A GENERIC FRAMEWORK FOR THE EVALUATION OF CONTENT-BASED IMAGE AND VIDEO ANALYSIS TASKS IN THE CORE TECHNOLOGY CLUSTER OF THESEUS

Stefanie Nowak, Peter Dunker, Ronny Paduschek

Fraunhofer Institute for Digital Media Technology IDMT
Ehrenbergstrasse 31, 98693 Ilmenau, Germany
nwk@idmt.fraunhofer.de, dkr@idmt.fraunhofer.de, pdk@idmt.fraunhofer.de

ABSTRACT

THESEUS is a German research program that aims at the development of sophisticated algorithms and web-based infrastructures for the acquiring, processing and seeking of knowledge available on the web. The research focuses on text recognition, privacy, ontologies, user interfaces, video and image analysis, visualization techniques and evaluation strategies. All developments in THESEUS are evaluated by an independent work group, the Fraunhofer Institute for Digital Media Technology. In this paper we want to present our evaluation framework that is developed in THESEUS for the evaluation of the video and image analysis algorithms. Its key features are an easy extension to new formats and measures, the storing of previous test results for comparison and measurement of improvement, sophisticated visualizations for interactive reviewing and the generation of descriptive test results.

1. INTRODUCTION

THESEUS is a German research program that focuses on the development of a Internet-based infrastructure that better provides access to knowledge stored in the world wide web. The five years long project is divided into the core technology cluster (CTC), responsible for developing the core technologies, and the use cases that utilize these technologies in an application scenario. One of the major objectives of the THESEUS project is the application oriented research and it focuses on text recognition, privacy, ontologies, user interfaces, image and video analysis and evaluation strategies to name only a few central developments.

The Fraunhofer Institute for Digital Media Technology (IDMT) leads the evaluation task which is responsible for the quality assessment of the core technologies developed by several research partners involved in the CTC of the THESEUS program. The evaluation assures the quality control and the measurement of the ongoing research results. The task *Picture Analysis* of the evaluation workpackage focuses on the evaluation and benchmarking of CTC tasks that concentrate on image and video retrieval, classification, annotation and segmentation.

In this paper we would like to introduce our work in the evaluation of the image and video analysis algorithms in THESEUS. We therefore present our concept to evaluate different kinds of image and video analysis techniques in one general evaluation framework. Its key features are an easy extension to new formats and measures, the storing of previous test results for comparison and measurement of improvement, sophisticated visualizations for interactive reviewing and the generation of descriptive test results. We defined abstract *test cases* that cover the evaluation of all developments in the image and video analysis tasks. Test cases are concepts that encapsulate similar multimedia retrieval procedures and are used to generalize the evaluation framework for different evaluation needs at the conceptual level.

2. DATABASES AND MULTIMEDIA CONTESTS

One objective of the THESEUS project is to develop state-of-the-art algorithms and evaluation strategies to judge these algorithms regarding their quality or their performance in comparison to other algorithms. For this a ground truth is needed that decides if a result is correct or wrong. In the past many people put a lot of effort in collecting and annotating multimedia databases to get test data and execute contests. We therefore present an overview of both, existing databases and contests, in the area of image and video analysis.

2.1. Databases

To allow for a comparison of different algorithms, many research groups collected databases and made them publicly available for research purposes. In video retrieval the TRECVid database originated by the TRECVid benchmark [1], is established as a golden standard and widely accepted in the research community. For image retrieval, there exists no standard database yet. An often used commercial database is the Corel Database or subsets of it, although this database was often criticized in the community (see e.g. [2]).

Mainly the high ambiguity in annotating images makes it difficult to collect databases and define a ground truth. In contrast, the requirements concerning the features of the databases are commonly high and strongly task dependent.

CONTEST	AREA	TASK
PASCAL: Visual Object Class Challenge (VOC)[a]	Image	VOC 2008: 1) Classification (presence / absence of objects) 2) Object Detection 3) Pixel-wise object segmentation 4) Person Layout (detecting head, hands, feed)
Caltech Challenge[b]	Image	Caltech 2007: Classification (1 out of 256)
ImageCLEF[c]	Cross Language Image Retrieval	ImageCLEF 2008: 1) Photographic Retrieval 2) Medical Retrieval 3) Photographic Concept Detection 4) Automatic Medical Image Annotation 5) Image Retrieval from a Collection of Wikipedia Images
ImagEVAL[d]	Image	ImagEVAL 2006: 1) Recognition of transformed images 2) Text/Image mixed research 3) Detection of text areas 4) Detection of objects 5) Semantics Extraction
Berkely Segmentation Benchmark[e]	Image Segmentation	Image segmentation and boundary detection
TRECVid [f]	Video	TRECVid 2008: 1) Surveillance event detection pilot 2) High level feature extraction 3) Search 4) Rushes summarization 5) Content-based copy detection
VideOlympics[g]	Video	Showcase: Video Retrieval

