i.e. patients showing similar anatomical and pathological characteristics. Investigating the anamnesis and the successful treatment could then give good advice for the case considered. The ability to compare images with those obtained in other patients has the potential to provide real-time decision support to practicing radiologists by showing them similar images with associated diagnoses and, where available, responses to various therapies and outcomes.

2. MATERIALS AND METHODS

We currently use whole-body CT images from our clinical partner for lymph node inspection and lesion search. 100 images have been semantically annotated with terms from the MEDICO-ontology,[3] combining background knowledge represented in medical ontologies such as the *Foundational Model of Anatomy* (FMA)[4] and RadLex.[5] For lesions, an expert annotated lesions within liver, spleen and kidney in 186 images.

To avoid large efforts in annotating images we recently proposed a *semantic reporting process*[6] which makes use of an image parsing system[7] and a semi-automatic semantic reporting tool. The image parsing system automatically detects anatomical structures and generates an initial annotation list, whereas the reporting tool allows the radiologist to complement them. The semantic reporting tool provides the user with term suggestion, fast volume navigation through directly jumping to or zooming into an anatomical region and hyperlink report text passages with the appropriate image location.

*The spleen is subordinated the abdominal organs, even if physicians consider it a lymphatic structure not an organ.

For search we currently provide two complementing mechanisms: *query by concept* enables the user to query the image database by the use of regular expressions where the terms are coming from the MEDICO-ontology. The second search mechanisms is called *query by scribble*: the query interface provides the user with a drawing tool to define arbitrary regions. In our case, we use it to enclose a reference lesion. Combining these two mechanisms, the query language is tremendously extended versus classical content-based image retrieval systems (CBIR). Subsequently, we explain the mechanism with the following sample query:

Find all patients with similar lesions in the liver and with thoracic lymph nodes enlarged.

The image database can be searched for images containing similar regions based on the visual appearance. This is a close approach to classical content-based image retrieval systems (CBIR). The main advantage of our system is that the CBIR results can be restricted by the *query by concept* (here: *enlarged thoracic lymph nodes*). This mechanism furthermore allows to fully automatically limit the results to lesions within the organ which is currently of interest (here: the liver).

2.1 Query by Concept

The semantic annotations are stored in the *Annotation Ontology* (see Figure 1), which is part of the MEDICO ontology stack. The blue arrows are used to depict properties, rectangles for classes and black arrows are inheritance dependencies. The annotation ontology scheme moves the patient to the center. Every patient owns some studies defined by a unique identifier and a specific time period. The MEDICO study is more than just a DICOM study: it is a container for all annotations from images, texts, clinical data within a given time period. This is the cornerstone to enable temporal queries as well as queries considering multiple modalities.

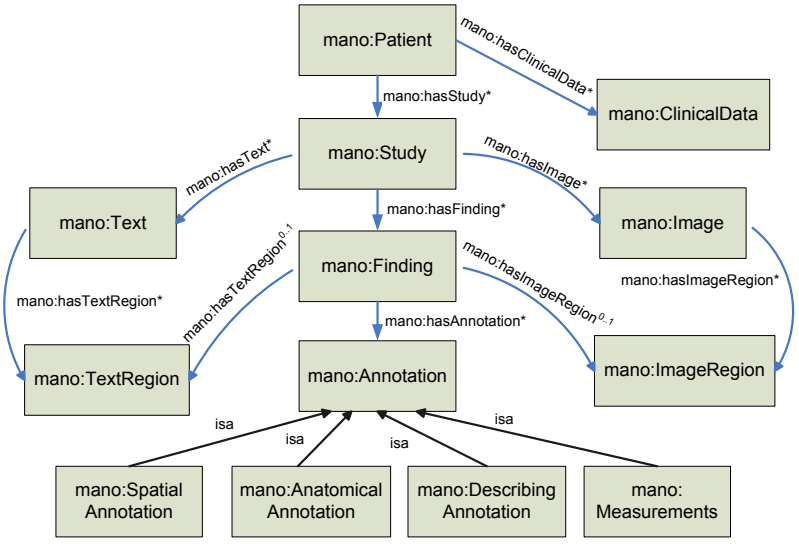


Figure 1. Annotation ontology scheme supporting temporal, multi-modal and text-to-image relations.

The scheme is illustrated in Figure 1 and the design was driven by the following requirements:

- *Link report text passages with related image regions:* Annotations from images and texts must be stored in the same model which should consider the fact that reports summarize annotations from multiple images.
- *Disease progression:* Changes to anatomy due to a pathology over time should be represented. A combined examination of studies with their pre-studies needs temporal relations.
- *Multi modality:* Diagnosis often needs a synoptic view of images acquired with different modalities, e.g., CT, MRI, US. Therefore, the underlying annotation ontology should link annotations not only across time, but also across different modalities.
- In order to adopt hospitals preferred wording, the stack of used ontologies should be extensible, e.g., some of the hospitals have already made experience with SNOMED CT or AIM.[8] Therefore, the annotation scheme should not only incorporate RadLex and FMA but also support further ontologies. For the ontology alignment we developed the KEMM-methodology.[9]

An image region is an arbitrarily shaped spatial sub image which is defined as landmark point, triangulated mesh or image mask. The triangulated meshes are currently used to describe organs, detected by the image parsing system,[7] and image masks to define scribbles.

