Indexing and retrieval of interpreted video contents

OTKA PD-83438, 2011.01.01.-2013-12.31.
Final Scientific Report

The main focus of this research work is on indexing, retrieval and visualization schemes for content-based retrieval over video pools, concentrating on feature extraction, indexing structures, automatic feature selection, descriptor evaluation, and visualization. The overall goal is to produce new theoretical and practical results in video analysis and content-based indexing and retrieval.

Short project and progress overview

Generally, the work performed followed the original work plan, including dissemination at conferences and in scientific journals, where the goals set for the 3 years have been achieved (with a second submitted scientific journal article accepted with major revisions [R] on Jan. 27, 2014). As of yet, a total of 10 project-related papers have been published [1-10]. During the project I have been working with 1 PhD student (research and publications) and 2 regular students (programming-related tasks). Regarding financing, total spending remained below the planned budget.

Presentation of scientific results

In the following the main scientific results produced during the project will be presented, which cover for main, interconnected areas: i). evaluation of content descriptors and automatic feature selection, along with a descriptor- and data-independent ranking framework to aid the selection of best performing features for any given dataset; additionally, a data- and descriptor-independent parallel multi-tree indexing method was created, for indexing and querying large datasets with automatic feature selection from an arbitrarily large pool of features; ii). automatic and lightweight pattern and shape segmentation and recognition for mobile devices; iii). visual detection, recognition and tracking of moving airborne targets; iv). query and retrieval result visualization for image and video content based retrieval engines along with a web-based evaluation, annotation and visualization framework.

i. Descriptor evaluation and feature selection

As one of the most important results of the project, a new approach has been developed for descriptor evaluation and feature selection, based on graph structure analysis [7, 5, 3]. We investigated and laid the foundations of an automatic image and video feature descriptor evaluation framework, based on several points of view. First, evaluation of distance distributions of images and videos for several descriptors were performed, then a graph-based representation of database contents and evaluation of clique formation the appearance of the giant component was performed. The goal was to design an evaluation framework where different descriptors and their combinations can be analyzed, with the goal of automatic feature selection. Starting from random geometric graph structure analysis based on the works of Erdős-Rényi, we built and analysed descriptor graph behaviour in graphs where nodes are

database elements (i.e. videos and images) and vertices are weighted based on the distances of the elements based on different descriptors and their difference metrics. The basic idea is that descriptor performances can be evaluated based on their clustering properties, which in our case translates to the behaviour of the descriptor graphs in the process of the appearance of the so called giant component. This is also a heavily investigated generic graph analysis field, since there are no known thresholds or limits regarding the estimation of the point of appearance of the giant component in any geometric graph. The novelty of this approach lies in that we do not follow the generic path of building a graph then try to cluster the graph regions, but we approach from the opposite end by investigating the graph building process itself and use this information not to partition the graphs, but to try to rank the descriptors which are being used to build the graph itself. Thus, we analyse the graphs of different descriptors to find the point in the process (i.e. phase transition Fig.1) where this component appears (Fig.2), which will be different for each descriptor, and use this information to produce a rank of the investigated descriptors. Then, this rank is used in the content-based retrieval process to produce results with higher relevance to the query. We also created a fitness function (Eq. 1) which we use as the basis of descriptor ranking (Fig.3(a)), and which takes into consideration the critical edge weight where the giant component appears (w_{crit}), the sizes of the largest and second largest components at this point (C_2/C_{max}), and the number of components ($nrcomp$) at the point of appearance of the giant component. Also, numerical evaluation results based on this fitness function have been presented, along with visual retrieval results which show the viability of this approach.

$$F(\cdot) = w_1 \cdot w_{crit} + w_2 \cdot \frac{|C_2|}{|C_{max}|} + w_3 \cdot \frac{nrcomp}{n} \quad (1)$$

A parallel data- and descriptor-independent indexing scheme has also been developed [1] providing a flexible and modular indexing scheme for evaluating large descriptor sets with an associated result ranking retrieval step for servicing high-precision multi-feature content-based queries. Evaluations on multiple datasets shows fast indexing times, and competitive retrieval capabilities as well (Fig.3(b-c)).