Table 1. Overview of contests in multimedia retrieval, classification, annotation and segmentation

[a] http://www.pascal-network.org/challenges/VOC/
[b] http://www.vision.caltech.edu/CaltechChallenge2007/
[c] http://www.imageclef.org/
[d] http://www.imageval.org/e_presentation.html
[e] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/
[f] http://www-nlpir.nist.gov/projects/trecvid/
[g] http://staff.science.uva.nl/∼cgmsnoek/VideOlympics/index.php

Several requirements can be defined that should be fulfilled by an appropriate database for multimedia information retrieval. So these databases have to be **representative** to the databases of the end users to ensure to test performance under realistic testing conditions. Important is the **availability** of the database. Optimal is a free of charge available database without any copyright restrictions. It is necessary to collect a **huge diversity** of multimedia documents. The database has to be **general** enough to cover a large range of semantics from a human point of view and large enough to be **statistically significant**. (compare also [3])

Annotating multimedia documents is very time-consuming. So several collaborative initiatives evolved to solve this problem. Examples are the LabelMe[1] initiative from the MIT where objects in an image dataset are annotated with keywords and polygonal object contours or the ESP game[2]. Here images are collaboratively tagged in a game scenario.

In table 2 (see appendix), a list of the most important image and video databases is presented. It does not claim to cover all databases because a huge amount of databases, especially small ones, exist. Furthermore many databases are constructed by combining (parts of) other databases. One comprehensive overview of available multimedia databases was summarized by the MUSCLE project[3].

As it is apparent from table 2, the multimedia databases are characterized through a huge diversity in their characteristics, be it the size of items or the way they are labelled (format, language etc.). The lack of a golden standard database in content-based image retrieval and object classification, leads to the utilization and collection of this huge amount of diverse databases specific for the retrieval task. One result is, that the proposed retrieval algorithms are often incomparable and it is hardly possible to decide, whether one approach outperforms another or not. As a result several contests for different tasks in image and video retrieval became popular.

2.2. Multimedia Analysis Contests

Contests define challenging tasks with the goal to objectively measure the performance of algorithms and to establish a baseline for comparing systems.
Some criterions are important to consider by defining the tasks of a contest:

- Objectivity
  Evaluation has to be objective and unbiased concerning any algorithm or methodology.
- Scalability
  The contest also has to examine the systems scalability to huge databases.
- Processing Times
  The processing times of the evaluated algorithms have to be reported. Although processing times are not always important in research, especially for a proof of concept, when dealing with real world applications, they become the limiting factor. Generally one has to tradeoff between the accuracy of a retrieval algorithm and its speed. So for the judgement for or against an individual approach processing times are of great interest.

[1] http://labelme.csail.mit.edu/

[2] http://www.espgame.org/

[3] http://muscle.prip.tuwien.ac.at/data_links.php. Other listings of databases can be found in http://www.pascal-network.org/challenges/VOC/databases.html, http://www.cs.cmu.edu/∼cil/v-images.html or http://peipa.essex.ac.uk/benchmark/databases/index.html to name only a few.

- User's interest
  The tasks have to be defined in awareness of the user needs apparent for the examined technique.

- Expected real-world scenarios
  The tasks of a contest ideally should cover the entire spectrum of expected real-world scenarios. This issue also has to be regarded by deciding about an appropriate test database. [3]

Table 1 lists in short the most important contests in image and video retrieval. The column *Area* describes the application area of the contest in general; the column *Task* gives a short summary of the latest tasks. As most contests are repeated yearly with adopted and new tasks, this listing refers to the most recent challenges.

3. IMAGE AND VIDEO ANALYSIS IN THESEUS

The partners within the THESEUS CTC tasks research on a diversity of image and video analysis algorithms. Individual technologies are spread from image segmentation to frameworks for video event detection. In this section some selected approaches should be introduced.