The class mano:Finding relates anatomical annotations, such as *liver*, *spleen* with the anatomy qualifying describing annotations, such as *enlarged*, *hypodense*, *jagged margin*. Currently, FMA and the anatomical tree of RadLex are used to define the anatomy and the imaging observation and visual modifier trees of RadLex for description. If an anatomical term of a finding is missing in the existing vocabulary, spatial annotations allow the user to paraphrase it with spatial relations such as *nearTo* or *inBetween*, e.g., the lymph node *near to* renal hilus. If the finding is a specific area or volume, we can add a mano:Measurement to store the values of the parameter. All other additional information can be archived by mano:FreeText. To free the user from selecting the right anatomy term, we added a query expansion mechanism which recursively infers sub-classes in FMA. Thus, the sample query *Thoracic lymph node* results in 90 sub-classes:

Thoracic lymph node → Mediastinal lymph node
→ Pretracheal lymph node
→ Esophageal lymph node, ...
→ ...

2.2 Query by Scribble

The goal of MEDICO's visual similarity search is to allow the user to quickly outline a region of interest (ROI) and to ask the system for similar ROIs, without having to take the time for an exact segmentation. We call such a quick ROI specification a *scribble*. We support 3D scribbles, however, in favor of a faster query specification, we expect most queries to be posed as 2D selections. Since 2D image features will have the highest descriptive power due to their maximized image resolution, we represent one 3D annotation by a collection of 2D image features. We treat such a collection like a classical multi-instance problem, where one object is represented by an unknown number of instances of a fixed representation. On the one hand, this causes a loss of information, since we discard the slices' order. On the other hand, the multi-instance perspective allows the comparison of various slice permutations, which can very well contribute to lesion similarity.

In our experiments, we found a combination of grey value histograms and Haralick texture features[10] to perform best for the given problem. For each 2D ROI r_i of the complete region of interest we generate one histogram of 150 bins (HIST) over the given Hounsfield space, as well as a Haralick descriptor for the 9 subwindows of a 3 by 3 grid imposed on each r_i (HAR). Since one Haralick descriptor for 5 different pixel distance values $(1, 3, 5, 7, 11)$ contains $13 \cdot 5 = 65$ statistics, this amounts to descriptors of sizes 100 and $65 \cdot 9 = 585$ for each slice covered by the lesion's bounding box. Finally, we add a third lesion representation (SIZE) representing the extension of the lesion's bounding box in all three dimensions.

For each slice representation j, we define a feature-wise lesion distance d_j on the instances $\{\mathbf{a}_{0,j}, \ldots, \mathbf{a}_{s-1,j}\}$ representing the s slices of a lesion as the Sum of Minimum Distances (SMD), using the Manhattan distance as instance distance on the slices' feature vector representations.

We form the overall distance between two lesions as a weighted sum over the single representations' distances. We require two kinds of weights: s_j, representing the standard deviation of distances of representation j for ensuring a comparable distance scaling. Note that this simple distance joining procedure assumes that the single representations' distances follow comparable distributions. Additionally, the weights w_j represent an actual weighting factor to be assigned to representation j. The combined distance measure d for a set of lesion representations R is: $d_{\text{combined}} = \frac{1}{\sum_{j \in R} w_j} \sum_{j \in R} w_j s_j d_j$.

Even though this distance measure is well-suited for lesion comparison, similarity among medical images remains a difficult application. The appearance of a CT image depends on the setting of the image kernel and the time and kind of the applied contrast agent. In many cases, the latter is not even available to the computer. Therefore, image similarity alone can hardly be a significant indication of a similar patient case.

The MEDICO system thus exploits all available, manually specified and automatically generated meta-information of the query volume for restricting the search space to annotations which are actually relevant. MEDICO can automatically determine the position of the ROI w.r.t. a number of organs and landmarks[6] or within a standardized body atlas,[11] as well as any available information on the patient's prior history in the accessible database collection.

Figure 2 shows an example scribble and the position of a query by scribble in the combined search workflow.