A visual showcase of how the introduced feature selection and descriptor ranking method can benefit retrievals on large video pools is shown in Fig.4, which shows how the precision and relevance of the retrieved results increases if the best performing descriptors are selected and used from an arbitrary large set of descriptors.

A scientific journal paper detailing all the main finding of this part of the research is in progress [R].

ii. Automatic lightweight pattern and shape recognition

Research has been performed regarding the possibilities of content based pattern recognition and retrieval on mobile devices, with the goal of performing computations locally, with network non-dependence [9, 6]. The main goal for such methods are lightweight implementation on mobile devices locally, not relying on network connection or server side processing (Fig.5(a)). In such circumstances the focus needs to be on effectiveness and simplicity in a constrained hardware environment, while still preserving high level of

functionality (i.e. good recognition rates, Fig.5(b)). Application areas involve offline object recognition and template matching (e.g. for authorization and blind aid applications), and various object categorizations either in pre-processing or in full local processing. The shape extraction process is based on a robust, extended Harris detector, while the recognition process uses a custom index scheme. The recognition process here is transformed into a high-precision retrieval task, and Fig. 5(b) shows evaluation results proving that our retrieval method can always produce high precision (i.e. recognition) rates.

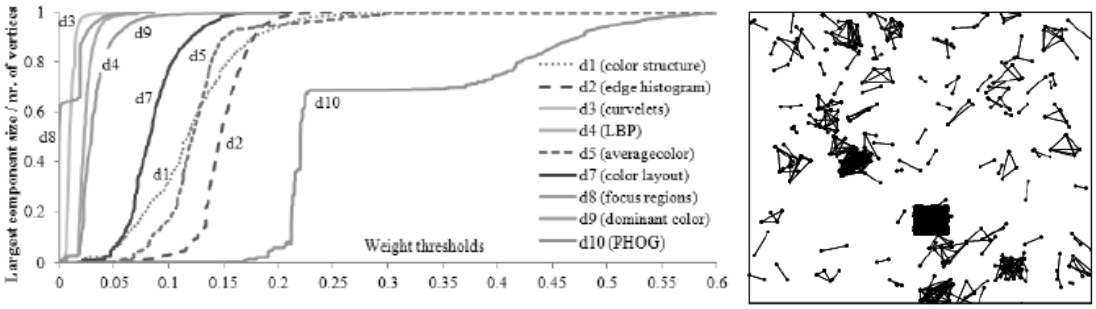


Fig. 1: Left: A phase transition graph for multiple descriptors over a dataset, showing the behaviour of graphs components w.r.t. changing weight thresholds in the retrieval process. Right: A snapshot of a graph structure during the evolution of the components.

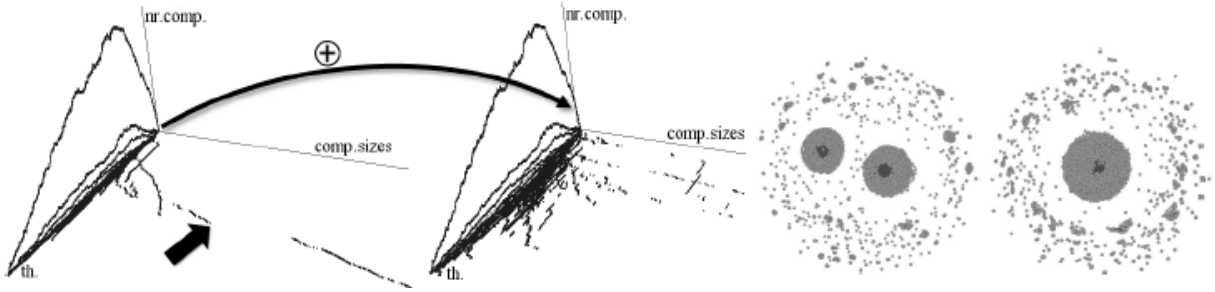
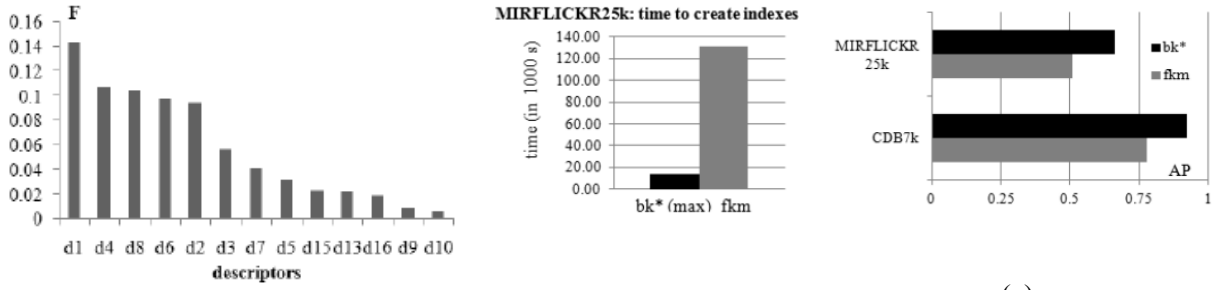