A task that combines image and video analysis is the fast image and video identification based on perceptual hashing. Typical evaluation measures for video identification incorporates false-positives and false-negatives as e.g. in TRECVID 2008 task *Content-based copy detection(CBCD)*[4]. The definition of the concrete measure depends on the kind of application, e.g. the TRECVID 2008 CBCD weights false-negative more than false-positive. An industrial related challenge organized by the Motion Picture Association of America (MPAA) for filtering of user-generated content[5] weights false-positive more than false-negative. A comparison of various algorithms and a concrete evaluation can also be found in [4].

Further investigation in the video domain concentrates on temporal shot, subshot and scene change detection, on video genre estimation and video analysis for event detection. The video event detection technologies concentrate on a flexible framework for a fast and easy integration and optimization e.g. for different surveillance tasks. Next to image segmentation algorithms, partners research on spatio-temporal segmentation algorithms which incorporates the time dimension to analyse moving regions. Image classification or named entity retrieval is a task that is processed in different steps e.g. by research on new visual features or image or object representations, new classification approaches and fast indexing methods. These methods concentrate on generic approaches e.g. to be agnostic for use with medical images or user generated content. A specific image classification task concentrates e.g. on a robust face detection. Finally, partners research on different machine learning algorithms e.g. for parameter learning or improvement of other technologies.

4. CONCEPT

One challenge in developing a generic evaluation framework in THESEUS is constituted through the huge diversity of image and video applications that have to be evaluated and the generation of adequate test corpora and relevance judgments. In THESEUS, research in image and video analysis focuses on 1) image and video identification and similarity retrieval, 2) spatial and temporal segmentation, 3) face detection, 4) scene classification and annotation, 5) context detection, 6) video genre detection, 7) different query paradigms and 8) automatic quality assessment and correction of videos as already pointed out in section 3. All these scenarios have to be considered in the conceptualization of the framework.

Requirements on databases like representativeness, availability, diversity, generality and size (see section 2.1) have to be taken in mind by performing the evaluation. The evaluation has to deliver representative and reliable results, so criteria like the objectivity, scalability, processing times, users interest and expected real-world scenarios (see section 2.2) play a major role.

We defined abstract **test cases** that cover the evaluation of all developments in the area of image and video analysis. Test cases are concepts that encapsulate similar multimedia retrieval procedures and are used to generalize the evaluation framework for different evaluation needs at the conceptual level. Altogether we defined three test cases: 1) Retrieval 2) Keyword or Segment Indexing and 3) Multimedia Enhancement.

- Retrieval:
  Retrieval describes the scenario where one multimedia document serves as input into the analysis application and a list of similar documents is the output. This list can be further enriched with holistic annotations or segments and segment-based annotations of the single documents. Applications are low- or high-level based search scenarios.

- Keyword or Segment Indexing:
  The test case Keyword or Segment Indexing covers all scenarios, where one media item is the input into an application and a description of this item is computed as output. These descriptions are holistic annotations, segment information or segment-based annotations. This case is applied for the evaluation of face or object detection as well as classification algorithms.

- Multimedia Enhancement:
  Last, Multimedia Enhancement deals with all cases where the input multimedia document is processed and an

[4] `http://www-nlpir.nist.gov/projects/tv2008/Evaluation-cbcd-v1.3.htm#eval`

[5] `http://opinion.latimes.com/bitplayer/2007/03/filtering_userg.html`

enhanced version of this document serves as output like in automatic distortion corrections in images or videos.

Depending on the current test case, adapted evaluation measures are chosen and different views for the visualization and interpretation of the results are available. Most important are the test cases Retrieval and Keyword or Segment Indexing.

Besides the test cases, the evaluation framework has to contain modules for the query topics, performance tests, result visualizations and track the performance improvement.

Query Topics

Query topics have to be defined in following development iterations of the evaluation system (e.g. for the evaluation of image annotation algorithms) and are reused over the duration of the project. For this a query module will be developed and integrated into the framework. The query topics, the keywords and examples for testing, as well as the data-basis are documented to ensure comparability and measurement of improvement over time. We consider to update the databases and topics step by step, so the evaluation will be more comprehensive at the end of the project.

Performance Tests

In the performance tests we will for example monitor how fast a classification and retrieval task is done on different scale databases.

Visualization

Visualization offers a way to intuitively get insight into the characteristics of the performance of an algorithm. Various views on the results strengthen the awareness of positive and negative aspects of the algorithms performance. We therefore would like to provide different views to facilitate a deeper analysis of the results.

Track performance improvement

In the framework the improvement of the iterative refinement of the algorithms is measured by a statistically relevant measure. Significance tests are used to document this improvement .

