



**SALERO**

# **Report on Search System Evaluation**

**SALERO Deliverable D5.7.1**





# Report on Search System Evaluation

SALERO identifier: SALERO-D5.7.1-UG-Report on Search System  
Evaluation-v02.doc

Deliverable number: D5.7.1

Author(s) and company: R. Villa, P. Swamy, V. Strathopoulos, Y. Feng,  
H. Misra, F. Hopfgartner, K. Athanasakos, I. Arapakis,  
M. Halvey, D. Hannah, A. Goyal, R. Ren, J. Jose (UG)

Work package / task: WP05

Document status: Final

Confidentiality: Restricted

## DOCUMENT HISTORY

Version	Date	Reason of change
1	2010-02-12	Version for internal review
2	2010-02-22	Final Version

The work presented in this document was partially supported by the European Community under the Information Society Technologies (IST) priority of the 6th framework programme for R&D.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

This document contains material, which is the copyright of certain SALERO consortium parties, and may not be reproduced or copied without permission. All SALERO consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the SALERO consortium as a whole, nor a certain party of the SALERO consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

## Table of Contents

---

<b>1</b>	<b>Executive Summary .....</b>	<b>1</b>
<b>2</b>	<b>Introduction.....</b>	<b>2</b>
2.1	Purpose of this Document .....	2
2.2	Status of this Document.....	2
2.3	Related Documents .....	2
<b>3</b>	<b>Automatic Evaluation of Search Systems .....</b>	<b>3</b>
3.1	The Cranfield/TREC Evaluation Methodology.....	3
3.2	TRECVID Test Collections .....	4
3.3	TRECVID 2006.....	5
3.4	TRECVID 2007.....	6
3.5	TRECVID 2008.....	7
3.6	TRECVID 2009.....	9
3.7	Summary .....	10
<b>4</b>	<b>Automatic Image Annotation .....</b>	<b>12</b>
4.1	Introduction .....	12
4.2	An Example Approach to Image Annotation.....	12
4.3	A Framework for Evaluating Automatic Image Annotation Algorithms.....	14
4.3.1	<i>Sampling Procedure</i> .....	14
4.3.2	<i>Content Descriptors</i> .....	15
4.4	Results .....	15
4.5	Conclusion .....	17
<b>5</b>	<b>Performance on the Alan Online Image Collection.....</b>	<b>18</b>
5.1	The Retrieval Problem .....	18
5.2	Requirements .....	19
5.3	Initial Approach .....	19
5.4	Second Approach .....	20
5.5	Evaluation .....	22
5.6	Issues.....	25
5.7	Summary .....	26
<b>6</b>	<b>Image Search via an Intermediary .....</b>	<b>27</b>
6.1	Introduction: The Problem .....	27
6.1.1	<i>Searching via an Intermediary Collection</i> .....	28
6.2	Searching via an Intermediary using the AspectBrowser Interface.....	29
6.3	Offline Evaluations of the Technique.....	31
6.3.1	<i>COLLECTIONS and SYSTEMS</i> .....	31
6.3.2	<i>Automatic Retrieval via an Intermediary</i> .....	32
6.3.3	<i>Manual Intermediary</i> .....	34
6.4	User Interface Evaluation using a Specialised Interface .....	35
6.4.1	<i>The Search Interfaces</i> .....	36
6.4.2	<i>Procedure</i> .....	37
6.4.3	<i>Results</i> .....	37
6.4.4	<i>Discussion</i> .....	40
6.5	Conclusions .....	40

<b>7 Vigor: Evaluation of a Grouping Interface</b> .....	<b>41</b>
7.1 Motivation .....	41
7.2 The Vigor Search Interface .....	41
7.3 User Studies .....	42
7.4 Exploratory Task Evaluation .....	43
7.4.1 Exploratory Task Evaluation Results .....	44
7.5 TRECVID Evaluation .....	45
7.5.1 TRECVID Evaluation Results .....	46
7.5.2 User Feedback .....	47
7.6 Conclusions .....	49
<b>8 Conclusions</b> .....	<b>51</b>
<b>9 References</b> .....	<b>53</b>
<b>10 Glossary</b> .....	<b>59</b>

## 1 Executive Summary

---

This document describes the evaluation efforts which have occurred over the course of the SALERO project, in relation to the content-based search system which have been developed over the period of the project.

The report starts by presenting a range of the retrieval techniques and associated automatic evaluation results, undertaken as part of the TRECVID effort. These efforts have been to improve the performance of the underlying retrieval system, on which the other interfaces and systems presented in later sections are based. Such automatic evaluation of retrieval performance has become a vital part of Information Retrieval research, and has also provided the bedrock on which other parts of WP5 have been built.

Following on from this, we present work concerning automatic image annotation, first briefly describing the evaluation of an approach developed as part of SALERO, before describing a framework for evaluating annotation systems. Automatic image annotation attempts to attach high-level, semantic concepts to individual images based on low-level image features. If such annotations are provided, searching can then be carried out on the annotations, which have some semantic significance, rather than the visual properties of the images. Given the difficulty of this problem, it is still very much an ongoing research problem, but similar to the basic research carried out while taking part in TRECVID, such research has a considerable application to the problem of search in SALERO.

Building on the systems built during the TRECVID evaluations, we then present work which has taken place to improve retrieval in a single application: that of the Alan online art installation. This problem – the retrieval of the “best” matching image from a small collection of images, based on a user-drawn query – is specific to SALERO and the Alan online system and its associated physical installation. Chapter 5 describes in detail the new work which has gone into solving this problem, carried out since the SALERO review at IBC in Amsterdam.

Chapter 6 discusses a technique for searching image databases without textual annotations or metadata, as an alternative to the automatic annotation methods discussed in Chapter 4. This is the result of an offshoot of the AspectBrowser interface and takes advantage of the “contextual” design of the interface which allows multiple aspects to be created and results from one search to be used as a query to another search. The AspectBrowser interface has already been described in detail in previous reports (e.g. D5.5.4, The System Design and Implementation of the Context-based Search System). The problem of searching purely image databases has come up frequently in the SALERO project – i.e. how can we search a large collection of un-annotated images using only content-based search? Examples of such databases are the asset collections of images maintained by PGP. One approach is to use automatic image annotation, but given the provisional state of research in that area, this is not currently practical. The intermediary technique discussed and evaluated in Chapter 7, on the other hand, is one available to users now within the SALERO search system.

The next chapter, Chapter 7, describes the evaluation of the ViGOR search interface. ViGOR is a grouping interface, which provides an alternative approach to the problem of searching and organising material in video and image databases. It may be considered as complementary to the AspectBrowser interface, providing a different view of the “context” of a search. In the AspectBrowser interface, each search aspect can be considered as existing within the context of the larger sequence of aspects, which each aspect providing both search history and allowing the saving of search results. In ViGOR, a single search history exists, with the user able to organise images into “groups”, all of which exist within the context of the user’s search workspace. The underlying search system is the same as that used in the AspectBrowser interface, based on the TRECVID research presented in Chapter 3.

The document finishes with a summing up of the SALERO evaluation work, carried out on the developed search interfaces.

## 2 Introduction

---

### 2.1 Purpose of this Document

---

This document reports the numerous evaluations which have taken place with the SALERO retrieval system over the course of the project, and importantly, provides an updated system and evaluation of a system for searching the Alan Online image database supplied by TAIK.

It should be noted that this document is not an exhaustive listing of all the evaluations which have taken place on the SALERO search system. In particular, the evaluation of the AspectBrowser interface has already been described in document D5.5.4, Chapter 6; the integration of the search system with two SALERO experimental productions has also been described in document D5.5.4, Chapter 7. Neither of these will be repeated here; we instead refer readers to the appropriate chapters of the aforementioned deliverable.

### 2.2 Status of this Document

---

This is the final version of the deliverable D5.5.7.

### 2.3 Related Documents

---

Before reading this document it is recommended to be familiar with the following documents:

- D2.3.1 User Requirements Document
- D5.5.1 Context-based retrieval system and user interface
- D5.5.2 Retrieval algorithms based on contextual features
- D5.5.4 The System Design and Implementation of the Context-based Search System

## 3 Automatic Evaluation of Search Systems

---

Before dealing with user evaluations in Chapter 6 and 7, we first start with descriptions of the evaluation of the underlying image and video retrieval system developed during the SALERO project. This system is the one underlying the interfaces and demos presented elsewhere, and has been extensively developed and evaluated by participation in the international TRECvid effort.

The dominant evaluation paradigm in the Information Retrieval research community is the Cranfield/TREC evaluation methodology, which provides a mechanism for evaluating the performance of a retrieval system within a well defined lab environment. No user feedback is required, instead a set of search topics and associated relevance judgements, which are generated manually, are used to evaluate the performance of a retrieval system. This basic methodology has been used, and continues to be used, in TREC and TRECvid for the evaluation of text and video retrieval systems.

Before considering user evaluations in later chapters in this report, we first outline the results of four years of participation in the TRECvid video retrieval effort, between 2006 and 2009. The results of these efforts have been used to improve the underlying backend retrieval system used in the SALERO search system, and have been of fundamental importance in the overall development of the system.

### 3.1 The Cranfield/TREC Evaluation Methodology

---

The dominant approach to the evaluation of Information Retrieval systems is typically called the “Cranfield Paradigm” [Voorhees 2001], although it is perhaps more accurately described as the TREC approach. The evaluation methodology is based on the idea of a test collection, which is composed of the following three components:

- A document collection, which in the case of image and video retrieval, is composed of a large set of images or videos
- A set of *topics* or “queries”. Each topic describes an information need of a user, and may be composed of text, image or video data depending on the test collection
- A list of relevance judgements, often called *qrels*. This list specifies which documents in the collection are relevant or non-relevant to which topics, and is typically generated manually

Given these resources, the evaluation of a retrieval system can then be carried out automatically: each topics can be submitted to the retrieval system, the results returned and then automatically evaluated against the relevance judgements which provide the “ground truth” by which the system is evaluated.

This approach was first used in the classic Cranfield evaluations ([Cleverdon 1967] and [Cleverdon 1991]), which aimed to investigate different indexing languages. The approach was used in a number of studies in an ad-hoc manner, while in 1975 Spark Jones and van Rijsbergen [Sparck Jones 1975] suggested the definition and creation of an “ideal” test collection. In 1992 the National Institute of Standards and Technology (NIST) and U.S. Department of Defense started the Text REtrieval Conference (TREC), which became the standard resource for the evaluation of text retrieval engines. Since by this time text collections used in evaluation were too large to exhaustively evaluate all topics against the whole collection, a “pooling” method is used, where the top  $n$  documents from each retrieval engine which takes part in TREC are pooled into a single collection of documents which are then manually judged for relevance. Any documents not judged are then normally considered as non-relevant to the topic.

In 2001 the “TRECvid” (Video Retrieval Evaluation) effort started as an offshoot of the TREC text evaluation, the aim of which was to evaluate video retrieval systems. This effort has provided an internationally recognised forum from which the development of the backend SALERO retrieval system could be investigated. The techniques used during the four years of the project, and the associated evaluation results will be presented in Sections 3.2 to 3.5. In the next section, we briefly describe a typical TRECvid test collection from the point of view of the search task.


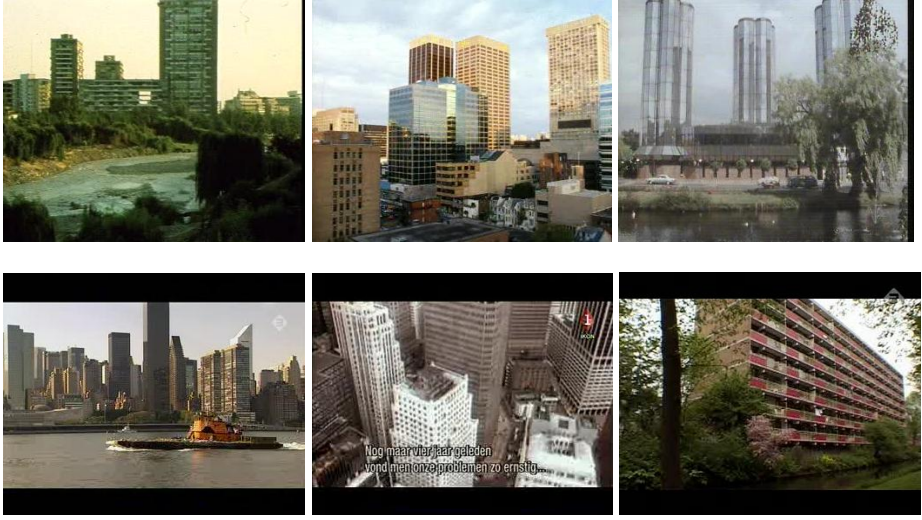
Topic Number	0271
Text	Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible
Image examples	
Video examples (frames from the video)	

Table 1: An example of a TRECvid topic, from the TRECvid 2009 collection

### 3.2 TRECvid Test Collections

A TRECvid test collection is composed of the same parts as previously discussed: data collection (videos), list of search topics, and list of relevance judgements. Since the aim of TRECvid. For TRECvid 2009, the most recent collection to date, is composed of the following main components:

- 400 videos from Dutch TV, supplied by the Netherlands Institute for Sound and Vision, totalling roughly 280 hours of material, including news magazines, science, news reports, documentaries, educational programming, etc.
- Master shot reference files, which specify for each video the extent of each video “shot”
- 25 topics, which include a short text query plus video and/or image examples of the topic information need
- A list of relevance judgements generated by NIST

In addition to these resources, various other elements, such as high level features, may also be provided, either by NIST or other groups contributing in the TRECvid effort. One of the more important resources which has been provided, in some TRECvid efforts, is the output from a automatic speech recognition system.

The master shot reference specifies, for each video, where the shot boundaries between separate video shots lie. The aim of the TRECvid search task is to return a ranked list of video shots for each topic. Having a single common shot list ensures all groups who take part in TRECvid use the same shots.