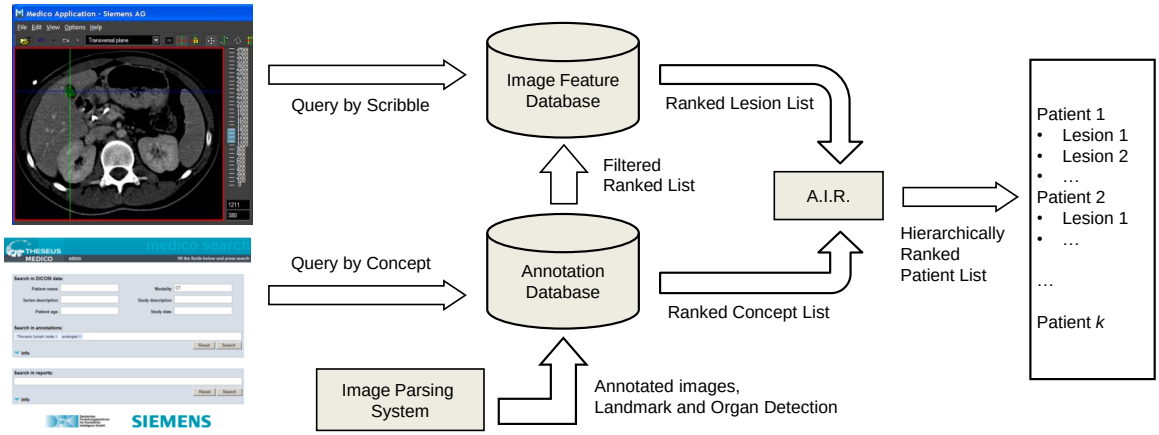


Figure 2. Query by Scribble: retrieve similar regions of interest (ROIs) via a quick selection mark on an ROI. Query by Concept: retrieve image list using semantic filtering criteria, some of which can be automatically generated. Combined Search: use (some of) the output of a Query by Concept as filter list for a Query by Scribble and combine the resulting lesion lists.

2.3 Combined Search

In our example, the *query by concept* and *query by scribble* interact in two ways: first, we restrict the lesion search to the area of the liver which is automatically detected by the MEDICO image parsing system in less than 2 min.[7] This significantly reduces the number of search results and in parallel it increases the search quality since, e.g., lesions in the spleen or the kidneys are a-priori eliminated. The added search for thoracic lymph nodes results in about 90 SPARQL queries due to the built-in query expansion mechanism.

Currently, the result list from semantic search is used as a filter for similarity search (Fig. 3). This not only returns a merged list but also reduces the runtime in querying the *Image Feature Database*.

2.4 Search Infrastructure

MEDICO provides the user with an easy-to-use web-based form to describe a search query. Currently, a search consists of a semantically rich data set composed of DICOM tags, image annotations, text annotations and gray-value based 3D CT images as reference. This leads to a heterogeneous multimedia retrieval environment with multiple query languages for retrieval: DICOM information is stored in PACS systems, image and text annotations are saved in a triple store and the CT scans are accessible by a image search engine performing a similarity search.

Apparently, all these retrieval services are using their own query languages for retrieval (e.g., SPARQL or SQL) as well as the actual data representation for annotation storage (e.g., OWL). Beside all differences, these describe a common (semantically linked) global data set. To fulfill a meaningful semantic search, these interoperability issues had to be solved. Furthermore, it is essential to formulate queries that take the aforementioned diverse retrieval paradigms into account. For this purpose, MEDICO integrates the AIR[12] multimedia middleware framework which implements the MPEG Query Format (MPQF)[13] which is currently the most specific query language for multimedia retrieval. This framework has been especially designed to serve as a mediator between a search interface and an arbitrary amount of backends. AIR is able to support both, distributed query processing as well as local query processing.

The lesion representations used for the visual similarity ranking are all stored in the *Image Feature Database*, which uses SQL tables for enabling a quick retrieval of candidate lesions via the specification of a filter set of candidate volumes provided by the *Annotation Database* with the *Filtered Ranked List*. If no filter is specified, our system supports spatial indexing structures for accelated ranking queries.