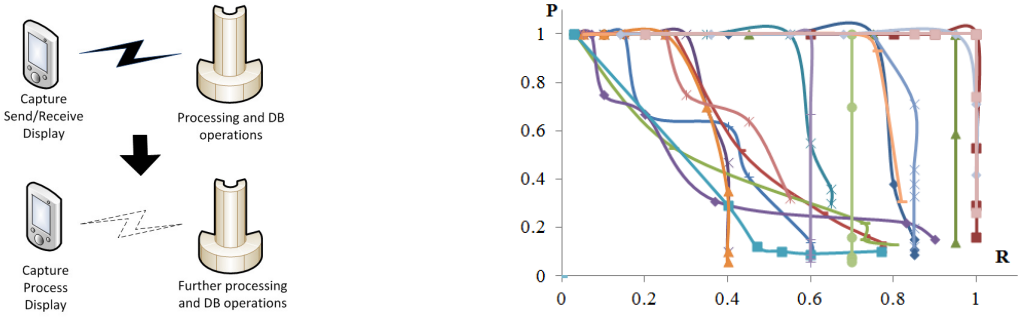
Fig. 2: Left: 3D visualization of the evolution of graphs for homogenous texture descriptor according to changing weight thresholds (*th*), existing component sizes at a given threshold (*comp.size*) and the number of such sized components (*nr.comp.*). Small black arrow points to the area of interest for the critical weight. Right: Component structure right before and at the GC formulation.



(a) (b) (c)

Fig. 3: (a) Descriptor ranking. (b) Fast index-building for the introduced scheme (*bk**). (c) Average precision values for different datsets.

Fig. 4: Retrievals (top: using full descriptor set, bottom: after descriptor ranking) using the descriptor ranking procedure produce higher precision results (top-left image is the query video's representative frame, others are results).



(a) (b)

Fig. 5: (a) Top: online: capture data on the device, send, wait for results, then receive and display. Bottom: offline: capture, process and display on the device, and only send upstream if necessary. (b) Precision/Recall curves for evaluated queries.

Related results were introduced in [2] where we presented a method for local processing of image contents and associated sensor information on mobile devices. The goal was to lay the foundations of a collaborative multi-user framework where ad-hoc device groups can share their data around a geographical location to produce more complex composited views of the area, without the need of a centralized server-client architecture. Sensor (camera, image) locations were processed by analysing so called *vision graph* structures (Fig.6), clustering related images using device sensor information and image contents, to find images with similar views. Then, as a showcase, composite views, warped images and quasi-panoramas were produced to show the viability of the approach (Fig.7).

Fig. 6: Location of sensors (cameras, images) (left); built and clustered vision graph structure (middle); same graph structure also showing the orientation and location of the images (right).