An example TRECvid topic is shown in Table 1. It is composed of a topic number, a short text query, an optional list of images illustrating the topic, and a further optional list of video examples which further illustrate the topic. A retrieval engine can therefore use a combination of the text, image or video data in order to search the video collection. In the work reported here, the main focus is using the image and video examples to search the TRECvid collection of videos by content.

The list of relevance judgements, together with a ranked list, can be fed into a program such as “trec\_eval”<sup>1</sup> to generate various standard Information Retrieval evaluation metrics, including:

- Mean Average Precision (MAP)
- Interpolated Recall - Precision Averages at different recall levels (from 0.00 to 1.00)
- Precision after  $n$  documents retrieved (where  $n$  may be 5, 10, 15, 20, 30, 100, etc.). This is often written as P5, P10, P20, etc.

Further details on all of these measures can be found in [van Rijsbergen-1979].

### 3.3 TRECVID 2006<sup>2</sup>

---

This was the first year GU participated in the TRECVID search task. One interactive run and five fully automatic runs were submitted: we restrict ourselves here to discussing the automatic runs only. The automatic runs were a combination of text features only (UG F 1), visual features only (UG F 2 and UG F 5) and a combination of feature modalities (UG F 3 and UG F 4). The following list describes the submitted runs:

- UG F1 Text baseline: The Terrier retrieval system [Ounis 2005] was used with the BM25 retrieval model, with each shot represented using the automatic text transcripts from the six shots preceding and following the shot of interest
- UG F2 Automatic search based on visual features (optimised weighting): Five MPEG-7 standard visual features were used (Colour layout, contour shape, dominant colours, edge histograms, homogeneous texture), and positive and negative example keyframes extracted from the test topics. A linear combination of the scores from the different features are used, for this particular run, the feature being scaled based on the classification error on the training (topic) data
- UG F3 Graph model based on text query: Used a combination of text and visual features combined in the Image-Context Graph (ICG), proposed in [Urban 2006]. Querying in the ICG is implemented using the theory of random walks [Lovasz 1993], where querying involves choosing a set of starting or query nodes, and then computing the stationary distribution of the ICG based on a restart vector. In this run, a text query was issued to the ICG.
- UG F4 Graph model based on text and visual query examples: As for run F3 above, except this run also uses visual features represented in the graph.
- UG F5 Automatic search based on visual features (equal weighting): As for run F2, except all features are treated as equal
- UG I1 Interactive search run (text and visual features): This used a custom video retrieval interface, presented in [Urban 2006c]

The results for these runs are shown in Table 2 and Table 3. Results using only visual features (F2 and F5) are, as can be expected, poor relative to the text. The baseline text run appears roughly in line to other submissions, suggesting a similarity of approach with other participating organisations. UG F4 performs slightly better than UG F3, although neither can improve on the text baseline. In hindsight, both the use of query expansion (QEX) and the choice of issuing one random walk per query term rather than using one overall restart vector containing all query terms, have decreased the performance of the ICG in run UG F3. The MAP score of the original run submitted was 0.0183. A run using one overall restart vector and QEX results in a MAP score of 0.0242, while using one overall restart vector without QEX results in a MAP score of 0.0315. This shows that the ICG can improve the text-baseline (MAP = 0.0298).

---

<sup>1</sup> Available at [http://trec.nist.gov/trec\\_eval/](http://trec.nist.gov/trec_eval/)

<sup>2</sup> The TRECVID notebook paper giving full details is available at <http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/glasgow.pdf>

Run ID	MAP	P(10)	P(NR)	Recall
UG_F_1	0.03	0.121	0.077	0.148
UG_F_2	0.005	0.075	0.023	0.051
UG_F_3	0.018	0.146	0.052	0.11
UG_F_4	0.021	0.05	0.072	0.143
UG_F_5	0.004	0.1	0.025	0.059
UG_I_1	0.047	0.558	0.076	0.067

Table 2: Evaluation Measures for various runs in TRECVID 2006

Topic	UG_F_1	UG_F_2	UG_F_3	UG_F_4	UG_F_5	UG_I_1
173	0.0132	0.0013	0.0065	0.0139	0.0026	0.0051
174	0.0007	0.0027	0.0008	0.0023	0.0126	0.0106
175	0.0009	0.0004	0	0.0002	0.0004	0.0394
176	0.0118	0	0.0066	0.0012	0	0.0006
177	0.0437	0.0003	0.0092	0.0075	0.0006	0.0222
178	0.1854	0.0001	0.0863	0.1112	0.0002	0.139
179	0.0689	0.0001	0.0224	0.006	0.0001	0.1178
180	0	0	0.0005	0.0002	0.0002	0.0225
181	0.0066	0.0001	0.0034	0.0051	0	0.1559
182	0.0564	0.0015	0.0071	0.0072	0.0115	0.0274
183	0.0094	0.0014	0.002	0.0094	0.0018	0.036
184	0.01	0.0036	0.0026	0.0027	0.0107	0.0165
185	0.0028	0.0002	0.0004	0.0062	0.0015	0.0115
186	0.0028	0.0015	0.0053	0.0025	0.0009	0.012
187	0.0371	0.0094	0.0065	0.0195	0.0005	0.0627
188	0.1341	0.0002	0.0675	0.0549	0.0009	0.0595
189	0	0.0161	0.0006	0.0039	0.025	0.0035
190	0	0.0001	0.0003	0.0006	0.0006	0.0229
191	0.002	0.0027	0.0018	0.0018	0.0057	0.0279
192	0.0011	0.0018	0.0012	0.0017	0.001	0.0034
193	0.0019	0.0002	0.0028	0.0064	0.0025	0.0042
194	0.0185	0	0.0667	0.0664	0.0001	0.1659
195	0.0577	0.0798	0.0291	0.0337	0.0252	0.0849
196	0.0504	0.0023	0.11	0.1355	0.0029	0.0878
all	0.0298	0.0052	0.0183	0.0208	0.0045	0.0475

Table 3: MAP for runs and Topics for TRECVID 2006

### 3.4 TRECVID 2007<sup>3</sup>

This year GU participated in the summarisation and automatic search task, whereas in the previous year, automatic and interactive search results were submitted. Two fully automatic runs were submitted, amongst the automatic runs, one (UG\_F\_Sys1) is based on matching SIFT (Scale-invariant feature transform, [Lowe 1999, 2004]) features and the other (UG\_F\_Sys2) is based on adapted SIFT features

<sup>3</sup> The TRECVID notebook paper giving full details is available at [http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/glasgow\\_university.pdf](http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/glasgow_university.pdf)

and annotated data. The major feature for both runs are the SIFT features. For UG\_F\_Sys1, we used one representative keyframe from the shots. For UG\_F\_Sys2, from each shot, irrespective of their length, five representative keyframes, at regular intervals depending on the shot length, were used for retrieval.

The results of the submitted runs are given in Table 4. The run UG\_F\_Sys1, which was solely based on SIFT features, performed poorly. However, UG\_F\_Sys2 which used information from keywords/ annotations, performed better for a few topics, such as 220 (gray scale shots of a street with one or more buildings and one or more people), 206 (Shots with hills or mountains visible), 214 (shots of very large crowd of people filling more than half of field of view), and was at its best for 218 (people playing musical instruments). For all other queries, the performance was average, being roughly at the medium for the systems performing at TRECVID 2007.

Run ID	MAP	P(10)	P(NR)	Recall
UG_F_Sys1	0.001	0.025	0.008	0.041
UG_F_Sys2	0.017	0.046	0.040	0.139

**Table 4: Evaluation Measures for submitted runs in TRECVID 2007**

Topic	UG_F_Sys1	UG_F_Sys2
197	0.0005	0.0002
198	0.0005	0.0020
199	0.0022	0.0103
200	0.0016	0.0019
201	0.0003	0.0036
202	0.0002	0.0002
203	0.0001	0.0004
204	0.0002	0.0027
205	0.0004	0.0002
206	0.0040	0.0231
207	0.0018	0.0068
208	0.0000	0.0004
209	0.0014	0.0003
210	0.0001	0.0009
211	0.0002	0.0002
212	0.0001	0.0053
213	0.0008	0.0004
214	0.0000	0.0184
215	0.0006	0.0016
216	0.0026	0.0031
217	0.0011	0.0126
218	0.0008	0.2639
219	0.0000	0.0015
220	0.0006	0.0547
All	0.0008	0.0173

**Table 5: MAP per Topic for TRECVID 2007**

### 3.5 TRECVID 2008<sup>4</sup>

In TRECVID 2008 GU submitted five fully automatic runs and one interactive run:

- UG-ASR-6 Text baseline (required to participate in TRECVID): This used the same techniques as in the text baseline run in TRECVID 2006

<sup>4</sup> The TRECVID notebook paper giving full details is available at <http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/glasgow.pdf>

- UG-AnLLF\_5 Automatic search using only visual features (weighted feature selection): Only visual features were used (the same MPEG-7 features as used in TRECVID 2006). A weighted feature selection technique was developed which used the diversity of query keyframes
- UG-TYRun1\_2 Automatic search using classifiers: Keyframes were classified using an SVN into seven categories: “city”, “human”, “indoor”, “nature”, “night”, “outdoor”, and “vehicle”. These categories correspond to the semantic meaning of the topics of TRECVID 2008.
- UG-TYRun2\_3 Automatic search using classifiers and indexing structures: Uses the same approach as in Run1\_2 above, but using a reduced feature dimensionality
- UG-AnHLF\_4 Automatic search using visual features and high level features: This run combined the techniques used in UG-AnLLF\_5 with the use of High-level features as provided by the TRECVID organisers, and the output of a face detection system. The retrieval and fusion used was similar to UG-TYRun1\_2 and UG-TYRun2\_3
- UG-Int\_1 Interactive search run based on text and visual query examples. This used an early version of the “VIGOR” interface, described in Chapter 7.

The overall results are shown in Table 6, show automatic runs using high level features performed better than the other runs. From the overall performance, UG-Int has the best P@10. Precision at total relevant shots, for UG-Int and UG-AnHLF is almost the same, with UG-AnLLF and UG-ASR having similar performance. UG-Int has the best overall MAP, closely followed by UG-AnHLF and UG-ASR. When it comes to Recall, one can see that UG-AnLLF and UG-TYRun1 have almost the same performance achieving 19% and 17% of recall respectively, but, when these are compared with respect to the time response, UG-TYRun1 outperforms, any runs as it consumes hardly 2% of the time taken by UG-AnLLF. The best recall achieved is 25% from UG-AnHLF run submission.

Considering specific topics, we have a varied performance as can be seen from Table 7. UG-ASR has the best results for the topic #245, “shots of a person watching a television screen - no keyboard visible” with MAP of 0.3109. MAP for any other topic is below 0.1. ASR has also higher MAP for #246, #247 and #256 which are the topics “shots of one or more people with one or more animals” and “shots of one or more people, singing and/or playing a musical instrument”. UG-AnLLF is best for topics #221, “shots of a person opening a door”, #225, “shots of a bridge”, and #227, “shots of a person's face filling more than half of the frame area” which are more natural and consistent in colour, texture and edge features.

UG-AnHLF worked better for many topics” #222, #230, #246, #248, #249, #250, #257, #262, and #263, which were shots related to, “3 or fewer people sitting at a table”, “one or more vehicles passing the camera”, “one or more people in a kitchen”, “a crowd of people, outdoors, filling more than half of the frame area”, “classroom scene”, “an airplane exterior”, “a plant that is the main object inside the frame area”, “one or more people in white lab coats”, and “one or more ships or boats in the water”. Most of these topics being based on number of people benefited from the face detector. The other topics benefited from the high level features, specifically: classroom, boat-ship, and airplane.

UG\_Int also performed better for many topics #255, #256, #257, #258, #260, #261, #262, #264, #267, #268, such as “just one person getting out of or getting into a vehicle”, “one or more people, singing and/or playing a musical instrument”, “one or more people sitting outdoors”, “one or more animals - no people visible”, “one or more coloured photographs, filling more than half of the frame area”, “the camera zooming in on a person's face”, “one or more signs with lettering” which are more semantic driven and difficult to retrieve with only low level.

A few topics have the same MAP irrespective of the method and feature used, for instance, UG-ASR and UG-AnHLF for #246, UG-Int and UG-AnHLF for #262. Though, UG-TYRun1 and UG-TYRun2 do not have the best MAP, it still falls in slightly below with the performance of UG-AnLLF and has the same MAP for topic #226.