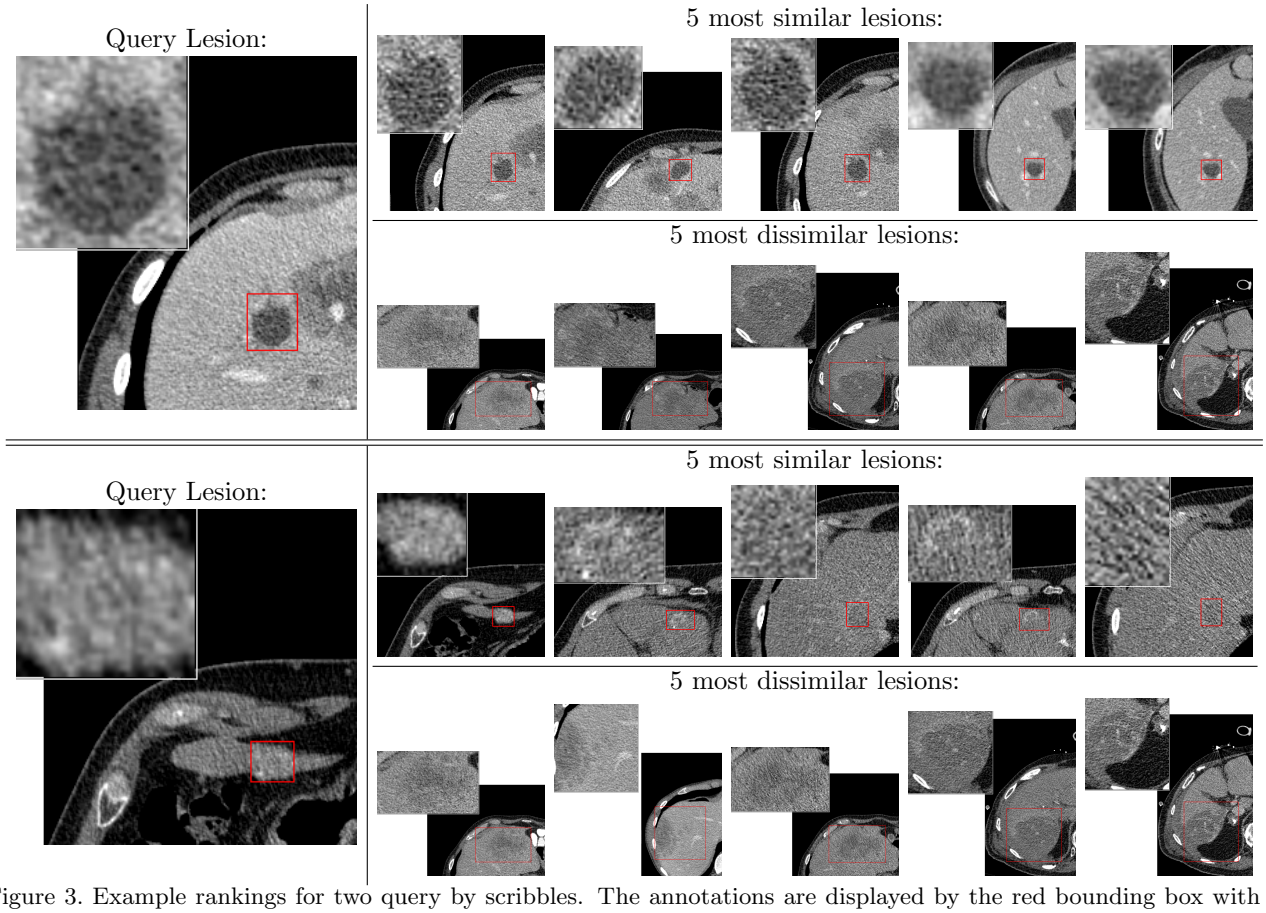


Figure 3. Example rankings for two query by scribbles. The annotations are displayed by the red bounding box with a close-up to the top left. These excerpts only show the center slice of the annotations, which may heavily vary in height.

The *Annotation Database* stores the semantic image and text annotations. It is implemented using a Jena text database (Jena TDB), which directly supports OWL/RDF and SPARQL. We selected the Jena library because of its good scalability and runtime performance.[14] See Figure 1 for the OWL developed to store the semantic annotations.

3. RESULTS AND CONCLUSIONS

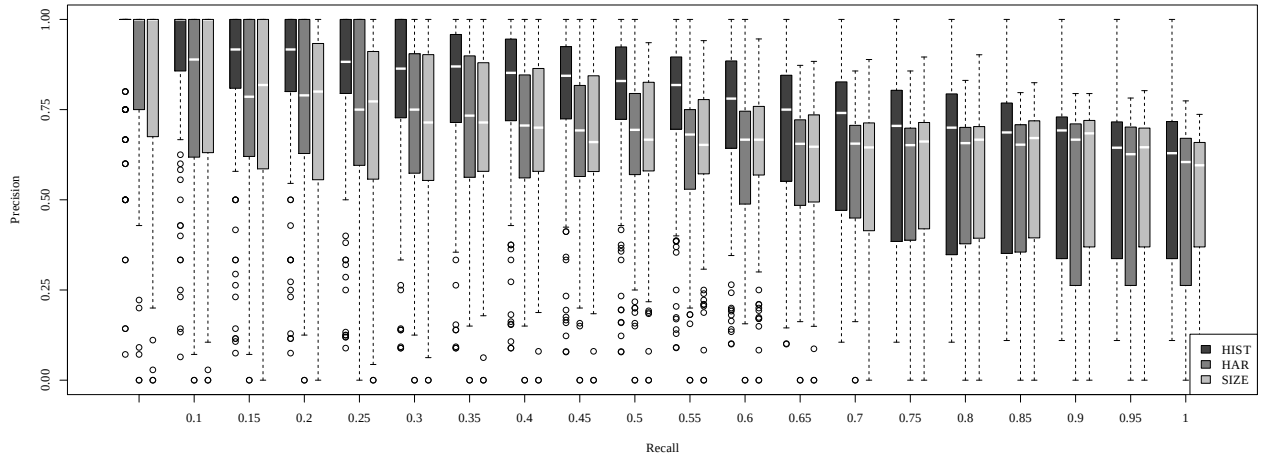
We validated our search procedure with respect to the quality of the visual similarity ranking, as well as w.r.t. the gain achieved by combining the visual query with automatically-derived and manually-specified semantic queries.