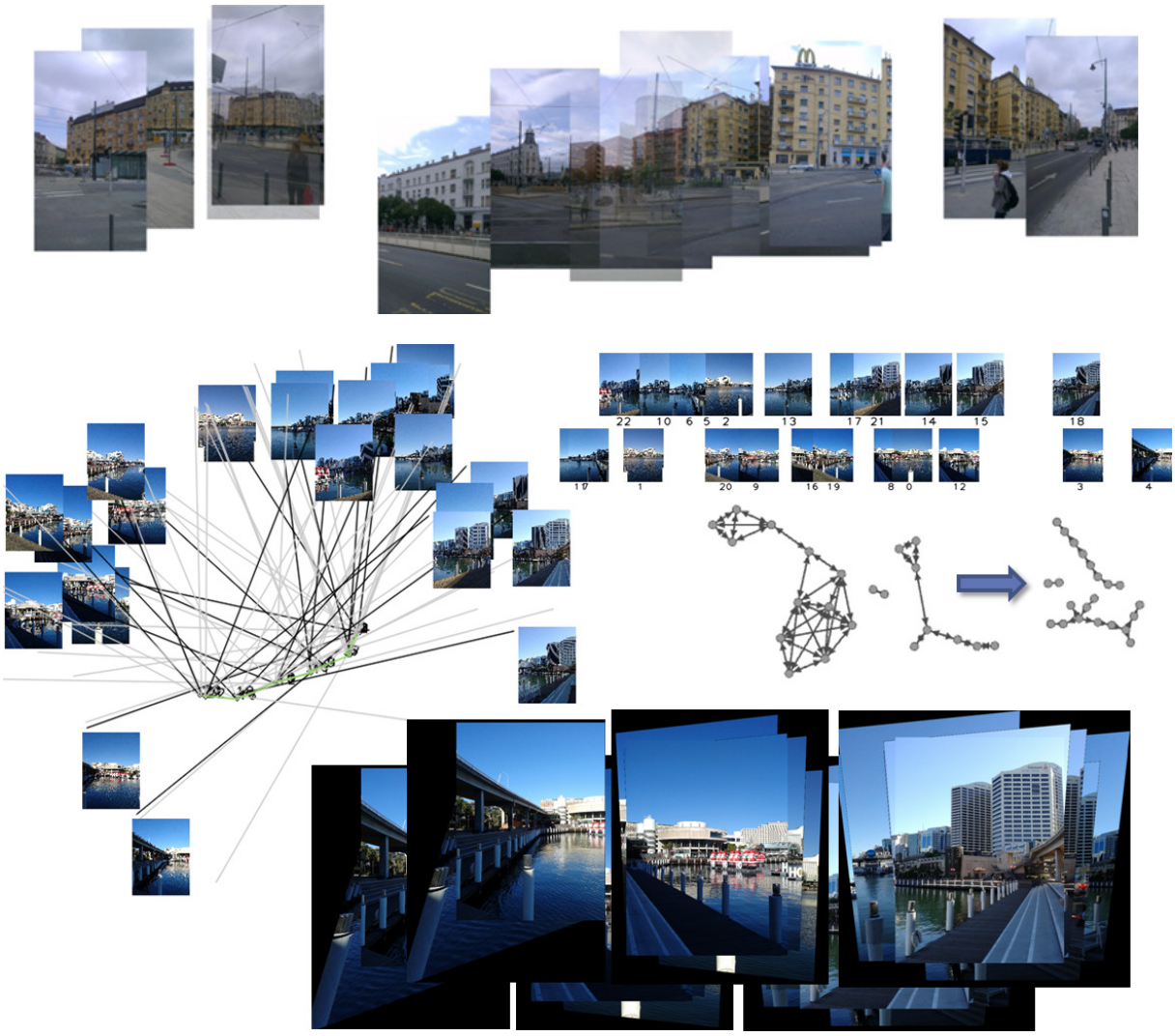


Fig. 7: Top: Images placed in clustered groups according to image contents device sensor information (Móricz Zs. square Budapest). Below: Graph structure, image relations and composite views at another location).

iii. Visual object detection and recognition

Connected to recognition/retrieval topics, a visual detection and recognition method for flying targets was created, based on automatically extracted shape and object texture information, for recognition and tracking tasks [10, 4]. Target objects are extracted based on robust background modelling and a novel contour extraction approach, and object recognition is done by comparisons to shape and texture based query results on a real dataset, with moving targets and moving cameras. Application areas involve passive defence scenarios, including automatic target detection and tracking with cheap commodity hardware, and also generic object detection and tracking based on fused shape and texture features. Novelties were introduced in the object detection and shape extraction process by using a new background and foreground modelling based on Gaussian mixture models combined with a Markov model in a Bayesian framework. Here, a global background modelling approach was fused with a per-object local background update process and localized foreground extraction to segment moving objects in a robust method that automatically rejected non-object image clutter

(Fig.8). For recognition, shape and texture features have been used, and evaluations were carried out on an approx. 9000 shape dataset in 26 target classes that we collected from real video footage. This work has also been the basis of research in retrieval based on multiple features in the feature selection and multi-tree indexing-retrieval framework discussed earlier.

