Review of dissertation proposal

Proposal title: **Improving Quality of Content-Based Image Retrieval**
PhD candidate: **Petra Budíková**

Proposal content:

The subject of dissertation proposal is generally the content-based image retrieval within the framework of the metric similarity search model. The author covers the topic widely, including efficiency and effectiveness problems, that is, how to search an image database quickly and how to search the database such that the user is satisfied. The proposal is structured into 4 sections, where the first one introduces into the problem of image retrieval, while the last one details the thesis plan. The middle two sections provide a kind of survey over various problems within the whole topic, namely, the metric space model and its role in indexing and search, the search task and similarity function definition, the presentation/postprocessing of query results, the relevance feedback, and finally similarity query languages.

Review:

The proposal is well written, in fact, it could be used as a concise survey on the entire topic of similarity search in image databases. Also the language and presentation style are above the standard. The scope and details presented in this survey make evident that the author has a wide and deep enough knowledge of the topic, and that the thesis that is about to follow will be of a high quality. I have only one serious comment – it seems to me that the author tries to shoot three birds with one stone. I believe that three objectives for a PhD thesis (here postprocessing, multi-queries, and query languages) is a bit too much to be investigated in a satisfactory detail in only 2.5 years. I informaly recommend the author to choose only one of the objectives (probably the postprocessing/ranking?).

Minor comments, hints:

1) In Section 4.2, to prevent „reinventing wheel", please read the papers about metric cache, as your idea appears very similar.

   F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti, "A metric cache for similarity search," in LSDS-IR '08: Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval. New York, NY, USA: ACM, 2008, pp. 43–50.

   F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti, „Caching content-based queries for robust and efficient image retrieval," in EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology. New York, NY, USA: ACM, 2009, pp. 780–790.

2) Concerning efficient processing of multiple queries (Section 3.1.1), consider also an earlier approach

   B. Braunmuller, M. Ester, H.-P. Kriegel, and J. Sander, "Multiple similarity queries: A basic DBMS operation for mining in metric databases," IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 1, pp. 79–95, 2001.

3) Concerning multi-object queries (Section 4.2), consider also the metric skyline concept.

   L. Chen and X. Lian. Efficient processing of metric skyline queries. IEEE Trans. on Knowl. and Data Eng., 21(3):351-365, 2009.

4) Section 1.1 – publications [51] and [45] are a bit outdated (10 and 8 years).

5) Section 2.1 – you forgot reflexivity d(x,y) = 0 <=> x=y, which is more important than it appears :-)

6) Section 3 – please, do not interchange top-k queries and k-NN queries. While the latter one is regular query (as meant in this section), defined by a query point and an extent k, the top-k queries is rather an operator over multiple database sources. In your case the database sources could be the results of previous queries and the top-k operator a means of reranking/aggregation with respect to some additional criteria.

In the following, I comment my impression from the area of similarity search applied to image databases. I do intent these comments as an input to the discussion during the defense, so do not consider them as a critique. They are more or less "philosophical" concerns on the entire topic of content-based similarity search in image databases.

7) On various places (Sections 1.1, 3.1.2, 3.3) the author talks about learning the similarity (by means of various techniques, as machine learning, user feedback, etc). Here it should be emphasized that, since we are in metric space model, the learning/user feedback is allowed to affect just some weights of the underlying metric or the linear combination of partial distances. Otherwise, if we admit more „aggressive" learning, the similarity would likely become a nonmetric. For example, in section 3.1.2 you mention an approach [49] that considers only certain portions of the compared objects, such that the result distance is minimal. Such an approach usually leads to nonmetric behavior (the triangle inequality is not satisfied). For more on nonmetric similarity, see

   T. Skopal, Unified framework for fast exact and approximate search in dissimilarity spaces. ACM Transactions on Database Systems 32(4):1-46. 2007.

8) The previous comment leads to a related one. If you improve the entire similarity model (also including the object description) to become more effective, you have two possibilities. You either leave the object representation as is (e.g., the five MPEG7 features) and change the similarity measure, or you change the objects representation, or both. Here we have to think about the purpose of modeling image retrieval systems in metric space. In the vast majority cases (and CoPhIR database is not an exception), the semantics of the model is hidden just in the image representation. Typically, the features that form the image are high-dimensional vectors. The similarity measure is then „stupid" Euclidean or L1 distance (of linear complexity). So when talking about the semantic gap, we should address also this discrepancy – we assume general metric space model, but any semantic knowledge is put into the vectors, anyways. So why all the noble statements concerning general metric space, when L2 (or Lp at the best) is always the standard? Is there a possibility to put some of the semantics also into the similarity/distance measure, making it thus more expensive but smarter? Or is the distance „sentenced" to be the last and most trivial part of the model... On the other

hand, if we try to use smarter similarity measure, we may often obtain a nonmetric distance (e.g., like the dynamic time warping distance when generalizing L2 to tackle with time-local shifts and stretches in vectors/time series).

9) Finally, following the previous two comments, we move to another related concept. At the end of section 3.2.2 you talk about cheaper similarity measure used for global search and then more expensive one for the additional reranking of query result. Here it would be interesting to investigate whether it is better to model just one expensive similarity measure and perform an approximate search having query with small selectivity (e.g. 10NN), or (as you mention) perform exact search with cheap distance having query with large selectivity (say 1000NN) and subsequent reranking of the result by an expensive similarity. Moreover, unlike the former approach, the latter one could be enhanced by searching by nonmetric measure because the reranking is done sequentially in the small result set.

Conclusion:

In summary, the PhD candidate proved she is able to submit an excellent PhD thesis in a foreseeable time. Although the candidate's research record is not very large up-to-date, the following 2.5 years give enough time to disseminate the results prior to the actual PhD thesis submission.

I definitely **recommend** the proposal to be accepted as dissertation proposal, as well as the rigorous thesis.

In Prague, 15th of March, 2010

Doc. RNDr. Tomáš Skopal, Ph.D.
reviewer