5. EVALUATION FRAMEWORK

In general, the framework consists of the Evaluation Toolbox itself, a Graphical User Interface (GUI) to visualize and invoke the evaluation processes and an Input Interface that converts the output of any evaluated system to an internal file format. Additionally a binding to a database that holds the ground truth data and saves all evaluation results will be established. So it can be used in later runs for comparisons and to judge the improvement over time.

In the overall workflow, the system from the CTC task is executed and its results are saved on the file system. Therefore the output format of the external system is used. A loader module reads the output file and converts it to the internal processing format. In dependence from the data and provided annotations, an evaluation process is defined and computed. The results of the evaluation are displayed in the GUI and all results are saved in a database.
The Evaluation Toolbox therefore is divided into different modules that are shortly presented in the following:

- Evaluation Manager:

  The Evaluation Manager is used to connect all modules and to control the workflow.

- Convert Input Data Module:

  This module is responsible to load the output files from the systems and convert them into the internal processing format.

- Visualization Module:

  The Visualization Module serves as interface between the evaluation results and their graphical display. Its task is to provide different possibilities of result visualization and it has to deliver the necessary data to the Graphical User Interface. One key aspect of evaluation is to present a comprehensive insight into the performance of a system. Therefore different visualizations are provided. This can be e.g. ROC curves, the n best ranked photos of a photo retrieval task or a visual history of results for the same test case.

- Load Data Module:

  The Load Data Module is used as interface between the database and the evaluation processes. It provides the functionality of loading ground truth data and annotations from the database. Additionally it allows to access former test results of a test case.

- Evaluation Measures:

  Different evaluation measures e.g. rank-based measures or Precision / Recall measures are implemented. The evaluation toolbox supports a plug-in mechanism that allows adding different evaluation measures step by step.

- Significance Tests:

  Throughout the THESEUS project the Evaluation Toolbox will be complemented with significance tests. With the help of a significance test it is possible to compute if the improvement of the systems results over time is significant.

The first evaluation tests are performed in September 2008 and evaluate the performance of the image segmentation and face detection algorithms. In the next sections both evaluation tasks are described more in detail.

5.1. Image Segmentation

One of the first evaluation tasks in THESEUS is the evaluation of a segmentation algorithm. For the evaluation of image segmentation two things need to be taken into consideration. First, which image data is used for testing and is ground truth available for the data? Second, which measure should judge the performance of the algorithm?

5.1.1. Test corpus

We decided to utilize the Berkeley Segmentation Dataset provided by Martin et. al. [5] as one test corpus for evaluating the algorithm. It altogether consists of 1000 images from the Corel collection from which 300 images are publicly available. Martin et. al. collected between 5 and 10 human segmentations for each image from different persons with the objective to deal with the subjective factor of segmentation and different levels of granularity in the segmentation process. 200 images are expected for the training of the algorithms, 100 for testing purposes.

5.1.2. Evaluation measures

In general, image segmentation can be evaluated *parameter-based*, *boundary-based* or *region-based* when ground truth is available. The evaluation framework incorporates two measures for segmentation evaluation at the moment, one boundary-based and one region-based measure.

The first measure is the F-Measure that is computed for corresponding boundaries of the segmented and ground truth images. This measure is integrated from the Berkeley Segmentation Benchmark Toolbox [6] into our framework. The machines segmentation output is first converted into a boundary map. Due to several human segmentations for each image, each human boundary map is separately compared to the machines result and only edge pixels that match no human boundary are regarded as false positives. The output from the original benchmark are html files containing the results, that can be sent to the authors and are published in the web[6].

We integrated the boundary-based measure of the benchmark into our evaluation framework because of two reasons:

1. The ground truth was carefully collected and subjectivity was minimized by regarding several human segmentations as possible and correct results. So different granularities of segmentations are present in the ground truth and therefore mirror different user expectations concerning a correct segmentation result.
2. The published results on the webpage allow us to compare the segmentation results from THESEUS to other segmentation results that were computed under the same conditions.

Additionally, we implemented a region-based measure, the *normalized Hamming Distance* proposed by Huang and Dom in [7].