Run ID	MAP	P(10)	R-prec	Recall
UG-ASR	0.0124	0.0787	0.0390	0.0149
UG-AnLLF	0.0092	0.0792	0.0413	0.0191
UG-TYRun1	0.0058	0.0479	0.0285	0.0174
UG-TYRun2	0.0019	0.0250	0.0134	0.0094
UG-AnHLF	0.0153	0.0937	0.0517	0.0253
UG-Int	0.0243	0.2792	0.0535	0.0071

**Table 6: Resultant performance of various runs in TRECvid 2008**

Topic	UG-ASR	UG-AnLLF	UG-TYRun1	UG-TYRun2	UG-AnHLF	Topic	UG-ASR	UG-AnLLF	UG-TYRun1	UG-TYRun2	UG-AnHLF	UG-Int
221	0.0010	0.0202	0.0042	0.0007	0.0202	245	0.3109	0.012	0.0294	0	0.012	0.0051
222	0.0046	0.0054	0.0057	0.0012	0.0353	246	0.0215	0.0037	0.005	0.0005	0.022	0.0063
223	0.0101	0.0011	0.0004	0.0004	0.0007	247	0.0351	0.0059	0.0009	0.0004	0.0123	0.0194
224	0.0008	0.0019	0.0094	0.0029	0.0019	248	0.0007	0.0254	0.0172	0.0006	0.0855	0.0131
225	0.0026	0.025	0.001	0.0003	0.0023	249	0.0009	0.009	0.0021	0.0009	0.0351	0.0134
226	0.0046	0.0246	0.0236	0.0116	0.0108	250	0.0015	0.0079	0.0088	0.0008	0.0423	0.0132
227	0.0012	0.0194	0.0054	0.0019	0.0276	251	0.0173	0.0007	0.0006	0.0013	0.0007	0.0086
228	0.0011	0.0088	0.0127	0.0024	0.0088	252	0.0041	0.0039	0.0007	0.0033	0.0039	0.0542
229	0.0031	0.0043	0.0016	0.0014	0.0118	253	0.0021	0	0.001	0.0001	0.0001	0.012
230	0.0017	0.017	0.0087	0.0024	0.0328	254	0.0011	0.002	0.0015	0.0011	0.002	0.0145
231	0.0024	0.008	0.001	0.0004	0.008	255	0.0094	0.0128	0.0029	0.0003	0.0128	0.0819
232	0.0001	0.008	0.0069	0.0009	0.008	256	0.0226	0.0028	0.0016	0.0006	0.0028	0
233	0.0006	0.0004	0.001	0.0018	0.0004	257	0.0001	0.0211	0.0249	0.0037	0.0506	0.0266
234	0.0004	0.0028	0.0025	0.0013	0.0082	258	0.0008	0.0023	0.0014	0.006	0.0023	0.0235
235	0.0001	0.0032	0.0003	0.0006	0.012	259	0.0116	0.019	0.004	0.0007	0.0204	0.001
236	0.0002	0	0.0001	0.0014	0	260	0.0069	0.0019	0.0023	0.0007	0.0019	0.0238
237	0.0011	0.0029	0.0072	0.003	0.0042	261	0.0005	0.003	0.0029	0.0006	0.003	0.0293
238	0.0000	0.0008	0.0031	0.0003	0.0008	262	0.0272	0.0641	0.0013	0	0.0641	0.067
239	0.0023	0.0092	0.006	0.002	0.0164	263	0.005	0.0191	0.0121	0.001	0.0662	0.0227
240	0.0071	0.0061	0.0007	0.0001	0.0061	264	0.0061	0.0002	0.0008	0.002	0.0002	0.0385
241	0.0085	0.0048	0.0015	0.0015	0.0048	265	0.002	0.0144	0.0135	0.0051	0.0165	0.0031
242	0.0000	0.0013	0.0002	0.0003	0.006	266	0.0064	0.0047	0.012	0.0002	0.0047	0.0068
243	0.0052	0.0002	0.0002	0.0021	0.0002	267	0.021	0.0233	0.0148	0.0143	0.0246	0.0753
244	0.0096	0.0037	0.0028	0.0005	0.0168	268	0.0124	0.0045	0.0106	0.0046	0.0045	0.0249

**Table 7: MAP per topic in TRECvid 2008**

### 3.6 TRECvid 2009<sup>5</sup>

In 2009 GU submitted six fully automatic runs:

- UG-PURun1\_1 Search results using visual features (Colour Histogram, Edge Histogram and Homogenous Texture) on a reduced search domain combined with high level features with weighted late fusion.
- UG-PURun2\_2 Search results using MPEG7 visual features (Colour Histogram, Edge Histogram and Homogenous Texture), with adaptive feature weighting. The proposed method comprises of three stages. The first stage deals with the feature selection mechanism which selects the feature that preserves the diversity of visual features of the query examples. A mechanism to push the relevant yet diverse results towards the top of the result list frames the second stage. Finally, the third stage combines the results originating from various query examples and various features.
- UG-PURun3\_3 used the same techniques as in UG-PURun2\_2 combined with high level features as extracted by the top five teams in TRECvid 2008.
- UG-RRRun4\_4 Search using bag of words generated for each low level visual features.
- UG-HERun5\_5 Search results using LDA (Latent Dirichlet Allocation) based image retrieval approach using SIFT features.

<sup>5</sup> The TRECvid notebook paper giving full details is available at <http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/glasgow.pdf>

- UG-HERun6\_6 Search results using LDA based image retrieval approach, similar to UG-HERun5\_5

Due to the huge size of the collection, we used one frame per shot for our runs. Table 8 shows the results of various runs submitted by UG. The run **UG-PURun1\_1** has the best performance in terms of P10. The results generated in this run combined the low level features with the high level features. A smaller collection was comprised by pooling all shots annotated with the synonyms of the topic description. A search based on low level features was executed in this small collection to rank these results. If the number of shots in this collection was less than 1000, then the results from only the visual features were appended to the tail of the list. Since, the top of the list definitely consisted of the shots annotated with high level features, it resulted with better precision. UG-PURun3\_3, UG-HERun5\_5, UG-HERun6\_6 have almost the same MAP. This suggests that the LDA based retrieval using SIFT feature performs equally or even better than the run based on combining low level features and the top results of high level features.

Run ID	MAP	P(10)	R-prec	infAP	Recall
UG-PURun1_1	0.0094	0.1542	0.0245	0.0094	0.0395
UG-PURun2_2	0.0047	0.0958	0.0309	0.0045	0.0562
UG-PURun3_3	0.0119	0.1333	0.0424	0.0118	0.0626
UG-RRRun4_4	0.0003	0.0208	0.0063	0.0003	0.0129
UG-HERun5_5	0.0122	0.0833	0.0347	0.0113	0.0806
UG-HERun6_6	0.0132	0.0833	0.0358	0.0121	0.0785

**Table 8: Resultant performance of various runs in TRECVID 2009**

Topic	UG-PURun1_1	UG-PURun2_2	UG-PURun3_3	UG-RRRun4_4	UG-HERun5_5	UG-HERun6_6	Best of UG	Best in TRECVID09
269	0.0062	0.0072	0.0006	0.0005	0.0000	0.0000	0.0072	0.191
270	0.0198	0.0125	0.0790	0.0000	0.0331	0.0102	0.079	0.355
271	0.0009	0.0019	0.0007	0.0001	0.0012	0.0014	0.0019	0.202
272	0.0000	0.0007	0.0041	0.0005	0.0001	0.0000	0.0041	0.133
273	0.0012	0.0005	0.0076	0.0001	0.0002	0.0001	0.0076	0.257
274	0.0007	0.0021	0.0004	0.0001	0.0004	0.0006	0.0021	0.085
275	0.0006	0.0005	0.0002	0.0003	0.0004	0.0006	0.0006	0.019
276	0.0007	0.0022	0.0021	0.0000	0.0009	0.0006	0.0022	0.579
277	0.0046	0.0025	0.0004	0.0010	0.0001	0.0005	0.0046	0.222
278	0.0200	0.0090	0.0208	0.0004	0.0151	0.0117	0.0208	0.294
279	0.0000	0.0001	0.0000	0.0000	0.0000	0.0002	0.0002	0.006
280	0.0001	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.043
281	0.0466	0.0016	0.0085	0.0006	0.0001	0.0003	0.0466	0.111
282	0.0003	0.0024	0.0017	0.0001	0.0001	0.0023	0.0024	0.061
283	0.0001	0.0018	0.0070	0.0001	0.0004	0.0005	0.007	0.074
284	0.0044	0.0016	0.0582	0.0001	0.0016	0.0045	0.0582	0.346
285	0.0002	0.0376	0.0242	0.0006	0.2228	0.2693	0.2693	0.488
286	0.0001	0.0006	0.0002	0.0004	0.0022	0.0011	0.0022	0.209
287	0.0093	0.0057	0.0029	0.0002	0.0043	0.0010	0.0093	0.27
288	0.0000	0.0013	0.0016	0.0001	0.0002	0.0002	0.0016	0.038
289	0.0035	0.0139	0.0022	0.0019	0.0065	0.0063	0.0139	0.145
290	0.1068	0.0019	0.0571	0.0000	0.0022	0.0053	0.1068	0.369
291	0.0001	0.0003	0.0003	0.0001	0.0005	0.0003	0.0005	0.023
292	0.0001	0.0062	0.0067	0.0000	0.0001	0.0000	0.0067	0.018

**Table 9: MAP per Topic for TRECVID 2009**

### 3.7 Summary

The UG TRECVID evaluation effort has evaluated a number of different techniques and methods on the different years data. The first efforts were based on relatively simple global MPEG-7 visual features, text, and the ICG technique. As can be expected, text retrieval, with its long and successful history, works best when available. ICG approach was promising, but suffers from slow search times making its use in interactive systems difficult. Attempts were made to speed the algorithm up, but without significant gains being made.

TRECVID 2007 was used as a testbed for the use of SIFT features, which have a good reputation in the research community. The results from this year were disappointing, however, this work was restarted in TRECVID 2009 using the LDA approach, which does show promise. The second run in 2007 improved performance considerably, but relied on high-level features and annotations in the collection. Ideally, the aim is to perform well with automatically extracted visual features only, such as the global MPEG-7 features, or SIFT, although such extra runs do provide a useful comparison with fully automatic runs.

In TRECVID 2008 a number of different approaches were tried: a face detector was found to be useful for topics requiring the identification of people, a classification system was developed and evaluated which classed video shots into seven different categories, and a prototype ViGOR system was used in an interactive run. Unfortunately the classification approach did not perform well, which resulted in a different techniques being investigated in TRECVID 2009. In 2009 SIFT features were used again along with the LDA technique, and also clustered to generate “bad of visual words”: the former LDA technique shows promise, performing considerably better than the previous SIFT results in TRECVID 2007 (UG\_F\_Sys1 in Section 3.3<sup>6</sup>), and is the subject of further study.

---

<sup>6</sup> Please note that UG\_F\_Sys2 is not comparable to this later run, due it it's use of high-level features and annotations.

## 4 Automatic Image Annotation

---

One major current research issue in image and video retrieval is that of Automatic Image Annotation (AIA): i.e. the association of some high-level concept with an image or video shot. For example, an image containing a “picture of a car”, may be annotated with the concept “car”. Once such annotations are attached, retrieval then becomes the much simpler problem of indexing and searching text annotations. Achieving such automatic annotation, however, is a difficult ongoing research problem. In this Chapter, some results into AIA work carried out as part of the SALERO project is briefly presented, along with a framework for the evaluation of AIA systems.

### 4.1 Introduction

---

Multimedia content, like images, video and audio, does not lend itself to traditional indexing methods such those applied on text documents. The main difficulty arises from the nature of the information encoded in these types of media. While the basic structural units of a text document are words, which directly convey the message of the document in a human understandable form, there is no corresponding analogous to words in media such as images and audio. To make indexing and retrieval of multimedia content feasible, features similar to words have to be extracted from such media and queries using the same feature representation can be answered using the similarity of the query and the media with respect to these features.

Towards this direction, most Content Based Image Retrieval (CBIR) systems employ signal processing and image analysis techniques to represent images in terms of colour, texture and edge distributions. Querying such systems requires the presentation of an image example of the user's information need. Images which are similar to the query image in terms of colour, texture and edge distributions are returned to the user. Success of such systems, however, is difficult due to the semantic gap [Smeulders 2000]. A solution towards bridging this gap between visual similarity, as expressed by similarity of low-level image features, and semantic similarity is to directly associate to images semantic features such as words. Indeed, this is the approach followed by many commercial image archives such as GettyImages<sup>1</sup>, Corbis<sup>2</sup> etc. Moreover, users and communities are also interested in indexing their personal collections using semantic features in order to render them accessible by others. This is evident from the success of image sharing web applications such as Flickr<sup>3</sup>, Picasa<sup>4</sup> and others.

There is therefore a profound need to devise algorithms that will, even at least partially, automate the process of annotating images. Several algorithms in this direction have been proposed lately, with most notable examples those in [Duygulu 2002], [Lavrenko 2003], [Feng 2004], [Blei 2003] and [Carneiro 2007]. In Section 4.2 we briefly present the evaluation results of one AIA technique developed during SALERO.

After this, the rest of this chapter will deal with the evaluation of AIA systems: many AIA models have been traditionally compared only on the “easy” dataset provided by Duygulu et al. [Duygulu 2002]. Some of them [Carneiro 2007] have been evaluated on more realistic collections as well, such as the TRECVID News dataset in order to support certain assumptions and statements regarding real-life multimedia collections. However, it is unclear whether the reported results are due to the descriptive power of the model, or are simply artefacts of the discriminating power of the employed descriptor in combination with the collection.

In Sections 4.3 and 4.4, we describe an evaluation framework which addresses this problem. We argue that a more comprehensive evaluation of AIA models is needed in order to show that the models' assumptions actually hold and that results are neither collection nor descriptor-specific. In light of this, a framework for evaluation and comparison of AIA models is suggested, which incorporates various collections and standardized content descriptors. It essentially defines a set of test collections, a sampling method which attempts to extract normalised and self-contained samples, a variable-size block segmentation technique with varying degrees of overlapping and a set of multimedia content descriptors.