3.1 Datasets

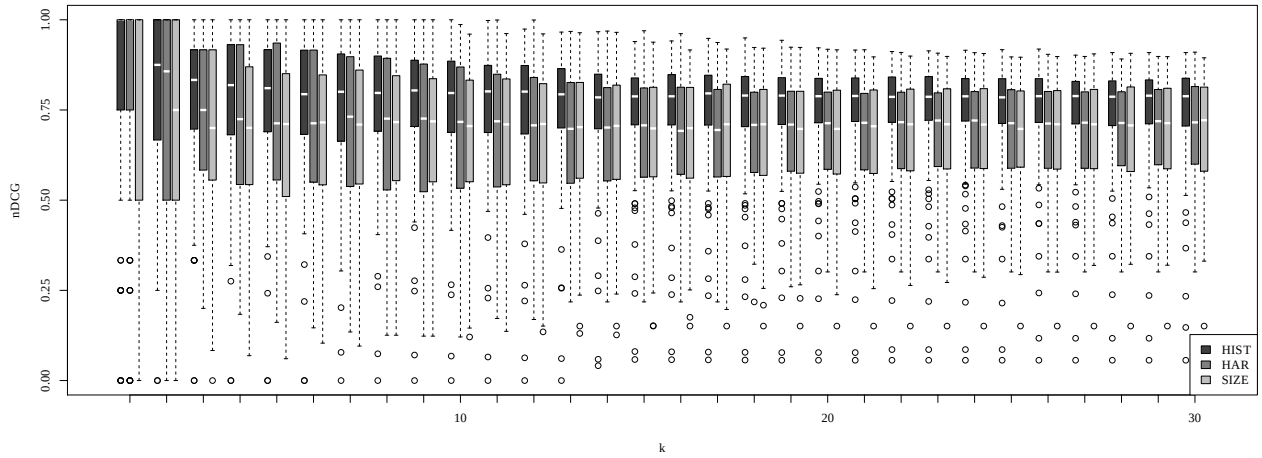
A medical expert annotated the 3D bounding boxes of 1293 lesions (973 liver, 130 spleen, 190 kidneys) in 577 CT scans for 92 patients. For verifying the quality of our visual similarity metric, we selected 111 liver lesions with a bounding box volume $\geq 5\,\mathrm{cm}^3$ as validation set V_1 (79 volumes of 26 patients) and a medical expert annotated them with pair-wise similarity scores on a 5-step scale from 0 (completely dissimilar) to 100 (same lesion).

Furthermore, we extended V_1 by 13 spleen lesions and 62 kidney lesions (all $\geq 5\,\mathrm{cm}^3$) as V_2 for testing the effect of omitting the automatically-derived location knowledge.

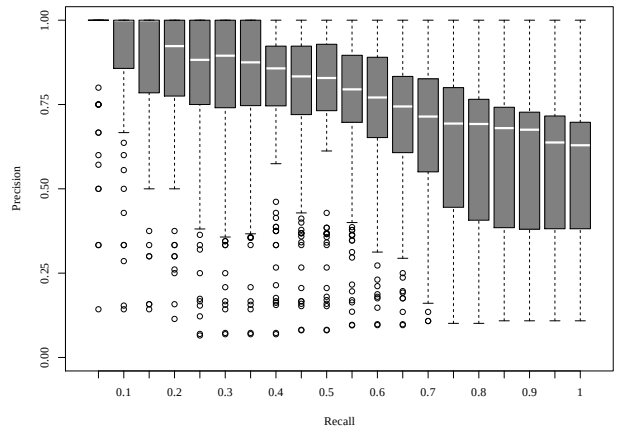
Additionally, our medical experts annotated 100 CT scans as set V_3 in a *semantic reporting process*[6] for visible radiological findings, mapped into the MEDICO-ontology.[3] This set of volume annotations can be queried by advanced semantic queries like *enlarged thoracic lymph nodes*.



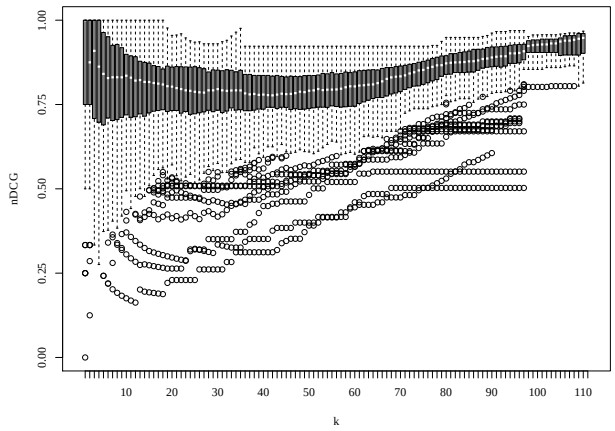
(a) Precision-Recall Curves of single descriptors.



(b) nDCG[15] Curves of single descriptors – for a better visibility, ranking scores for more than 30 objects are omitted.