The *directional Hamming Distance* between a segmentation $S = \left\{R_1^1, R_1^2, ..., R_1^m\right\}$ and a ground truth segmentation $G = \left\{R_2^1, R_2^2, ..., R_2^n\right\}$ for the same image is denoted by $D_H(S \Rightarrow G)$. For each region R_2^i from G a region R_1^j from S is associated so that $R_2^i \cap R_1^j$ is maximal. Then

$$D_H(S \Rightarrow G) = \sum\nolimits_{R_2^i \in G} \sum\nolimits_{R_1^k \neq R_1^j, R_1^k \cap R_2^i \neq \emptyset} |R_2^i \cap R_1^k| \tag{1}$$

with $|.|$ as the size of the set. This measure can be symmetrical computed for $D_H(G \Rightarrow S)$. The normalized Hamming Distance is then computed as

$$p = 1 - \frac{D_H(S \Rightarrow G) + D_H(G \Rightarrow S)}{2 * |S|} \tag{2}$$

where $|S|$ determines the image size and the resulting score $p \in [0, 1]$. In case of a perfect segmentation $p = 1$.

The directional Hamming Distance also allows to derivate two error rates: the *missing rate* E_R^m and the *false alarm rate* E_R^f. They are computed as

$$E_R^m = \frac{D_H(S \Rightarrow G)}{|S|} \text{ and } E_R^f = \frac{D_H(G \Rightarrow S)}{|S|}. \tag{3}$$

5.1.3. Result presentation

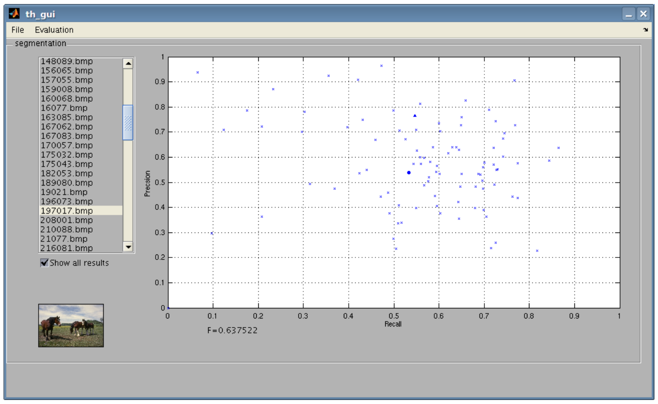
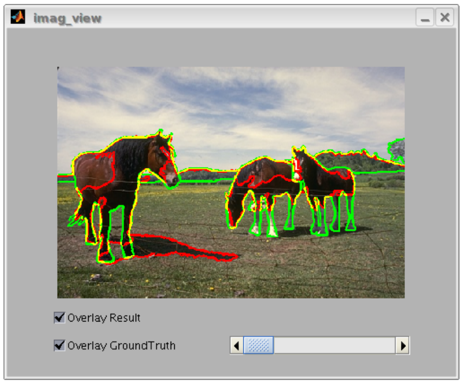


Fig. 1. Visualization of the overall results of a segmentation run.

One example of the graphical representation of the results in the segmentation evaluation task can be seen in figure 1. All

[6] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/bench/html/algorithms.html

segmentation results for the test corpus are shown as crosses in a Precision / Recall Plot. The mean segmentation performance is denoted by a filled circle. The evaluator can click on the crosses and the corresponding image is highlighted in the list on the left and displayed on the left bottom of the window. Additionally the f-measure for this particular image is shown on the bottom.



☑ Overlay Result

☑ Overlay GroundTruth

Fig. 2. Visualization of the segmented and ground truth boundaries of one image.

The image can be enlarged in a separate window (see figure 2) and the boundaries of the computed segments and the ground truth are marked in red and green respectively. Overlapping boundaries are drawn in yellow to show the correspondences of both segmentations. It is possible to step through the different ground truth annotations.

5.2. Face Detection

This task refers to the evaluation of face detection systems. A ground truth set of more than 350 images with a total of 1000 annotated faces is collected for the first evaluation analysis and will be increased in further treatments. The data is annotated manually and consists of bounding boxes, which are drawn around each face. The coordinates of the ground truth data are arranged as follows: upper left x coordinate, upper left y coordinate, lower right x coordinate, lower right y coordinate of the bounding box. Thus, each ground truth file contains one bounding box for each annotated face of the corresponding image. The algorithm that is evaluated in this task, is executed on the test data and also delivers bounding boxes around detected faces that are saved in the same way.

The evaluation workflow for the face detection is described below. All images, ground truth data and test data are loaded into the evaluation framework. The data for each image is arranged in an individual matrix separately for ground truth and detected data. Afterwards the matrices are compared with each other. The detected faces from the algorithm are compared with the ground truth data. To relate a detected face to a ground truth face, it has to fulfil two constraints which refer to the position and the size of a bounding box. First, the Euclidean distance $\Delta dist\_xy$ between the upper left corners of the bounding boxes, shown in Figure 3, is calculated

$$\Delta dist\_xy = \sqrt{\Delta x^2 + \Delta y^2}, \tag{4}$$

whereas $\Delta x^2 = (x''_{ul} - x'_{ul})^2$ and $\Delta y^2 = (y''_{ul} - y'_{ul})^2$.