### 4.2 An Example Approach to Image Annotation

---

In this section we briefly consider the problem of accurate density estimation in the domain of image annotation. The work is built upon that of Carneiro et. al. [Carneiro 2007], which has achieved the best so far published results on the Corel 5K benchmark collection, for AIA. In particular, we adopt the ideas

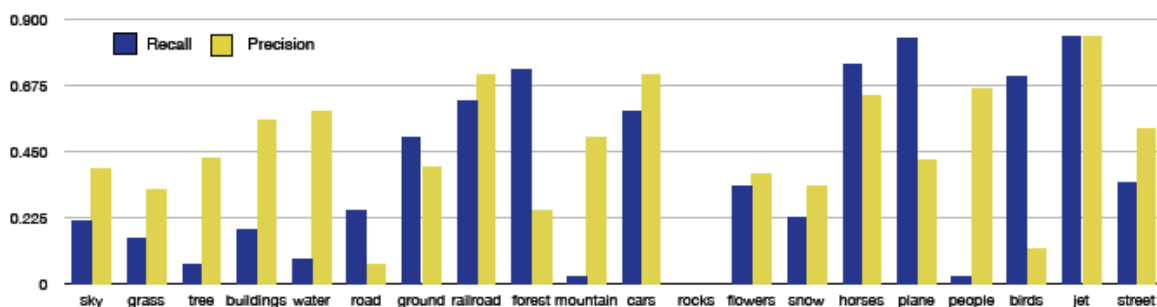
of learning mixture hierarchies of multivariate Gaussian components from [Vasconcelos 1998] and propose a Bayesian treatment which provides regularized solutions and does not suffer from the numerical difficulties of maximum likelihood methods. To provide an algorithm with low computational complexity we used approximate inference and in particular variational learning ([Jordan 1998], [Attias 2000]) which yields an algorithm with slightly more computations but with the same order of complexity as the maximum likelihood algorithm of Vasconcelos and Lippman [Vasconcelos 1998]. Furthermore, we incorporate our algorithm in the supervised learning framework of [Carneiro 2007] and [Yavlinsky 2005] and validate our approach by conducting experiments on the Corel 5K dataset.

The approach taken uses a Bayesian Hierarchical Gaussian Mixture Model (BHGMM), which is fully described in [Stathopoulos 2009]. To validate this Model, we incorporate it in the supervised learning framework of [Carneiro 2007] and [Yavlinsky 2005] and apply it on the Corel 5k collection, where the aim is to annotate the images with a set of annotation words, which correspond to different semantic concepts.

Results for the technique are shown in Table 10 and Table 11. In total the technique predicts 116 words with non zero recall. The mean per-word precision and recall over these 116 words are 0.36 and 0.48 respectively. A comparison of our approach (BHGMM) with the Continuous-space Relevance Model (CRM) [Lavrenko 2003] and the Supervised Multi-class Learning (SML) [Carneiro 2007] in terms of average per-word recall and precision is shown in Table 10. Averages are over all the 260 words in the test vocabulary. In Figure 1 the precision and recall for the 20 most frequent words in the vocabulary are shown.

Model	CRM	SML	BHGMM
Words with non zero recall	107	137	116
Mean per-word recall	0.19	0.29	0.21
Mean per-word precision	0.16	0.23	0.17

**Table 10: Performance comparison of automatic image annotation**



**Figure 1: Precision/Recall for the 20 most frequent words in the vocabulary obtained by BHGMM**

Although we extend the algorithm of [Carneiro 2007] we do not manage to achieve better results. We hypothesize that the reason is the simplistic approach for initializing the mixture models in our algorithm. It is well known in the literature [Bishop 2006] that the initialization of mixture models can have a significant effect in their performance. A common strategy for initializing Gaussian mixture models, is to use a k-means clustering to find initial estimates of the means and covariance matrices of the Gaussian components. In [Carneiro 2007] it is not reported how the mixture models are initialized while parameters of the algorithm are heavily optimized.

In summary: our method out-performs the previously published continuous-space relevance model [Lavrenko 2003] while it performs comparable to the method of Carneiro et. al [Carneiro 2007]. Even though we do not use sophisticated initialization methods and parameter optimization, we are still competitive with the current state of the art, which is encouraging for further work towards this direction. This experiment was conducted using a standard image database, the Corel 5k collection. In the next section we consider the problem of evaluating AIA systems within a broader, more realistic framework.

## 4.3 A Framework for Evaluating Automatic Image Annotation Algorithms

---

Many AIA systems have been traditionally compared only on the “easy” dataset provided by Duygulu et al. [Duygulu 2002]. Some of them [Carneiro 2007] have been evaluated on more realistic collections as well, such as the TRECVID News dataset, in order to support certain assumptions and statements regarding real-life multimedia collections. However, it is unclear whether the reported results are due to the descriptive power of the model, or are simply artefacts of the discriminating power of the employed descriptor in combination with the collection. We argue that a more comprehensive evaluation of AIA models is needed in order to show that the models' assumptions actually hold and that results are neither collection nor descriptor-specific.

Regarding multimedia collections, facts such as whether images depict single or multiple objects, and whether an annotation implies dominance of an object or simply its presence are some examples of these factors. Moreover a collection could be strongly or weakly labelled, depending on whether all instances of an object are annotated or not, while the existence of object hierarchies having tags such as “cat” and “tiger”, “car” and “exotic car” or “water” and “ocean” might not only affect the performance of the algorithm, but also the results that one would expect. Collections also define the level of semantics that an algorithm should target for. Searching for objects is a totally different task than searching for scene categories or emotional states. It would perhaps require a different way of treating images namely segmenting and representing, thus again modifying the overall setting on which the algorithm would have to operate.

As such, an evaluation of a set of image classification algorithms would simply be incomplete, if it did not involve testing these algorithms on various settings in order to prove their robustness, namely whether they perform equally well under various settings. Therefore, a set of three multimedia collections was selected to be incorporated in our evaluation procedure. These are the Corel 5K [Duygulu 2002], TrecVid 2007 [Ayache 2007] and Caltech 101 [Fei-Fei 2006] collections.

Corel 5K is considered a rather easy setting, since Global Colour Features alone are considered to provide enough discriminative power for this collection. It was first used by Duygulu et al. [Duygulu 2002] in the field of automatic image annotation algorithms, while since then, it has been used by each new model in the literature, in order for the results to be comparable to previously proposed models. The TrecVid 2007 dataset on the other hand comprises an extremely challenging setting. Since it is intended to be used for several high level tasks such as shot boundary detection and high level feature extraction, one can appreciate that using this dataset in the AIA domain will be equally difficult and unpredictable. Caltech 101 has a major advantage over other multimedia datasets, in that each image depicts a single object, thus removing any confusion associated with the multiple-labels paradigm. As such, it can be employed to learn precisely the class and non-class model of certain categories and objects. It is the only collection which can allow for a sample which is fair towards all categories, namely it has the same number of images describing each category, and still being consistent and self-contained.

### 4.3.1 Sampling Procedure

The afore-mentioned collections however were not used as a whole; rather we used a sampling procedure to extract a smoother and self-contained representative sample of each collection. By smoother, we mean that most of the tags would contain approximately the same number of images, and only a few, if any, would be described by significantly more example images. By self-contained, we mean that no matter how popular a tag already is, we would not discard any of its instances, as this would harm the class and non-class models. This sampling process was performed mainly because all of these collections have a highly unbalanced distribution of images over classes. There are a lot of classes which are inadequately described, a set of classes with a reasonable number of images belonging to them and a few which are very popular and frequent within each collection. Using the whole collections would probably create an easier setting for all of the algorithms for two reasons. When evaluating such an algorithm, popular tags would be more likely to be selected to be tested, while on the other hand, when classifying an image it would be more likely to annotate it with a more frequent tag. Moreover, we did not want to allow models to exploit attributes of collections which were unrelated to visual information, such as tag popularity. Hence, a sampling procedure was applied on all of the collections, which attempted to smooth these settings removing extreme conditions, namely classes

which were either inadequately or very precisely described, while at the same time preserving the rest of the attributes of these collections.

#### 4.3.2 Content Descriptors

Image representation and feature extraction is an important and definitive step when attempting to use an automatic image annotation algorithm. It is important to identify the appropriate set of features, one which would provide not only the appropriate level of discrimination among images, but also enough compactness, so that the algorithm itself will not suffer from the challenging problems of computational complexity, immense resource requirements and the curse of dimensionality. In addition, it is not unusual for a multimedia collection to be known to yield better results when used in combination with a specific set of features, while on the other hand, certain image classification algorithms also perform better when used with certain sets of features. Hence an evaluation of image classification algorithms incorporating various features sets representing different attributes and characteristics of the same images from the same collections might shed some light into the operation of these algorithms through their variation in performance when applied on various such settings of collections and features sets.

When deciding on the features sets which would be incorporated in the evaluation process, the objective was to use standardised features sets, no matter how well they would actually perform. The goal of the present work was not to get better results, but to investigate patterns in the relative performance and the presence of any consistency between certain image classification algorithms. As such, by using colour and texture features defined in the MPEG-7 Standard [Manjunath 2002], such as Colour Histogram (CH), Edge Histogram (EH), and Homogeneous Texture (HT), as well as SIFT features introduced in 2004 by Lowe [Lowe 2004], it would be clear that we did not act in favour of a specific algorithm, while the results of this work would still be meaningful in the future, as it would be straightforward to implement a new algorithm, run experiments on the same collections using these standardised features sets and get comparable results.

## 4.4 Results

---

The previously described framework was used to evaluate and compare two state-of-the-art image classification models, namely the Multiple Bernoulli Relevance Model (MBRM) [Feng 2004] by Feng et al. and the Supervised Multiclass Labelling (SML) introduced by Carneiro et al. [Carneiro 2007]. In Table 1, results of experiments with MBRM and SML using MPEG-7 and SIFT Features respectively are presented for the three collections. Our results are significantly lower than the ones reported in the original papers [Feng 2004, Carneiro 2007]. The reason for this is that we used normalised parts of the collections, as well as other sets of features. On the other hand, in Table 2, the MBRM is contrasted to the simpler Support Vector Machines (SVM) approach using the SVM-light implementation [Joachims 1999].

First of all, with respect to the collections, we would say that Corel was the most “extreme” setting, followed by that of TrecVid 2007, and then the completely normalised sample of Caltech 101. By “extreme”, we mean that only a few tags were more popular than others, while these had significantly more example images.

From Table 11, we can see that the variance of both Precision and Recall around the means was significantly high. We also see that only a small percentage of tags has Recall>0 and most of these tags are popular tags in the collection. This is similar to previously reported results [Jeon 2003, Feng 2004, Carneiro 2007] on the Corel 5K collection. However, since we have removed most of the popular tags the numbers tend to be significantly smaller. This shows that previous optimistic results on Corel 5K are actually due to the tag distribution rather than the descriptive ability of the models. Interestingly, MBRM would always return the most popular words when evaluated on Corel 5K and TrecVid 2007. On the contrary, in Caltech 101, in which tag frequencies were completely normalised, more words were returned and the diversity among them was high. Also, regarding the TrecVid dataset, we see that MBRM had exactly the same response across all descriptors, meaning that similarity across images was not taken into account by the model. On the other hand, SML achieved the best performance on TrecVid 2007, followed by Corel 5K and Caltech 101. The bad performance on Caltech might be due to the fact that it is a single-label environment, and the actual number of classes depicted in an image was considered during the annotation process. The difference in performance between Corel 5K and TrecVid 2007 might be either due to the visual content of the images, or due to collection-specific properties. Nevertheless, overall in all collections, our results are not as optimistic as previously reported ones, and this seems to be related to the normalised tag distributions of our samples.

However, although different categories of features were used with each model, the results between them are still comparable and can be interpreted in a generic way.

Moreover, we applied a Support Vector Machine using global MPEG-7 features on the Corel 5K collection and compared it with MBRM and SML. Results are presented in Table 12, where we can see that a simple SVM with global features achieves better results than MBRM and SML, which are considered state-of-the-art methods. We have also implemented a SVM with local MPEG-7 features by using k-means to cluster local features and create visual terms. The local features are associated to their closest visual term (cluster centroid) and images are represented by the frequency of the visual terms they contain, similar to a bag of word model used in Information Retrieval. Despite the quantisation errors introduced by the k-means algorithm, results are still better than MBRM and SML although not as good as using the SVM directly on the global MPEG-7 features.

Finally, with respect to SML, it was not feasible to combine it with local MPEG-7 Features. The image segmentation procedure which was used for extracting MPEG-7 local features led to a quite homogeneous representation of each image individually. The MBRM was not affected by this homogeneity since features were homogeneous only at the image level. SML however was not able to cluster the feature vectors representing each image with a mixture model of a reasonable number of components. As SML uses a mixture of Gaussians, it essentially makes strong assumptions about the nature and the properties of the features, thus making it feature-dependent. Hence, the SML would require a significantly larger dataset, and a descriptor which would provide an appropriate degree of heterogeneity at the image level.