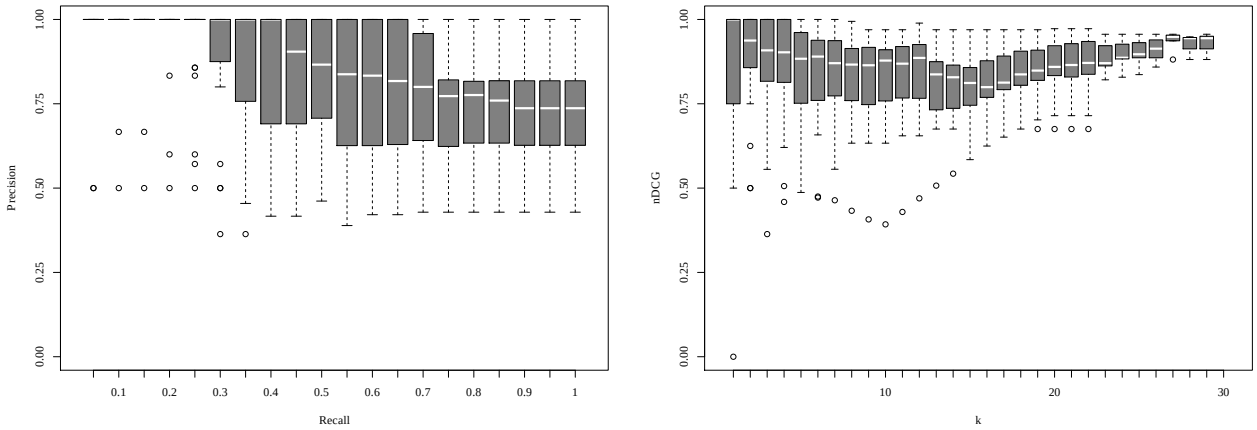


(c) Precision-Recall Curves of combined descriptors.



(d) nDCG[15] Curves of combined descriptors.

Figure 4. Evaluation of the rankings on V_1 via boxplots, displaying the median enboxed by the first and third quantile. The whiskers represent the farthest non-outliers. 4(a) and 4(b) display rankings based on single image descriptors, 4(c) and 4(d) show rankings for the combined distance measure d_{combined}. The single features' distance contributions are weighted HIST : HAR : SIZE $= 3 : 1 : 1$.



(a) Precision-Recall Curves of histogram descriptor.

(b) nDCG[15] Curves of histogram descriptor.

Figure 5. Evaluation of the rankings of the randomly-sampled lesion set V_1' ($\subset V_1$) of size 30 with the same validation setting as in Figure 4 using only grey value histogram features.

3.2 Ranking Performance

In contrast to our fast box-annotation scheme, Napel et.al.[1] proposed a retrieval scheme for liver lesions which requires the exact segmentation of the lesion and an additional, manual specification of 161 semantic properties. They tested their approach on 30 lesion annotations. Our goal is to achieve rankings of a comparable quality over a larger database with a considerably smaller annotation effort (only box annotations, no semantic properties).

In order to minimize the annotation overhead for our large set of lesions we decided to restrict the validation of the visual similarity search to subset V_1 of 111 liver lesions. For every lesion, we generated a ranking of the remaining lesions according to their automatically-determined visual similarity. The first and the last ranked lesions of two such example rankings are depicted in Figure 3. In both cases, the first (and best) match in the top row is actually the same lesion, only originating from CT scans taken on another day.

Figure 4 shows the precision-recall curves (a pair is considered to be relevant for a similarity score ≥ 75) and the normalized discounted cumulative gain (nDCG)[15] aggregated over the complete set of 111 lesions. 4(a) and 4(b) display the performance of the single features, whereas 4(c) and 4(d) validate distance measure d_{combined} based on all three representations. The combination of the slice-wise grey-value histogram features (HIST), and a haralick pyramid kernel (HAR) with the simple size measure SIZE results in a mean average precision of 0.74 and an average nDCG value for the 10^{th} retrieved lesion of 0.82.

This ranking does not completely reach the quality of the validation results by Napel et.al.,[1] however, this is due to the rougher annotation quality and due to the larger dataset (111 instead of 30 lesions). When restricting our dataset to a randomly chosen subset V_1' of 30 lesions, the mean average precision increases to 0.89 and the average nDCG value for the 10^{th} retrieved lesion becomes 0.85 only using grey value histogram features. The corresponding validation plots are displayed in Figure 5.

3.3 Benefit of the Combined Search

The quality of the above results gained a lot from our information combination approach. The information about the scribble's anatomic position enables to exclude all entities from the search space which are not localized within the liver. To test our hypothesis, we generated rankings on the dataset V_2 containing an additional set of 75 lesions in the spleen and the kidneys. When querying V_2 without using the semantic information about the organ context of the query lesion, 12 % of the top ten hits originate from foreign organs (cf. Table 1). The miss-placed lesions appear to be similar for the image descriptors, but they are not useful in the context of a lesion query. This is a major advantage of the MEDICO query system in comparison to other retrieval systems, where this

Table 1. Confusion matrix of ranking test on V_2 (62 kidney, 111 liver and 13 spleen lesions) and percentage of hits matching the query organ in the 10-nearest neighbors (excluding the query). 12 % of the top 10 hits are from a foreign organ.