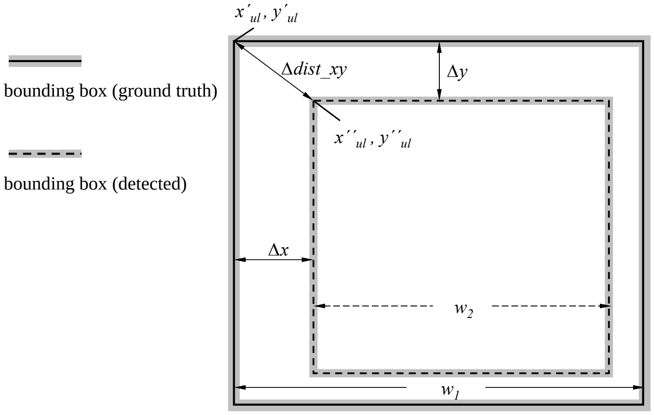


Fig. 3. *Bounding boxes. The coordinates are used to compare the ground truth with the detected face.*

Thus, the first decision function is $\Delta dist\_xy < \xi w_1$, where $\xi = 0.5$.

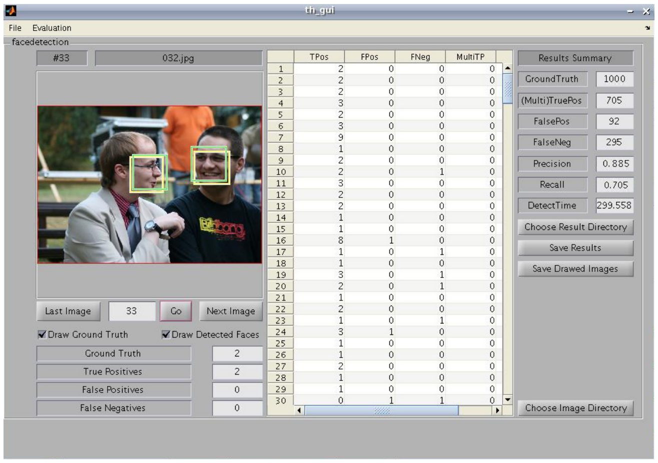
A general problem is the different size between bounding boxes of the detected faces and the rectangles around the annotated faces. We defined a tolerance value which is given by

$$w' = \frac{|w_2 - w_1|}{w_1}. \tag{5}$$

The second constraint is fulfilled if $w' < \psi$, with $\psi = 0.4$.

The variables ξ and ψ are heuristically determined and can be adapted subsequently if required. Consequently, a detected face is set as $true\ positive$, if both constraints are true. Otherwise, it is set as $false\ positive$. Faces, which are not detected are labelled as $false\ negative$. The sum of the criterions over all images are used to calculate the measurements precision and recall. According to our algorithm for comparison of the bounding boxes, Kasturi et. al. describe in [8] another approach which concentrates on the overlapping between the ground truth box and the box which is detected by the detection system.

Further features of the evaluation framework for face detection are the visual representations of the results. The user can visualize the results and can take a look at each image. It is also possible to draw and display the bounding boxes around each face arbitrarily. The images with bounding boxes and the analysis results can be stored on demand. A demonstration of the graphical interface of the face detection is given in Figure 4.



#33 032.jpg

33

☑ Draw Ground Truth ☑ Draw Detected Faces

Ground Truth	2
True Positives	2
False Positives	0
False Negatives	0

	TPos	FPos	FNeg	MultiTP
1	2	0	0	0
2	2	0	0	0
3	2	0	0	0
4	3	0	0	0
5	2	0	0	0
6	3	0	0	0
7	9	0	0	0
8	1	0	0	0
9	2	0	0	0
10	2	0	1	0
11	3	0	0	0
12	2	0	0	0
13	2	0	0	0
14	1	0	0	0
15	1	0	0	0
16	8	1	0	0
17	1	0	1	0
18	1	0	0	0
19	3	0	1	0
20	2	0	1	0
21	1	0	0	0
22	2	0	0	0
23	1	0	1	0
24	3	1	0	0
25	1	0	0	0
26	1	0	0	0
27	2	0	0	0
28	1	0	0	0
29	1	0	0	0
30	0	1	1	0