Collections	Corel 5K				TrecVid 2007				Caltech 101			
Models	MBRM		SML		MBRM		SML		MBRM		SML	
Descriptors	CH	EH	HT	SIFT	CH	EH	HT	SIFT	CH	EH	HT	SIFT
# of words in total	70				30				50			
# of words with Recall>0	4	4	4	6	8	8	8	9	18	14	13	2
Precision and Recall on all words												
Mean Per-word Recall	0.034	0.045	0.045	0.046	0.194	0.194	0.194	0.130	0.125	0.265	0.270	0.015
Variance in Recall	0.151	0.193	0.193	0.175	0.356	0.356	0.356	0.269	0.207	0.275	0.286	0.077
Mean Per-word Precision	0.020	0.010	0.010	0.003	0.163	0.163	0.163	0.073	0.127	0.286	0.251	0.0009
Variance in Precision	0.121	0.044	0.044	0.011	0.296	0.296	0.296	0.160	0.214	0.316	0.268	0.005
Precision and Recall on words with Recall > 0												
Mean Per-word Recall	0.569	0.750	0.750	0.495	0.534	0.534	0.534	0.397	0.222	0.377	0.422	0.360
Variance in Recall	0.284	0.238	0.238	0.303	0.295	0.295	0.295	0.320	0.206	0.182	0.174	0.125
Mean Per-word Precision	0.334	0.172	0.174	0.036	0.449	0.449	0.449	0.188	0.227	0.360	0.284	0.021
Variance in Precision	0.374	0.054	0.054	0.012	0.232	0.232	0.232	0.235	0.218	0.244	0.217	0.015

**Table 11 Mean Precision and Recall of MBRM (MPEG-7) and SML (SIFT).**

Collections	Corel 5K						Caltech 101				TrecVid 2007					
Models	MBRM		SVM		MBRM		SVM		MBRM		SVM		MBRM		SVM	
Descriptors	CH	GCH	CH	EH	GEH	EH	CH	GCH	EH	GEH	CH	GCH	EH	GEH		
# of words in total	70						50				30					
# words (Recall>0)	4	32	26	4	37	29	18	20	14	23	8	15	8	16		
Precision and Recall on all words																
Mean Recall	0.034	0.204	0.102	0.045	0.402	0.314	0.125	0.327	0.265	0.580	0.194	0.405	0.194	0.611		
Recall Variance	0.151	0.236	0.159	0.193	0.430	0.250	0.207	0.221	0.275	0.325	0.356	0.265	0.356	0.374		
Mean Precision	0.020	0.131	0.051	0.010	0.242	0.193	0.127	0.372	0.286	0.740	0.163	0.454	0.163	0.564		
Precision Variance	0.121	0.226	0.087	0.044	0.301	0.183	0.214	0.158	0.316	0.363	0.296	0.347	0.296	0.336		
Precision and Recall on words with Recall > 0																
Mean Recall	0.569	0.149	0.173	0.750	0.188	0.245	0.222	0.173	0.377	0.300	0.534	0.227	0.534	0.265		
Recall Variance	0.284	0.095	0.144	0.238	0.195	0.169	0.206	0.080	0.182	0.124	0.295	0.140	0.295	0.144		
Mean Precision	0.334	0.173	0.087	0.172	0.193	0.139	0.227	0.142	0.360	0.057	0.449	0.009	0.449	0.028		
Precision Variance	0.374	0.269	0.092	0.054	0.294	0.182	0.218	0.234	0.244	0.288	0.232	0.258	0.232	0.288		

**Table 12 Comparison between MBRM and SVM using MPEG-7 Descriptors.**

## 4.5 Conclusion

In this chapter, we very briefly presented some results from a AIA developed as part of SALERO, and then we considered the lack of proper evaluation in the domain of Automatic Image Annotation, this latter subject providing the main topic of the chapter. We found that the evaluation methodologies followed by AIA researchers are insufficient and do not support and prove the models' initial assumptions. Hence, we defined an Evaluation Framework, which is comprised by more than one multimedia collections and standardised descriptors, uses a sampling method to extract smoother, self-contained and representative samples and a multi-resolution block-segmentation method. We used this framework to evaluate and compare two state-of-the-art AIA models and we found that they heavily depend on the underlying test set. MBRM was found to return the most popular tags, while the SML was found to be extremely feature-dependent, and could not be integrated with standardised MPEG-7 Features. Thus, the high reported performance measures could be artefacts of the collections and not due to the descriptive power of the models. Finally, we have demonstrated that a simple SVM approach performs better than state-of-the-art models across several collections and descriptors.

We argue that as the number of experimental settings increases and as we keep their diversity high, we get more insight on a model's functionality, while strong and weak points emerge. As such, this study sets forward an evaluation paradigm for future annotation models, while the proposed framework should be integrated in the whole process of the development of a model, from the conceptualization and the development phases until the validation and evaluation.

## 5 Performance on the Alan Online Image Collection

While the scope of the content-based retrieval system designed and evaluated for the SALERO project is general, as part of the project there has been a need to design and evaluate a system for a specific system and task: that of retrieving images in the database used in TAIK's Alan online art installation. In this chapter the problem will be described, and the design and evaluation of a new TAIK retrieval algorithm given. It should be noted that this work is a response to the requests of the reviews from the IBC review, Amsterdam, 2009.

While it would appear that this is a straightforward problem, it will be seen that it is, in fact, much more complex than it at first appears, which would ultimately require a semantic analysis of the images drawn to retrieve the correct target images. Despite this, on a specially created set of hand drawn queries, the latest system is able to retrieve the correct image 82% of the time, rising to 96% if the search requirements are relaxed to allow the correct result to appear in the top 5 rank positions.

### 5.1 The Retrieval Problem

Image retrieval plays a central role in the multimedia art production of TAIK, called Alan online<sup>7</sup>, an online counterpart to the physical installation called Alan01. This system engages interactive audience in dialogue with a fictional Alan, as if Turing's consciousness had been coded into a machine at the time of his death. The interactor can "talk" to Alan online by a system of symbols, which we imagine to have been relevant to Alan Turing's life. The systems start with a way of allowing the user to draw an image, which is then passed to an image retrieval system, which returns one or more matching symbols from a hand-built collection. The search result or results then triggers the associational story in each of the productions.

The retrieval is made from a limited set of approximately 50 symbolic images, the graphical presentation and selection of which having been made while bearing in mind the nature of the interface. The symbols are similar to what a user might draw in a short time of five to ten seconds. Another requirement for the selection of the symbols is that they are connected to the context of Alan Turing's life. Some examples of the images: bird, hand, fish and heart are shown in Figure 2.

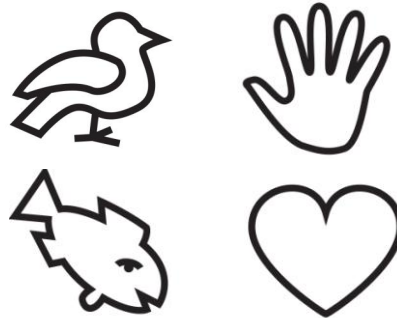


Figure 2: Examples of the symbolic images shared by the installation and its online version

The idea of the drawing interface is therefore to enable a non-textual input to an art piece, which can still be translated to symbols and their textual meanings. From there on, the associational narrative structure script of the art piece can start to function.

The first point to be made is that the visual retrieval system is only required to generate a single result: a single result is used by *Alan01* as the initialising point for the sequence of other actions; in the *Alan online* production, it is possible to display multiple results, to provide multiple responses to the input. This requirement calls for high precision results, with the added requirement that the system must always generate a result – even if there is nothing in the collection which is similar to the input query, the *closest* should always, ideally, be selected.

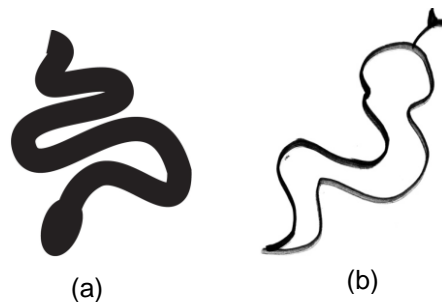
Secondly, the matching required by the productions is largely visual in nature – each query input, whether generated by the touch screen in *Alan01* or the online drawing canvas of *Alan online*, is only

<sup>7</sup> Installation website: <http://mlab.taik.fi/alanonline/index.html>

required to be visually similar to the result image in some manner. But this visual similarity should be *understandable*, i.e. users should be able to intuitively see why a result was produced for a query, or be able to determine a property of the search result which matches the input query. If the system cannot return results which cannot be interpreted as similar, the risk is that those interacting with the productions will be less likely to engage with them.

As already mentioned, for the specific needs of the installation a limited set of roughly 50 images were created (Figure 1). The retrieval system is expected to retrieve the most semantically similar image from this small collection of 50 images, but for an arbitrary hand drawing generated by *Alan01/ Alan online* via their respective sketch interfaces.

Since we have a very small collection of shapes and the queries are expected to be generated by *Alan01* and *Alan online* through an interface where a user can draw any arbitrary image, the similarity retrieval of an object from the collection becomes a big challenge. For instance, the image in the database for the object snake shown in Figure 3(a) is very different from the user drawn snake shown in Figure 3(b) and yet the system should ideally retrieve the snake shown in Figure 3(a).



**Figure 3: Examples of collection and query images**

The problem of recognising arbitrary shaped images irrespective of various geometric transformations such as, orientation, scaling, translation and shearing effects has been tackled in the fields of Robotics, object recognition and computer vision, among others. Object recognition with shape features deals with finding a match between certain features obtained from the shape of a query submitted with that of the different instances of the model objects in the database.

## 5.2 Requirements

---

The main requirements were:

- *Only a single result*: The image retrieval system was required to generate a single result, which is then used as to initialising a sequence of other actions. In the final AlanOnline production, due to performance issues, the final interface showed the top three results (Figure 1), although this was not considered ideal.
- *Visual matching (query-by-example)*: The matching required by the production is visual – each query input generated by the drawing canvas, is required to be visually similar to the result image in some manner. But this visual similarity should be understandable, i.e. users should be able to intuitively see why a result was produced, or be able to determine a property of the result which matches the query. If the system cannot return results which cannot be interpreted as similar, the risk is that those interacting with the productions will be less likely to engage with the production.
- *Small search collection*: The size of the collection which was to be searched was small, consisting of roughly 50 specially created images. The retrieval system is expected to retrieve the most semantically similar image from this small collection.

## 5.3 Initial Approach

---

The initial approach taken to this problem is given in document D5.5.4 (Chapter 4) and in the paper [Tuomola 2009]. Because the database to be searched are composed of black and white images colour, an important feature in content-based image retrieval could not be used. Instead, six other feature descriptors were created and investigated, with this collection:

- Edge Histogram descriptor: local and global edge histograms from the images
- Contour Shape Descriptor: identifies the “objects” in an image, and then extracts the longest contour from that object
- Block ratio: a feature which splits the image into blocks, and records the ratio of foreground and background pixels in each block
- Centroid Profile (“signature”): Computes the centre point of an object and then encodes the shape as distances from this centre point
- Object Signature with reference to Axis of least inertia: computes the block ratio and centroid profile relative to a reference axis for an object (axis of least inertia)
- Horizontal and Vertical projection profile: the object points projected to the x and y axis along the axis of least inertia of an object

A range of different fusion methods were also created: reciprocal rank, Borda count, Condorcet method, Weighted voting, etc. All of these techniques and features are described in D5.5.4.

In order to evaluate the different possible retrieval techniques which could be used, a small testing setup was created, composed of the target set of images, a set of hand-drawn query images, and a set of relevance judgements. The set of relevance judgements was manually created, where each query image was matched to its ideal target image (an example is given in Figure 4). In addition to an ideal “target”, we also defined for each query zero or more alternative images which were deemed to be acceptable results for the query, but not ideal (Figure 5). This latter list was defined to enable us to consider the matching between query and target(s) as a fuzzy mapping, in order to model what was thought to be “acceptable” in the implementation of *Alan01* and *AlanOnline*, where the output generated need not always be exact. Indeed, the collection may not necessarily contain any images similar to the query.



Hand drawn query



Target image from

**Figure 5: Example query image and associated target image****Figure 4: Alternative matches for the query in Figure 5**

Based on a set of hand drawn images as queries, it was found that the best combination of features and fusion techniques could retrieve the correct image 32% of the time; if we relax the retrieval to also allow the system to retrieve a set of “acceptable” target images, rather than the single correct image, the best technique was correct 47% of the time on the test set.

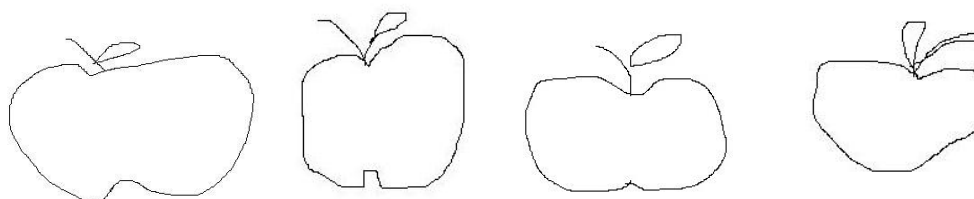
## 5.4 Second Approach

After the performance problems with the former approaches, and based on the feedback from the review meeting at IBC in Amsterdam, a different approach was tried which used expanded the TAIK produced set of images with an auxiliary set of hand drawn examples. Instead of attempting to match the user’s drawing with each individual TAIK image, the system matches the input against the expanded set of images.

For example, consider the problem of matching the apple image in the TAIK collection:



The first step in the process is to produce a range of human generated drawings of this image. This was carried out via an email campaign at Glasgow University, where ten users volunteered to manually draw each of the TAIK images. The drawing was carried out using a standard mouse, as would normally be used to draw on the Alan Online canvas. This resulted in a range of hand drawn images of various qualities, four of which are shown below for the apple shape.



As can be seen, these vary in shape, size and quality, as can be expected when non-artists draw apples using a computer mouse. Additionally, the width of all lines is one pixel wide, which also corresponds with the images produced by the Alan online system (although it should be noted the images displayed on the interfaces use a “pen” which is broader than a single pixel width, the final images sent to the retrieval system are composed of single pixel lines).

Our target “apple” image can now be represented in terms of the other hand drawn images: if the user’s input drawing matches one of the hand drawn images, then we achieve a match with the corresponding TAIK image.

Once we had collected together this expanded set of images, each one was indexed using three of the features as described in the previous section:

- Projection profile
- Edge histogram
- Homogeneous texture

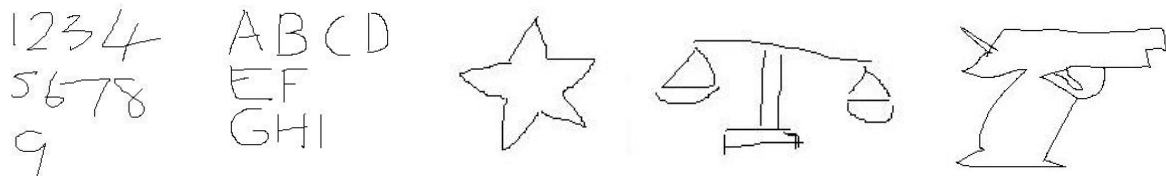
These three features performed best in our retrieval experiments. The same indexing structure was used as before, using the AspectBrowser search backend, but the method of retrieval was changed to better reflect the task at hand. Given the query image, for each of the three above features, the top 5 results were retrieved from the index, and pooled into a single set of 15 results. The number of instances of each symbol was then counted, and the symbol with the highest instance count is returned. If there is a tie, results from the projection profile feature are preferred over the results from the Edge histogram feature, which are preferred over Homogeneous Texture, reflecting the respective performance of the three different types of feature.