	10-nearest neighbors			
	Kidneys	Liver	Spleen	Total
Kidneys	560	55	5	62
Liver	25	1057	28	111
Spleen	3	110	17	13

	10-nearest neighbors [%]			
	Kidneys	Liver	Spleen	Total
Kidneys	90.3	8.8	1.1	62
Liver	2.3	95.2	2.5	111
Spleen	2.3	84.6	13.1	13

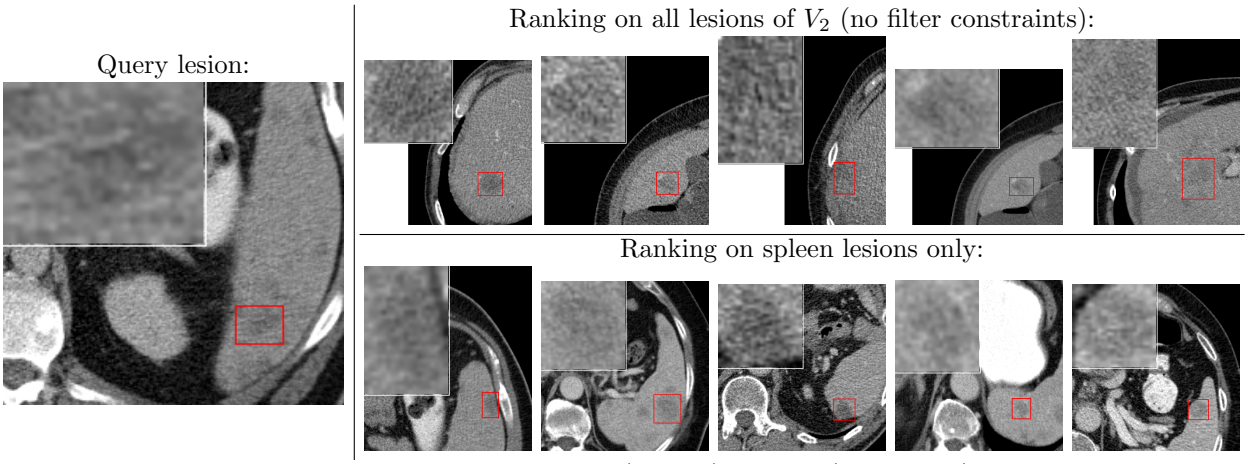


Figure 6. Example rankings for a spleen lesion query without (top row) and with (bottom row) organ constraints. Ranking all of V_2 returns only liver lesions in the top 5 hits.

information has to be filled in manually. The effect is exemplified by Figure 6, where similar spleen lesions will only appear in the top 5 ranks when applying an intelligent context filter.

The MEDICO system furthermore allows to specify manual semantic queries. In our example case the user wants to see all patients with enlarged thoracic lymph nodes. This query can be posed to the set of 100 semantically annotated volumes V_3 and it matches 34 patients in the *Annotation Database*. 10 patients have assigned lesion annotations and 9 of these patients show a total of 35 liver lesions in 26 volumes. The dataset V_1 can thus be restricted to a set of 35 instead of 111 liver lesions by requesting a similar patient history.

Besides the obvious benefit of restricting the result set to semantically valid items, the combination with semantic filter properties also speeds up the visual similarity ranking. A single query to V_2 takes 1330 ms when

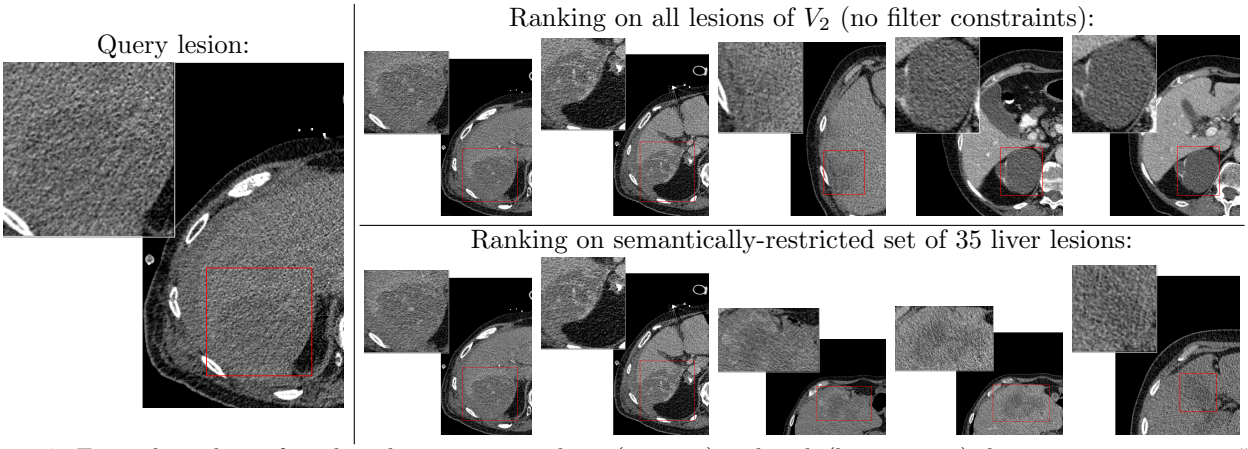


Figure 7. Example rankings for a liver lesion query without (top row) and with (bottom row) the semantic constraint "in liver, thoracic lymph nodes enlarged". The first two hits of both rankings each show the query lesion in various stages. When all of V_2 is queried, two kidney lesions are among the top 5 hits.

the database is not cached for a quick main memory retrieval, including the time required for generating the query lesion's features (266 ms). The same query takes only 1033 ms when adding the organ information "liver" (kidneys: 551 ms, spleen: 296 ms). When restricting the context to patients with enlarged lymph nodes, one query only takes 675 ms.