GroundTruth	1000
(Multi)TruePos	705
FalsePos	92
FalseNeg	295
Precision	0. 885
Recall	0.705
DetectTime	299.558

Fig. 4. *Graphical User Interface of face detection task. The current image is displayed with the detected and ground truth bounding boxes. A matrix over all image results is shown in the middle of the GUI. An overview over the summarized analysis is given on the right side.*

6. CONCLUSION AND FUTURE WORK

All in all we presented the concept for a generic evaluation framework that encapsulates methods for the evaluation of a huge diversity of image and video analysis algorithms. We introduced three types of test cases in which all developments of the CTC partners in the area of image and video analysis in THESEUS can be categorized in. Two initial evaluations already took place, namely image segmentation and face detection, and were presented here. The evaluation framework will be completed over the project duration with the proposed features.

A summary of already running contests and available databases was presented. It is not the goal to duplicate already accepted contest tasks in THESEUS. Due to this, the results of an algorithm for image classification will be submitted to the Visual Object Class Challenge 2008 by a CTC partner. In this case the contest task perfectly fits to the algorithms purpose. Additionally there is the advantage that automatically there is a comparison of this algorithms performance to other state-of-the-art algorithms. In the future we search for collaborations with existing contest organizators to organize a task in a contest, that fits to the characteristics of other image and video analysis algorithms from THESEUS that are not covered yet by a running contest.

7. ACKNOWLEDGMENTS

This work has been supported by grant No. 01MQ07017 of the German THESEUS program.

8. REFERENCES

[1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, (New York, NY, USA), pp. 321–330, ACM Press, 2006.

[2] H. Mueller, S. Marchand-Maillet, and T. Pun, "The Truth about Corel-Evaluation in Image Retrieval," *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 38–49, 2002.

[3] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," 2008.

[4] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video copy detection: a comparative study," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, (New York, NY, USA), pp. 371–378, ACM, 2007.

[5] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *Proc. Eighth Int'l Conf. Computer Vision*, vol. 2, pp. 416–423, 2001.

[6] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 530–549, 2004.

[7] Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," *IEEE International Conference on Image Processing*, vol. 3, pp. 53–56, 1995.

[8] R. Kasturi and et. al., "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1–17, 2003.

A. OVERVIEW OF MULTIMEDIA DATABASES

Database	Category	Annotations	Size	Objects	Remarks
		General			
IAPR TC12 Benchmark data[a]	Sports, city, landscape, animals, people, action	Word annotations in German, English and partly Spanish	20.000 color images, 20.000 thumbnails	-	license agreement
LabelMe	-	Word annotations in English	161.894 color images, 42.043 annotated objects	Object contours as polygons	Collaborative annotated, partly images not annotated or not completely
Caltech: 101 Object Categories[b]	101 categories with each 31 to 800 images	Keywords of objects through categorization	9144 color images	Bounding box and outline of objects	Not usable for segmentation evaluation because objects are centered
Caltech: 256 Object Categories[c]	256 categories with each 80 to 827 images	Keywords of objects through categorization	30.607 color images	-	-
Corel: 1000 Dataset[d]	Subset of Corel images: 10 image categories with each 100 images	-	1000 color images	-	Only part of the Corel dataset, publicly available
Princeton: Natural Scene categories[e]	13 image categories with each between 215 and 410 images	-	3859 Grayscale images	-	-
Princeton: Event Dataset[f]	8 sport event categories with each between 137 and 250 images	challenge level (easy etc.), distance to foreground objects	1579 color images	-	-
Washington Database[g]	22 image categories with each between 22 and 256 images	Word annotations of depicted objects (no positions)	1259 color images	Word annotations of depicted objects (no positions)	-
		Segmentation			
Berkely Segmentation Dataset[h]	-	- (no semantic description of segments)	300 images in color and grayscale (Public part)	Edge images of hand labeled segments	Subset of Corel images
Berkely Segmentation Dataset - Barnards enhancement[i]	-	Extended through semantic labels for each segment from WordNet	1014 images + segmen-tations + semantic labels	Edge images of hand labeled segments + Word-Net labels	Subset of Corel images
MRSC: Database 1[j]	9 object classes	each object labeled with an object class or *void*	240 color images	Pixelwise labeling of objects	-
MRSC: Database 2[k]	23 object classes	each object labeled with an object class or *void*	591 color images	Pixelwise labeling of objects	-
PASCAL: VOC 2007[l]	20 object classes	Class annotations, 24.640 annotated objects (bounding box + label)	9963 color images (training and test set)	Pixelwise object segmentation (632 images)	Standardized segmentations and annotations of different publicly available databases[m]
		Face			
Yale Face database B[n]	10 subjects under 576 viewing conditions (9 poses x 64 illuminations) + ambient illumination	-	5850 grayscale images	Coordinates of faces + coordinates of left + right eye + mouth	-