For example, consider the following query:



This query results in the following results for the three features:

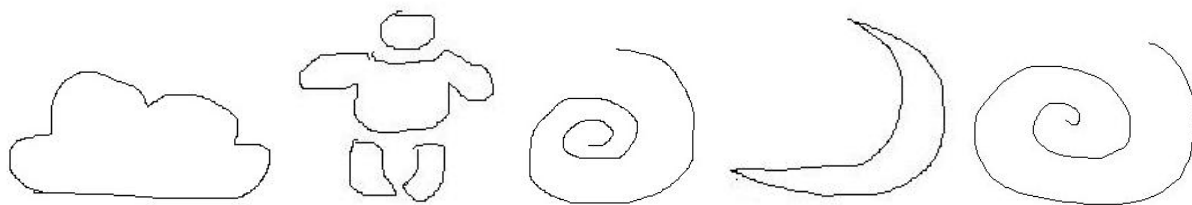
**Edge Histogram:**



**Projection Profile:**



**Homogenous Texture:**



It can be seen that in this case the projection profile feature is working much better than the other two features, although this can change depending on the query and the target image. Once these result lists have been retrieved, the next stage involves merging and counting the number of instances of each symbol. For the above query we have the following counts:

Symbol	Count
Letters	6
Spiral	2
Numbers	1
Star	1
Scale	1
Gun	1
Cloud	1
Child	1
Moon	1

The result of this search is therefore the “letters” symbol, which is correct in this case. This functionality was rolled into the backend retrieval system, which was reprogrammed to carry out this type of search on the TAIK set of images.

**5.5 Evaluation**

Similar to the previous evaluation, a new evaluation was carried out to investigate the performance of this new technique. A similar evaluation method was used as previously, with the same set of query images and collection. In this case, we evaluated the retrieval by considering the percentage of correct answers for each feature, and the combinations of pairs of feature, then the percentage of the queries in which the correct answer appears in the top five results, again for the three different features and combinations of features. Finally we also evaluated the search and fusion technique described in the last section, where top 5 results from each feature are fused and counted.

Since we have a range of different hand drawn queries generated by 10 different users, the evaluation was executed ten times, where in each case the hand drawn images created by one users were removed from the collection being searched, and instead being used as queries. E.g. the sketches drawn by user one were matched against users 2 to 10, and this was then repeated for each of the other users. The results are shown below, in Table 13.

Technique	Top 1	Top 5
Edge Histogram (EH)	36%	62%
Projection profile (PP)	47%	56%
Homogenous Texture (HT)	27%	68%
PP + EH	60%	80%
PP + HT	58%	78%
EH + HT	50%	76%
Top 5 plus fusion (Approach described in Section 5.4)	82%	96%

**Table 13: Results from the second method developed to retrieve images from the Alan online collection**

As can be seen in the table, the performance of the system has increased when compared to the formerly used methods: Projection profile (PP) when combined with Edge Histogram (EH) can, for the set of hand drawn images used for testing, return the correct image 60% of the time. This rises to 80% if we relax the evaluation to allow success to be if the target image appears in the top 5 results returned by the engine. Looking at the final row in the table, the fusion technique described above works particularly well, generating the correct result 82% of the time, or 96% when considering a correct result ranked in the top 5.

Considering each user generated set of queries separately, we also present the list of individual results for each user:

User	Top 1	Top 5
1	85%	91%
2	96%	100%
3	64%	85%
4	91%	100%
5	89%	97%
6	77%	98%
7	87%	98%
8	80%	98%
9	66%	94%
10	83%	100%

**Table 14: The percentage of correct results, for each individual user who generated hand-drawn images for the TAIK collection. Each user was then matched against all other drawings to generate the results.**

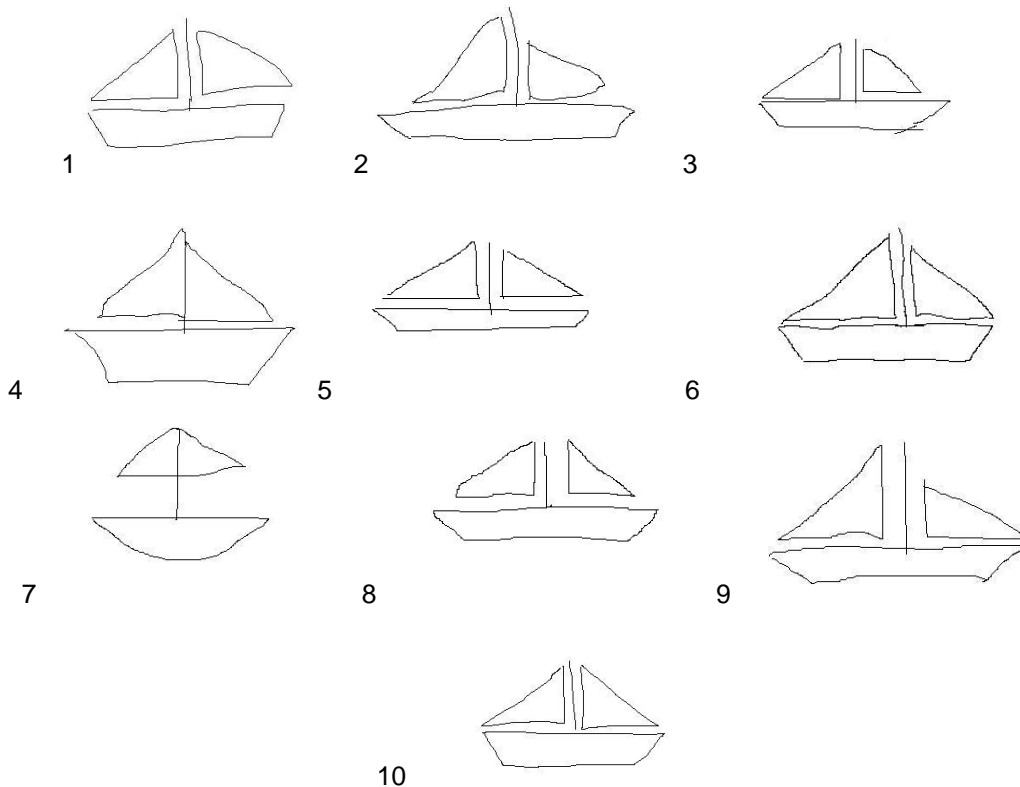
As can be seen, some user queries perform better than others, and probably reflect the visual similarity between the sketches created by different users. Some users can draw better than others, and are likely to render images with a greater visual similarity to other well-drawn images. Poorly drawn images will, on the other hand, be harder to match well. However, it should be noted that the greater the number and variety of the hand drawn images generated and indexed in the collection, the greater the chances of there being a correct match with poorly drawn images – i.e. increasing the number of hand drawn

images from the ten current examples may help generalise the matching to account for a greater range of drawing styles and qualities, where the aim is to match more poorly drawn sketches.

For example, the boat image below:

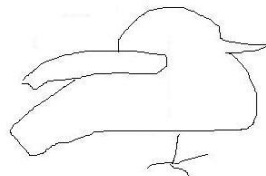


Was rendered by the ten users in the following way:



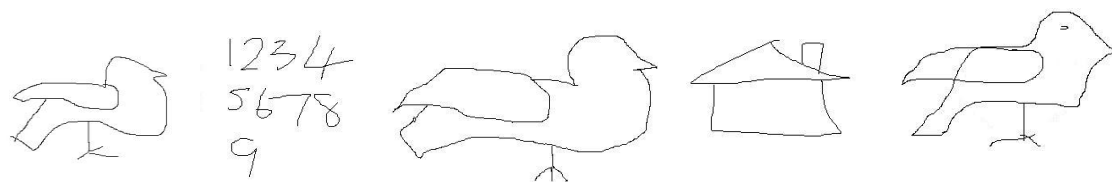
In this case, it can be seen that user 7 has produced a drawing which is markedly different from the others, and is therefore less likely to produce the correct answer.

While Projection profile is the feature which performs best overall, the improvement when using all three features suggests that there is a considerable benefit to using the implemented fusion technique, indeed, the three features may perform differently for the different query images. For example, the query:



Generated the following three ranked lists from the three features

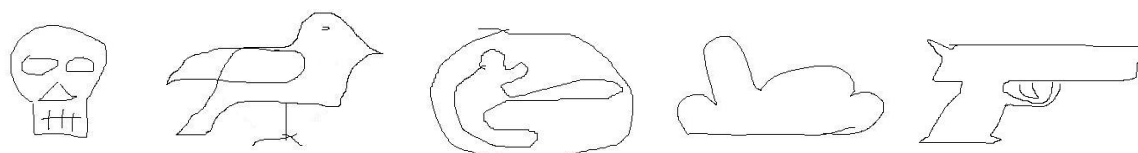
**Edge Histogram:**



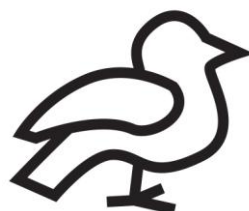
**Projection Profile:**



**Homogenous Texture:**



For this particular query, the Edge Histogram feature, which is on average the poorest, performed the best. Using the fusion presented in Section 5.4, the correct answer is again produced, i.e. the bird symbol:



## 5.6 Issues

In process of developing and evaluating the retrieval system, a number of issues arose. The AlanOnline production presented a very demanding challenge to an image retrieval system, where the retrieval is required to be consistent with the human visual perceptive system, and which presents high precision results.

In particular, the following issues will be a problematical for any system which aims solve this problem:

- (1) **High-precision:** One of the most difficult requirements for the search system was the limited number of search results required. In a conventional image retrieval system, if the user is presented with the top ten results, it is typically considered sufficient if a significant ratio of those results are relevant – the fact that the highest ranked result isn't relevant doesn't render the list unusable, which is the case here.
- (2) **Visual similarity is not just visual – semantics always gets in the way:** While the intention was for the matching of the query to the collection to be purely visual, a problem was that as viewers and judges of the quality of the search results, we do not naturally view objects while ignoring their semantics. For example, if one knows that the collection contains an apple, and you draw an apple, the expectation is that an apple is returned – even if there is another image which may be “visually” more similar in terms of its shape. The knowledge that an image is an “apple” or “computer” can override particular visual similarities which may be present between two images. This problem – which can be considered an example of the “semantic gap”, i.e. the gap between visual perceptions and our high-level semantics, is a current ongoing research problem.

- (3) **What to do when there's more than one possibility?** A further problem is when multiple, possibly similar images may all be matches for a query. E.g. given the query below (on the far left), we would like to retrieve the expected "human" image from the collection (centre image). However, the "devil" image is also very similar, and could appear equally likely. The reliable retrieval of one rather than the other is a difficult problem, complicated further when we consider how different the query must be to differentiate the two – e.g. if the user draws a human with a tail, should the "human" be retrieved, or the "devil"? Obviously, such considerations also involve a semantic interpretation that visual approaches lack.



- (4) **Predictability:** In AlanOnline, one of the central issues has become the predictability of the retrieval results. In an interface that uses this technology, the user is tempted to start to test the system or even play against it. Seeing the results which the system has delivered previously affects the imagery that a user starts to draw henceforth. If a user tries to replicate the images he/she has seen in previously, the following results need to be consistent to avoid the feeling of randomness, and to ensure that the productions illusion is not broken. A problem with current retrieval techniques used is that small changes in the query can result in large differences in the results, which makes predicting the result of a query difficult.
- (5) **What to do when there's nothing similar?** A further problem occurs when the query is unlike anything in the collection. The question is then: what does the retrieval engine return? Since a similarity measure will always return a score when two images are compared, we can always return the most "similar", even when the most similar is very dissimilar to the query. I.e. there is no "lower bound" on how dissimilar the query can be from the top result to be returned. But returning dissimilar images is likely to make the retrieval system appear more random to a user, rather than less, yet the retrieval system must return *something*, unless error conditions are allowed to be generated by the search.

## 5.7 Summary

In this chapter we have outlined the work which was carried out to create a retrieval system which would produce acceptable results for the Alan online experimental production, from TAIK. A test collection was created, tailored to the problem domain, and an initial set of techniques were developed which were able to return the correct image up to 42% of the time, for the test queries.

Further development, based on the comments from the SALERO reviews in the IBC review meeting in Amsterdam, was then undertaken to improve this performance. The second version of the system was able to return the correct result up to 82% of the time, for the test set, or 96% of the time when considering ranked lists 5 images long. This represents a considerable increase in performance over the initial system. The changes detailed in this chapter have been rolled into the search system, and are currently in operation on the Alan online art installation.

## 6 Image Search via an Intermediary

---

In developing and evaluating the AspectBrowser interface, a useful method of searching image databases which have no associated text was found: visual examples could be first found from a database which did use text, and then these examples could be used to search the target database for the required images. The ability of the interface to support this style of interaction led us to investigate whether this approach – using an “intermediary” image database in order to search a target database – could in its own right, be of use in the search. This led to the work reported in this chapter, where we specifically consider the problem of search image databases which have no associated text or other textual annotations – a common problem in the SALERO project.

### 6.1 Introduction: The Problem

---

Developing methods for searching image databases is a challenging and ongoing area of research. One common way of searching image databases is via text annotations or tags which have been manually attached to images. This is the standard method of search in online image and video systems such as Flickr<sup>8</sup> and YouTube<sup>9</sup>. Metadata can also be supplied manually by librarians or other human classifiers; alternatively in an effort to make the tedious task of providing annotations less of a chore for individuals.

When the images are part of a larger textual document, the text around an image (or in the case of the web, in the anchor tag) can also be used to characterise the image. Contemporary examples of this approach include the image search functionality of both Google<sup>10</sup> and Microsoft Bing<sup>11</sup>.