Fig. 8: Objects detected robustly over dynamic background (clouds, vapour trails, lighting).

The recognition part of the presented method employs a similar tree-based indexing-retrieval scheme in [1, 5], resulting in lightweight real-time recognition which we created to run in parallel with detection and tracking, and to not require the processing of all video frames, only a fraction of them, and still produce high recognition rates (Fig.9).

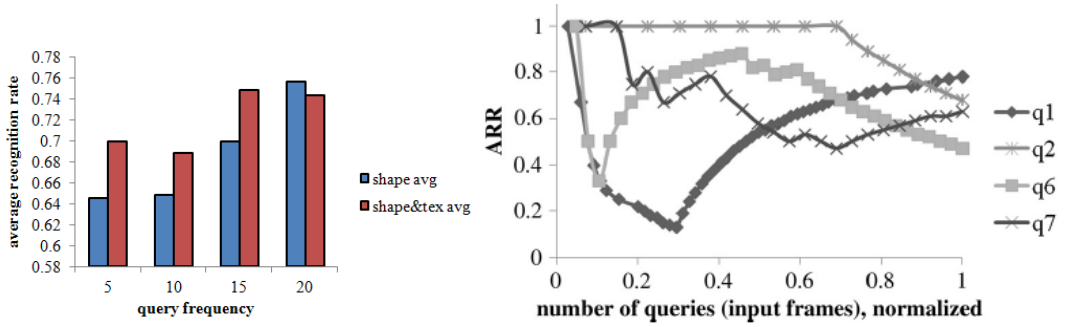
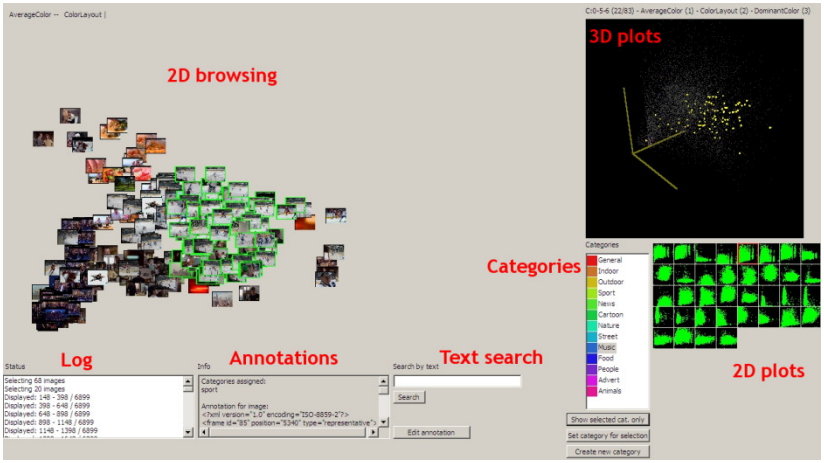


Fig. 9: Left: Recognition rates using shape and fused shape+texture features w.r.t. the frequency of processed video frames. Right: Average recognition rates improve in time as more instances of the same target are observed.