Naturally, the query process can be greatly sped up by caching the query database, however, in an environment not yet prepared for large-scale main memory storage, this procedure would interfere with other services. Thus, an intelligent filtering of the query database is an important step for a well-performing similarity query.

3.4 Outlook

We presented a comprehensive search framework which enables the user to accomplish visual similarity search combined with semantic search based on web 3.0 technology. This significantly extends the search capabilities compared with currently available content-based image retrieval systems and enables the system to answer real-world questions.

In our sample query, the total result set reduced by 12% taking the semantic information about the containing organ of the lesion into account. Another reduction is achieved from 111 to 35 results applying the semantic search criteria *thoracic lymph nodes enlarged*. With that, a meaningful result set for the given query is returned which from the physicians perspective shows similar patients having similar pathology located in the same anatomy.

In future work we aim towards improving the quality of the image-based query component query by scribble by testing further image descriptors and by incorporating a lesion segmentation step for detailing the imprecise box annotations. Furthermore, we will look for ways of refining the query combination mechanism and we plan to test our framework on larger sets of annotated data.

ACKNOWLEDGMENTS

MEDICO is part of the THESEUS Program funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. Thanks also to Mario Döller, Passau University, for his support in MPQF.

REFERENCES

[1] Napel, S. A., Beaulieu, C. F., Rodriguez, C., Cui, J., Xu, J., Gupta, A., Korenblum, D., Greenspan, H., Ma, Y., and Rubin, D. L., "Automated retrieval of CT images of liver lesions on the basis of image similarity: Method and preliminary results," *Radiology* **256**(1), 243–252 (2010).

[2] Allampalli-Nagaraj, G. and Bichindaritz, I., "Automatic semantic indexing of medical images using a web ontology language for case-based image retrieval," *Eng. App. of Artificial Intelligence* **22**(1), 18–25 (2009).

[3] Möller, M. and Sintek, M., "A generic framework for semantic medical image retrieval," in [*Proc. of the Knowledge Acquisition from Multimedia Content (KAMC) Workshop, 2nd International Conference on Semantics And Digital Media Technologies (SAMT)*], (November 2007).

[4] Rosse, C. and Mejino, J., [*Anatomy Ontologies for Bioinformatics: Principles and Practice*], vol. 6, ch. The Foundational Model of Anatomy Ontology, 59–117, Springer (December 2007).

[5] Langlotz, C. P., "RadLex: A new method for indexing online educational materials," *RadioGraphics* **26**, 1595–1597 (2006).

[6] Seifert, S., Kelm, M., Moeller, M., Mukherjee, S., Cavallaro, A., Huber, M., and Comaniciu, D., "Semantic annotation of medical images," in [*SPIE Medical Imaging*], (2010).

[7] Seifert, S., Barbu, A., Zhou, K., Liu, D., Feulner, J., Huber, M., Suehling, M., Cavallaro, A., and Comaniciu, D., "Hierarchical parsing and semantic navigation of full body CT data," in [*SPIE Medical Imaging*], (2009).

[8] Channin, D. S., Mongkolwat, P., Kleper, V., and Rubin, D. L., "The Annotation and Image Mark-up Project," *Radiology* **253**(3), 590–592 (2009).

[9] Wennerberg, P., Zillner, S., Mller, M., Buitelaar, P., and Sintek, M., "Kemm: A knowledge engineering methodology in the medical domain," in [*Proc. of the 5th International Conference on Formal Ontology in Information Systems (FOIS)*], (2008).

[10] Haralick, R. M., Shanmugam, K., and Dinstein, I., "Textural features for image classification," *IEEE Transactions on Speech and Audio Processing* **3**(6), 610–623 (1973).

[11] Emrich, T., Graf, F., Kriegel, H. P., Schubert, M., Thoma, M., and Cavallaro, A., "CT slice localization via instance-based regression," in [*SPIE Medical Imaging*], 762320 (2010).

[12] Stegmaier, F., Döller, M., Kosch, H., Hutter, A., and Riegel, T., "AIR: Architecture for Interoperable Retrieval on Distributed and Heterogeneous Multimedia Repositories," in [*Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*], 1–4 (April 2010).

[13] Döller, M., Tous, R., Gruhne, M., Yoon, K., Sano, M., and Burnett, I. S., "The MPEG query format: On the way to unify the access to multimedia retrieval systems," *IEEE Multimedia* **15**(4), 82–95 (2008).

[14] Stegmaier, F., Gröbner, U., Döller, M., Kosch, H., and Baese, G., "Evaluation of Current RDF Database Solutions," in [*Proceedings of the 10th International Workshop on Semantic Multimedia Database Technologies (SeMuDaTe), 4th International Conference on Semantics And Digital Media Technologies (SAMT)*], 39–55 (December 2009).

[15] Järvelin, K. and Kekäläinen, J., "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.* **20**(4), 422–446 (2002).