While searching image databases using annotations and other manually generated metadata such as tags has proved successful, it suffers from the problem that people must manually annotate the data. In the case of online repositories of images, where there are many users and many uploaders of data, annotation burden may be spread between many users. In the case of commercial image providers who have a commercial impetus to making their image collections easily searchable, the cost of manual annotation may make economic sense, especially when the number of users of the image collections is large<sup>12</sup>.

Manual annotation in this way can, however, take a considerable length of time and money, and this cost is not acceptable in many situations. There are many situations, however, where the effort or cost of manual annotations cannot be justified, such as when the collection of images is internal to an organisation, and used by relatively few users: a good example of this is the image asset collection maintained by PGP. Additionally, making the image collection public is often out of the question for commercial copyrighted images. Such private image collections may be made up of many thousands of images, and may also be temporally limited – for example, only used during the period of a particular animation of game. This temporal aspect can also limit the resources which can be spent on annotations.

---

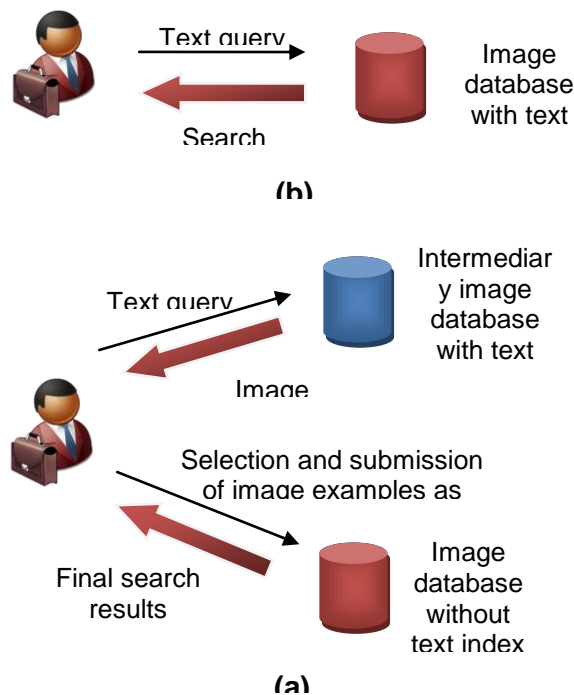
<sup>8</sup> <http://www.flickr.com/>

<sup>9</sup> <http://www.youtube.com/>

<sup>10</sup> <http://images.google.com/>

<sup>11</sup> <http://www.bing.com/images>

<sup>12</sup> For example, Getty Images ([www.gettyimages.com](http://www.gettyimages.com))



**Figure 6: (a) the target database to be searched has no associated text, therefore a user first searches an intermediary database for image examples, which he/she then uses as search queries to the original target database; (b) a user is able to directly search a multimedia database using a text query**

One alternative to metadata search is content-based image search [Smeulders 2000], i.e. systems which process the visual image data itself in order to extract features which can then be searched. Examples of current systems include QBIC [Flickner 1997] and EGO [Urban 2006b]. Such systems typically work with the input of an image example to initiate the search. If no pre-existing example image is available, random images from the collection may be presented to the user, or a sketch interface may be used. In query by sketch users are presented with a simplified paint-like interface in which to visually draw their information need [Tait 2001]. These techniques can be difficult to use – either users must be able to draw their information need, or already have available examples of the need, either of which is not always possible.

A further possibility is the use of automatic annotation, to automatically annotate images with high level concepts, as discussed in Chapters 4 and 5 of this report. As previously stated, the performance of such systems is unfortunately still low, and the number of concepts identified small. For example, LSCOM-lite used in TRECVID 2005 contained around 50 concepts, while only 20 concepts were used for test purposes in TRECVID 2008.

### 6.1.1 Searching via an Intermediary Collection

In this chapter, we consider an alternative approach where by a user is allowed to search for images through an intermediate database. In this approach, a user can search using text in the intermediate database as a way of finding visual examples of their information need. The visual examples can then be used to search a database that lacks annotations. This is illustrated in Figure 6.

A user first composes a text query and executes this query on the intermediary image database. This search is executed by text, using text annotations on a database of images with annotations. The results of this text search will be a list of images which the user can browse in order to find good examples of their information need. In the second stage, the user then submits one or more of these examples to the second “target image database”, in which the images do not have any textual annotations. This second search can be carried out using content based image retrieval methods, similar to those used in current query by example systems. In the rest of this chapter we will refer to *intermediary* and *target* when referring to the database with textual annotations, and the database without textual annotations, respectively.

Searching the target image database therefore becomes a two stage process, with the images of the intermediary database being used as proxies for the searching of the target database. When this

process is compared to directly searching an image database by text, as illustrated in Figure 6b, it may be considered as inferior – but this is only because of the lack of textual annotations.

It should also be noted that we assume that source of the image is important, i.e. the user is required to use an image from the target image collection, and only those images are acceptable results. This is not unusual in commercial environments, where the copyright and source of an image is important for it to be legally usable by the organisation. Journalists, for example, cannot use any image downloadable from the internet, but must use images which can be legally reproduced, and which are of a quality acceptable to a print publication.

Additionally, the images used from the intermediary collection do not themselves have to be relevant to the users information need *per se*, but rather represent visually an attribute of the relevant images required in the target collection. For example, if the user is looking for a red sunset, a visual search with a red sun rise may produce relevant results, due to the visual similarity of sunsets and sunrises.

Further, it may be thought that search tasks which are easier to visually represent as an example image are likely to be more tractable using such a scheme, although this is likely true for all image based searching. For example, the information need “find an image of a red Ferrari Daytona” is, by its very nature, visual: the red car has a relatively consistent visual form which can be identified by the user and submitted as a query to the target collection. More semantic search tasks such as “find romantic scenes from Hollywood movies” may work less well, given the greater range of very different visual content which may be used to represent the information need.

The question remains – how inferior, if at all, is using an intermediary when compared to searching a collection directly by text annotations? If by carrying out such an intermediary process a user is able to perform with an acceptable performance, then this may provide a useful method of searching image databases where the target collection cannot be searched by text.

For the purpose of the evaluations reported in this chapter, we are principally concerned with the initial stage of searching – i.e. the creation of the initial visual query and the finding of an initial set of relevant image results. Once one or more relevant images have been found in the target collection, relevance feedback techniques can be used to find other similar images using image retrieval. But making this initial first step – the finding of initial target images which are relevant – is, as described previously, a major barrier to searching un-annotated image collections.

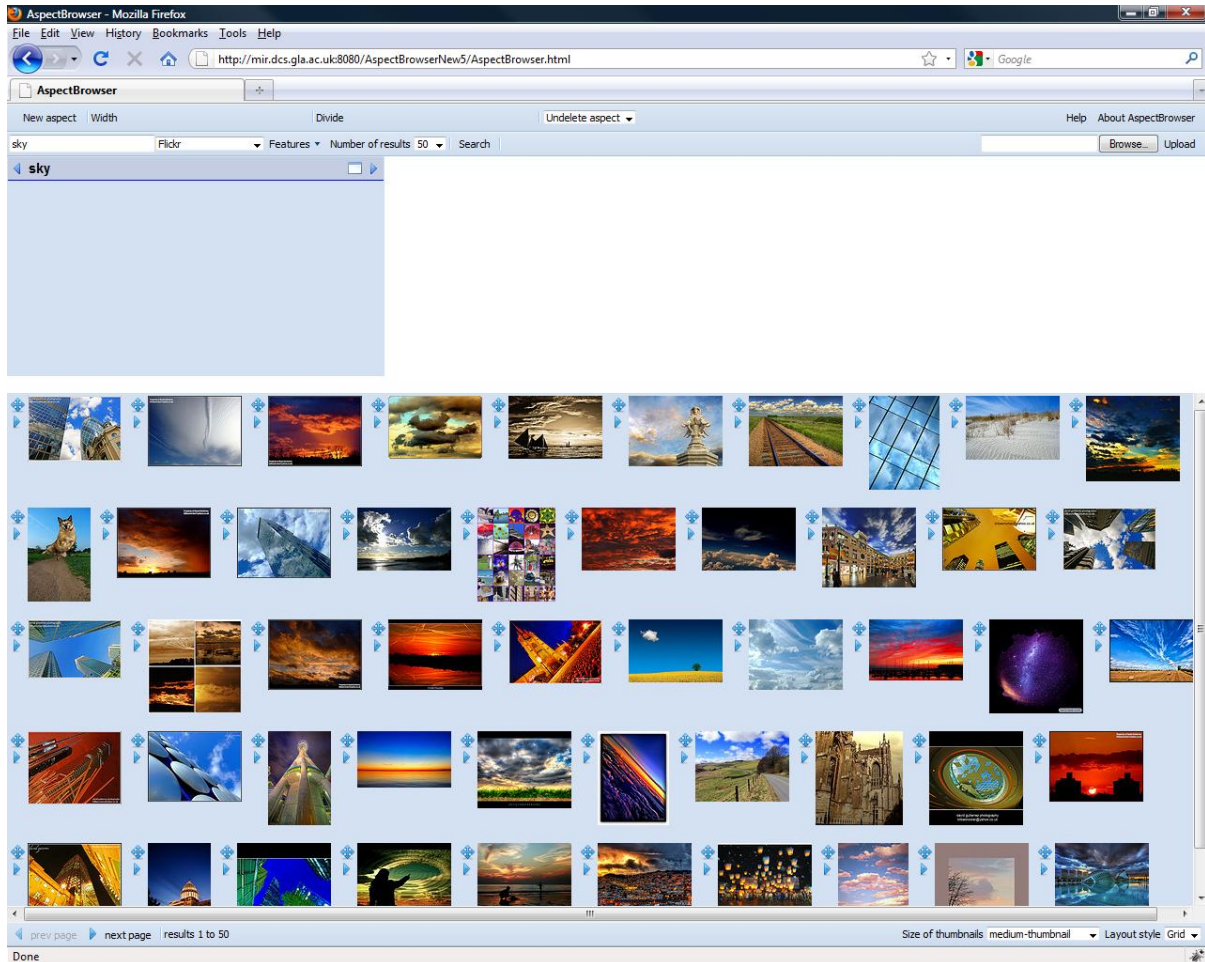
To investigate this problem, a set of offline evaluations was carried out: first an automatic study was carried out, which automatically used the image results from an intermediate database as queries for the target database; secondly, images manually found and selected by users from the intermediary database were used to find material in the target database; lastly, a user study was carried out to investigate the how user interaction with both sets of databases can influence the results produced. For this latter experiment, we present an interface which enables searching a target database via an intermediate, and compare this interface to users when searching a collection directly with text.

## **6.2 Searching via an Intermediary using the AspectBrowser Interface**

While not explicitly designed to allow searching via an intermediary database, the AspectBrowser interface (described in document D5.5.4) does, in fact, allow the easy use of an intermediary thanks to its concept of search aspects which exist within the context of an overall sequence of aspects. In particular, the features which support intermediary searching include:

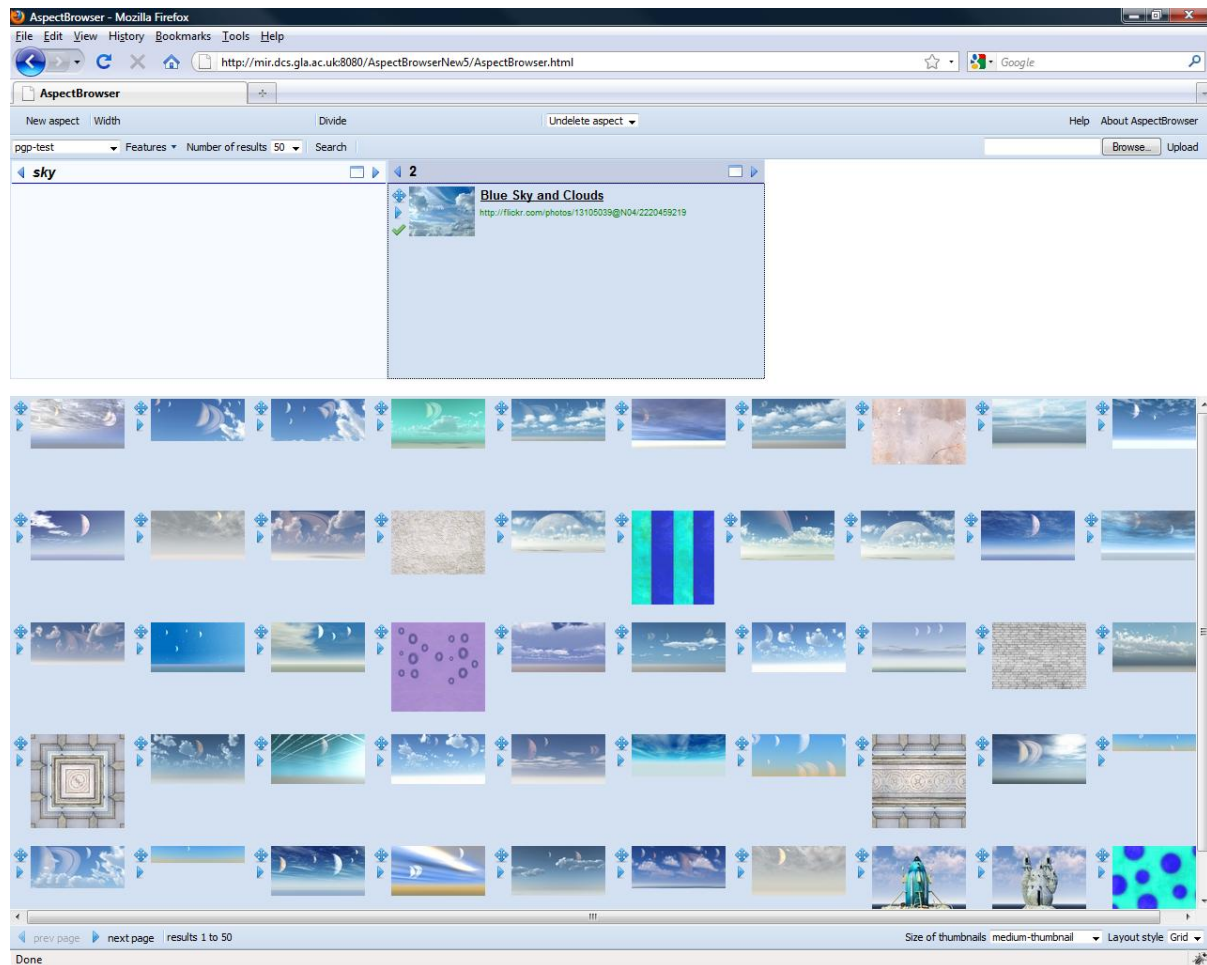
- The ability to search more than one collection
- The ability to use results from one search as a query to a new search, where the result and query may exist in different collections
- The ability to creation “aspects” which allow multiple queries to be handled at the same time

This will be illustrated with an example. First, shown below, a user can search the “intermediary” image collection by text: in this case the user executes the query “sky” on Flickr. The results are then displayed (Figure 7, below).