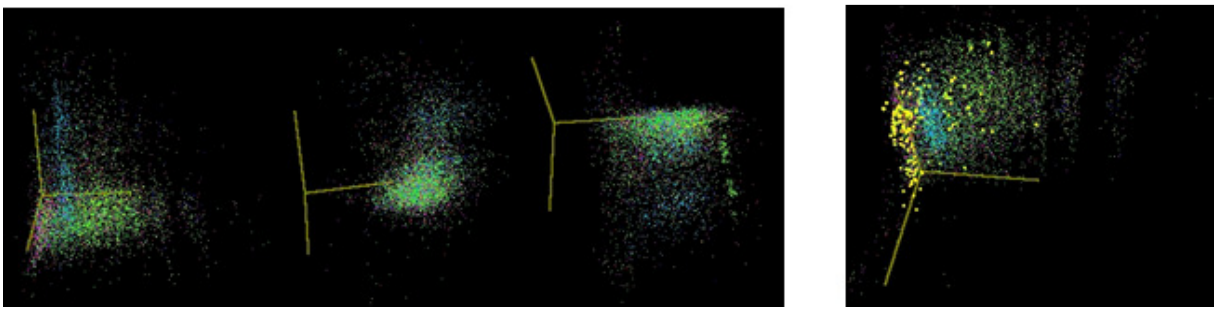
iv. Visualization of dataset distributions, queries and results

A query, search and result visualization application was created which is interactive and flexible, usable for image and video retrieval applications [8] (Fig.10). The main novelty of our approach is that, at the same time, it provides a text and model based search interface, a visual browsing interface, a distribution visualization interface based on a number of content based features, an annotation editing interface and a content classification interface, all combined together in an easy to use prototype. This framework has been used as a precursor for the automatic descriptor evaluation framework described earlier.

Later during the project, a web-based application framework (Fig.11) for more complex video indexing, querying, annotation and visualization tasks has been developed that aided us in evaluating the automatic descriptor evaluation and feature selection method that was developed. Although this framework has not been part of a publication yet, we use it internally continuously.



(a) Main view of the query-, result- and dataset-visualization application.



(b) 3D views of dataset distributions based on arbitrarily selected descriptors.

Fig. 10: Elements of the visualization framework for querying, viewing results, and visual selection of descriptors and data subsets.

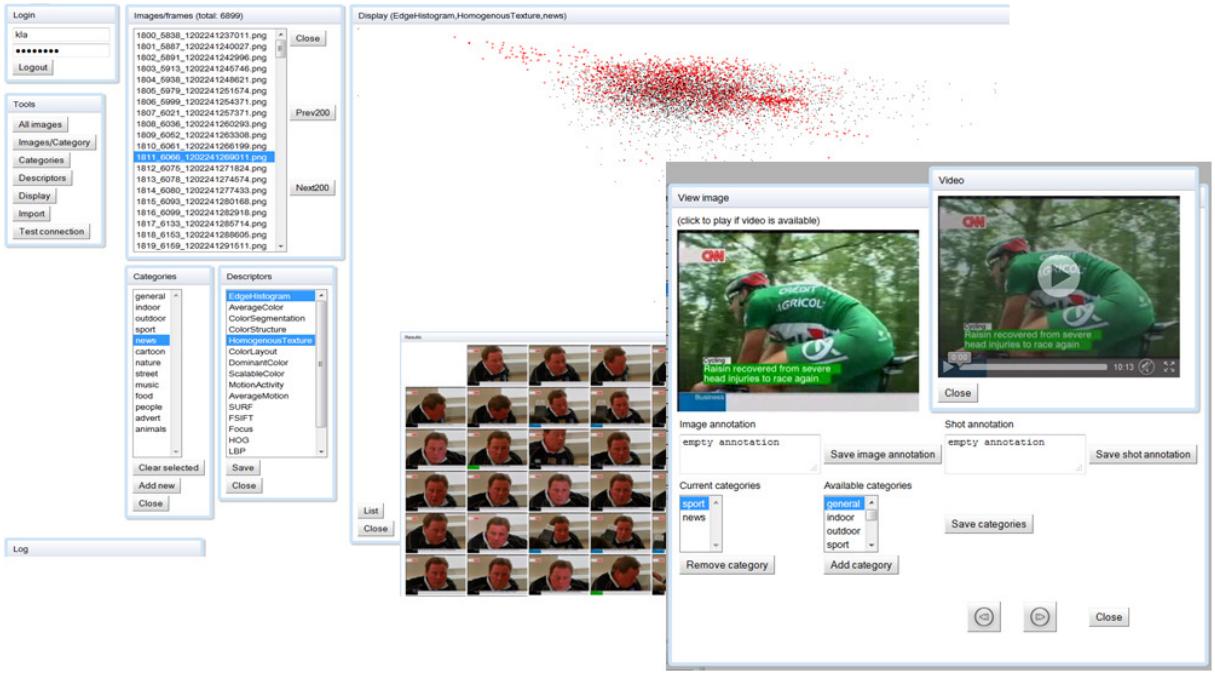


Fig. 11: A web-based application usable for query selection, descriptor selection, annotation of images and videos, playing of video queries and results, visualization of results and data distributions.