[a] http://eureka.vu.edu.au/~grubinger/IAPR/TC12_Benchmark.html

[b] http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

[c] http://www.vision.caltech.edu/Image_Datasets/Caltech256/

[d] http://wang.ist.psu.edu/~jwang/test1.tar

[e] http://vision.cs.princeton.edu/Datasets/SceneClass13.rar

[f] http://vision.cs.princeton.edu/lijiali/event_dataset/event_dataset.rar

[g] http://www.cs.washington.edu/research/imagedatabase/groundtruth/

[h] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

[i] http://vision.cs.arizona.edu/quanfu/semantic/index.html

[j] http://research.microsoft.com/vision/cambridge/recognition/MSRC_ObjCategImageDatabase_v1.zip

[k] http://research.microsoft.com/vision/cambridge/recognition/MSRC_ObjCategImageDatabase_v2.zip

[l] http://www.pascal-network.org/challenges/VOC/voc2007/VOCtrainval_06-Nov-2007.tar

[m] http://www.pascal-network.org/challenges/VOC/databases.html

[n] http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

Database	Category	Annotations	Size	Objects	Remarks
		Face			
Color FERET[a]	994 subjects, 13 poses	4 coordinate labels (eyes, nose tip, mouth centre) for a large part of db	11.338 facial color images	-	-
Grayscale FERET[b]	1209 subjects	4 coordinate labels (eyes, nose tip, mouth centre) for a large part of db	14.051 facial grayscale images	-	-
MIT-CMU PIE database[c]	68 subjects, 13 poses, 43 illuminations, 4 expressions	Labels concerning sex, age, glasses, moustache, beard, datum of capture, head, camera, flash locations, background images, color calibration images	41.368 color images	-	-
		Medical			
IRMA 2007 (see also data-sets from 2003, 2005, 2006)[d]	116 categories	-	12.000 radiographs	-	-
ImageCLEFMed 2007 [e]	Casimage (8725), MIR (1177), PEIR (32.319), PathoPic (7805), myPACS (15.140), endoscopic (1496)	Clinical case descriptions, metadata records, English, French, German	66.662 images	-	Only available for participants at benchmark
		Video			
TRECVid 2007(see also data of earlier years) [f]	36 high-level concepts	Shot boundary annotation, high level feature ground truth, search relevance judgments, video summary groundtruth	Sound and Vision: 106 hours, BBC: 50 hours	-	Only available for participants at benchmark
CAVIAR Test Case Scenarios [g]	different scenarios: people walking alone, meeting with others, window shopping, entering and exitting shops, fighting and passing out, leaving a package in a public place	Individual and group bounding boxes, additional: head, gaze direction, hand, feet and shoulder positions (7 sequences)	2 sets: a) 28 sequences, b) 26 sequences in corridor and in front view	Bounding box around persons	-

Table 2. Overview of databases for multimedia retrieval, classification, annotation and segmentation.
Category describes the domains the images/videos can be categorized in. *Annotations* summarizes if there are additional word annotations present for the images/videos. *Size* of the database determines how many documents belong to it. The term *Objects* depicts whether the objects in the documents are labeled through bounding boxes, object contours or pixel masks. *Remarks* summarizes special features relevant for the usage of the database.
Additionally the table is structured into the five subsections: *General*, *Segmentation*, *Face*, *Medical* and *Video*, determining the main application area for the database. General refers to traditional image classification tasks. Depending on the availability of additional annotations and their form, these databases can also be used for object classification or image annotation tasks. This is also true for the more specific databases. When having the information of a pixel-wise labeled object in an image and the corresponding object class, of course this database can also be used for object detection and object classification additionally to segmentation tasks.

[a] http://www.nist.gov/humanid/colorferet
[b] http://www.itl.nist.gov/iad/humanid/feret/
[c] http://www.ri.cmu.edu/projects/project_418.html
[d] http://www.irma-project.org/datasets_en.php?SELECTED=00007#00007.dataset
[e] http://ir.ohsu.edu/image/2007protocol.html
[f] http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html
[g] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/