Conclusions

The main planned topics - feature selection, visualization and indexing schemes - planned in the original work plan have been addressed during the three years of the research, with results published in conference and journal papers. Work regarding the tackled areas, especially descriptor evaluation and feature selection, continues beyond the grant period, having already produced results that proved to be both interesting and useful. I would like to take the opportunity to thank the Hungarian Scientific Research Fund for their support.

Publications

[R] L. Kovács, A. Keszler, T. Szirányi: Phase Transition and Component Evolution in Descriptor Graphs, submitted to Elsevier Digital Signal Processing (http://www.journals.elsevier.com/digital-signal-processing/), Status on Jan. 27. 2014: revised version submission in progress.

[1] L. Kovács: Parallel Multi-Tree Indexing for Evaluating Large Descriptor Sets, in Proceedings of IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 173-178, doi:10.1109/CBMI.2013.6576581, 2013.

[2] L. Kovács: Processing Geotagged Image Sets for Collaborative Compositing and View Construction, in Proceedings 2013 IEEE International Conference on Computer Vision Workshops (IEEE Intl. Workshop on Computer Vision for Converging Perspectives in conjunction with ICCV 2013), pp. 460-467, doi: 10.1109/ICCVW.2013.67, 2013.

[3] L. Kovács, A. Keszler, T. Szirányi: Óriáskomponensek megjelenése képi leírók gráfjaiban, és alkalmazásuk leírók kiválasztásában, in Proc. of KÉPAF 2013 – Képfeldolgozók és Alakfelismerők Társaságának 9. országos konferenciája, pp. 475-482, 2013.

[4] L. Kovács, A. Kovács, Á. Utasi, T. Szirányi: Flying Target Detection and Recognition by Feature Fusion, Optical Engineering, SPIE, Opt. Eng. 51 (11), pp. 117002-1-13 (November 02, 2012); doi: 10.1117/1.OE.51.11.117002, 2012.

[5] A. Keszler, L. Kovács, T. Szirányi: The Appearance of the Giant Component in Descriptor Graphs and Its Application for Descriptor Selection, in Proc. of CLEF (Conference and Labs of the Evaluation Forum - Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics), Lecture Notes in Computer Science vol. 7488, pp. 76-81, Springer, doi: 10.1007/978-3-642-33247-0_9, 2012.

[6] L. Kovács: Local shape recognition for mobile applications, in Proc. of IEEE International Conference on Image Processing (ICIP), pp. 501-504, doi: 10.1109/ICIP.2012.6466906, 2012.

[7] A. Keszler, L. Kovács, T. Szirányi: Graph Based Descriptor Evaluation for Automatic Feature Selection, in Proc. of VISAPP (Intl. Conf. on Computer Vision Theory and Applications), vol. 1, pp. 375-380 (ISBN 9789898565037), 2012.

[8] L. Kovács: Interactive Search and Result Visualization for Content Based Retrieval, in Proc. of IMAGAPP&IVAPP, pp. 266-269, at IVAPP (Intl. Conf. on Information Visualization Theory and Applications), (ISBN 9789898425461), 2011.

[9] L. Kovács: Shape retrieval and recognition on mobile devices, in Proc. of MUSCLE International Workshop on Computational Intelligence for Multimedia Understanding, Lecture Notes in Computer Science vol. 7252, pp. 90-101, Springer, doi: 10.1007/978-3-642-32436-9_8, 2011.

[10] A. Kovács, Á. Utasi, L. Kovács, T. Szirányi: Shape and texture fused recognition of flying targets, in Proc. of Signal Processing, Sensor Fusion, and Target Recognition XX, SPIE vol. 8050, pp. 80501E-1-12, at SPIE Defense, Security and Sensing, doi: 10.1117/12.883765, 2011.