**Figure 7: The AspectBrowser interface after executing the query "sky" on the external collection "Flickr"**

After perusing these results, the user can then select one or more good examples of "skies" for use in the next step of the process. In this case, a new aspect will be created. A single Flickr result called "Blue Sky and Clouds" is then dragged and dropped onto the tabbed area of the new aspect. This adds underlying Flickr image to the aspect, and allows that image to be used as a content-based query. The target collection is then selected, which in this case is a database of Bing and Bong image assets supplied by PGP. By pressing search a content based query will then be executed on this target collection of images, which may be indexed by low-level image feature only, such as is the case here. The results for this content-based search are shown below (Figure 8). As can be seen, not all images returned are relevant – but the majority are, and there is the potential for further improvements in the results by using further image examples from the target (PGP) collection.



**Figure 8: The AspectBrowser interface after executing a content-based search on a collection composed of PGP sourced asset images**

To investigate this approach further, a set of evaluations was carried out, first using automatic techniques, and then a user study was designed to evaluate the potential of this technique on a custom-designed interface.

## 6.3 Offline Evaluations of the Technique

### 6.3.1 COLLECTIONS and SYSTEMS

For the experiments reported in this chapter, two collections are required, a target image collection, and an intermediary image collection. For the target collection, the CLEF 2007 image collection was chosen [Grubinger 2007, 2007b], due to its standardised nature including topics and relevance judgements. CLEF 2007 is a set of 20,000 images, 60 search topics, and associated relevance judgements. As part of the CLEF 2006 effort, which shared the same set of topics as used in CLEF 2007, the topics were categorised into a number of different categories, including: easy/hard, semantic/visual, and geographic/general [Grubinger 2007]. Four example CLEF topics are shown in Figure 9 (these four topics were used in a user study, described in a later Section)

The image search system used to search the CLEF 2007 collection is the SALERO backend system. In this Chapter, five different visual features are used to represent each image:

- Colour Layout: special distribution of colours in an image
- Colour Histogram: colour distribution of the image
- Edge Histogram: the spacial distribution of edges in the image
- Homogeneous Texture: region texture representation
- Colour Structure: represents an image by a combination of colour distribution and the local spacial structure of the colour



Figure 9: Four example CLEF 2007 topics

For the intermediate collection, Flickr<sup>13</sup> was chosen. Flickr provides us with a very large, broad collection of images which can be searched by text (e.g. by image title, description, and supplied tags). The size and broadness of the images available makes this a good collection for finding examples of many different kinds of topic and means there is little need to ensure that examples of the CLEF topics are available. Additionally, Flickr has the advantage of being a well known repository of images known by users, and has an open API which allows searching by external systems such as our intermediary interface. The Flickr “flickr.photos.search” API was used<sup>14</sup>, free text Flickr queries being executed against titles, descriptions and tags, producing results similar to those produced when searching from Flickr’s own web site.

### 6.3.2 Automatic Retrieval via an Intermediary

The aim of the first experiment was to investigate the performance when using an intermediary automatically, i.e. when the image results returned from a search to Flickr are used directly to search the CLEF collection. In this case, the title text from each of the 60 CLEF 2007 topics were used to search Flickr, and the top 40 returned image results were downloaded. The resulting text queries are relatively short, averaging 4.7 terms per query (SD 1.34). Four examples of the queries are shown in Figure 9. The top  $n$  images were then used as a content based query to the CLEF collection, where  $n$  was set to 1, 5, 10, 20 and 40. The returned results were then evaluated, using Mean average precision (MAP) and precision at 20 (P20).

<sup>13</sup> [www.flickr.com](http://www.flickr.com)

<sup>14</sup> [www.flickr.com/services/api/flickr.photos.search.html](http://www.flickr.com/services/api/flickr.photos.search.html)

	MAP	P20
Text	0.1365 (0.1827)	0.1608 (0.2227)
Intermediary results at various Flickr ranking depths		
Top 1	0.0027 (0.0129)	0.0083 (0.0321)
Top 5	0.0015 (0.0046)	0.0033 (0.0126)
Top 10	0.0015 (0.0046)	0.0025 (0.011)
Top 20	0.0015 (0.0046)	0.0025 (0.011)
Top 40	0.0013 (0.0045)	0.0025 (0.011)

**Table 15: MAP and P20 results for a baseline text run, plus automatic intermediary results at different Flickr ranking depths; Mean (SD)**

For comparison purposes, a baseline run was also carried out, which used the text of the CLEF topics to directly search the image annotations in the CLEF collection. This run used the Lemur<sup>15</sup> information retrieval system, using the BM25 ranking model, Porter stemmer, and the standard stop word list. It should be noted that this run is not state of the art – for comparison, the best CLEF 2007 automatic text runs had MAP 0.2075 and P20 0.2533, however it does represent a standard and generalisable approach which can easily be applied to any collection.

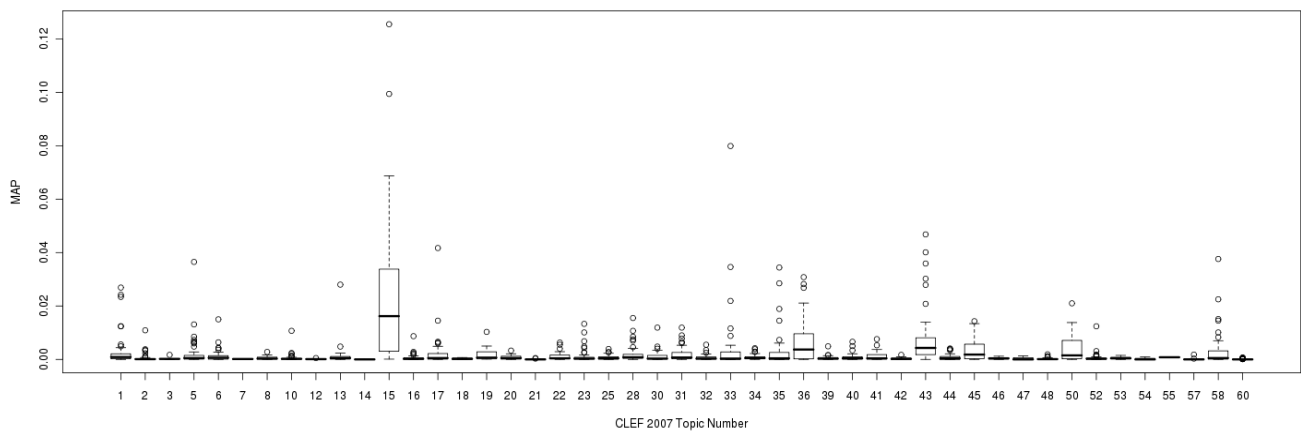
Results of this first evaluation are shown in Table 15. Out of the 60 topics, 13 topics did not return any images from Flickr; over the 60 topics the mean number of image results was 31, with Standard Deviation 14. It can be seen that the performance when using an intermediary is lower than the equivalent text, with the intermediary performing at only 1 to 2% of text as measured by MAP (2 to 5% measured by P20).

Using the intermediary, the best performance was achieved by taking the top ranking Flickr image result, but at the cost of a large standard deviation (SD) between the different topics. I.e., for some topics, using the top ranked Flickr result may return some relevant material, while for others, no relevant material is returned. When using more of the ranked list, both MAP and P20 values are reduced, as is the SD – i.e. the average performance across all 60 topics decreases, but for any individual topic there is potentially a better chance of relevant material being found. The relatively large SD (when compared to performance) suggests that there are images which perform relatively well within the ranked lists returned in the searches.

To investigate this further, the performance achieved by each individual Flickr result was evaluated, to check the maximum performance possible on the selection of a single Flickr image. Boxplots for each of the 60 topics are shown in Figure 10, showing the performance range for each query. The best performing Flickr result was for Topic 15, with MAP 0.1255; 11 topics had a maximum performance greater than 0.02, suggesting that while the average performance across all Flickr images is poor, there are images that perform better than the average.

While still poorer than the equivalent text scores, this does suggest that with accurate image selection, performance can be increased. A second experiment was therefore carried out which, instead of automatically using the Flickr ranking, asked users to select relevant images manually, to investigate whether user selection of images as queries improves the performance of the intermediary approach.

<sup>15</sup> <http://www.lemurproject.org/>



**Figure 10: Performance of Flickr results (MAP) for the different CLEF topics, illustrating topic differences**

Topic	Type	Text		CLEF topic images		User-selected Flickr images	
		MAP	P20	MAP	P20	MAP	P20
22	Vis/ Easy	0.0037	0.0000	0.1671 (0.0973)	0.5833 (0.2754)	0.0019 (0.0028)	0.0000 (0.0000)
53	Vis/ Med	0.1396	0.1500	0.0341 (0.0514)	0.1000 (0.1000)	0.0004 (0.0006)	0.0000 (0.0000)
14	Vis/ Diff	0.0034	0.0500	0.0790 (0.0924)	0.1667 (0.1756)	0.0011 (0.0020)	0.0042 (0.0144)
17	Med/ Easy	0.3671	0.5000	0.0352 (0.0432)	0.0500 (0.0500)	0.0024 (0.0043)	0.0125 (0.0311)
34	Med/ Med	0.0753	0.0500	0.0103 (0.0059)	0.0500 (0.0500)	0.0017 (0.0025)	0.0167 (0.0258)
49	Med/ Diff	0.0000	0.0000	0.0649 (0.0378)	0.3333 (0.1258)	0.0012 (0.0014)	0.0154 (0.0315)
39	Med/ VDiff	0.0006	0.0000	0.0002 (0.0002)	0.0000 (0.0000)	0.0002 (0.0005)	0.0000 (0.0000)
57	Sem/ Easy	0.9842	0.5500	0.0633 (0.1056)	0.0333 (0.0577)	0.0008 (0.0014)	0.0000 (0.0000)
8	Sem/ Med	0.5906	1.0000	0.0044 (0.0018)	0.0500 (0.0500)	0.0012 (0.0008)	0.0100 (0.0211)
26	Semc/ Diff	0.0000	0.0000	0.0112 (0.0050)	0.0500 (0.0000)	0.0024 (0.0055)	0.0125 (0.0354)
40	Sem/ VDiff	0.0005	0.0000	0.0085 (0.0091)	0.0500 (0.0500)	0.0024 (0.0030)	0.0143 (0.0244)
All		0.1364 (0.1827)	0.1608 (0.2227)	0.0435 (0.0669)	0.1333 (0.1947)	0.0013 (0.0025)	0.0068 (0.0206)

**Table 16: Retrieval results for the user selected Flickr images, along with the text and CLEF topic image baselines, Mean (SD); Topic type is the CLEF classification of topics Visual/Medium/Semantic and Easy/Medium/Difficult/Very Difficult**

### 6.3.3 Manual Intermediary

After the results obtained from automatically using the intermediary technique, a second experiment was conducted where three users were asked to search for relevant Flickr images. Only the user selected images were then used to search the CLEF collection. The three users were members of our organisation and expert users in of information and image retrieval systems, but were not involved in the experimental study.

To keep the time required to carry out this manual search down to a reasonable level, eleven CLEF topics were selected: three visual topics (22, 53 and 14); four medium-visual topics (17, 34, 49 and 39); and four semantic topics (57, 8, 26, and 40). The eleven chosen topics were selected to cover each visual/semantic and easy/difficult CLEF category (the visual category did not have any topic judged as “Very difficult”, and therefore only three visual topics were used). For each topic, the users were asked to search Flickr for up to 5 relevant images, which resulted in up to 55 manually selected images. In practice, two of the users were not able to find five relevant images for all topics, the number of manual images found being 55, 41 and 39 images respectively, for each user.

For each user selected Flickr image, we then carried out a content based search on the CLEF collection, which was then evaluated against the CLEF relevant judgements. The results, for each of the eleven topics, are shown in Table 16, along with the topic type. Again, to provide a baseline for comparison, the text results for each topic are also shown (the same automatic method as described in the previous section was used). In addition to this text baseline, a further image baseline was carried out, using the CLEF topic examples. Each CLEF topic has three corresponding image examples, as shown in Figure 9, which were used to search for images using exactly the same image retrieval technique as used with the Flickr images. This allows us to compare the results of using external images with images from within the collection.

It can be seen from Table 16 that across the eleven topics, performance was again very low for the intermediary method. Average performance was roughly equivalent or worse than that found using the automatic technique used in the previous section – i.e. asking users to manually select relevant Flickr images did not necessarily result in good performance when those images were submitted to CLEF. The intermediary results are again lower than both baselines, roughly performing at 1% of the level of the text baseline, and 3% of the CLEF image baseline.

The difference in performance between the CLEF topic images and Flickr images, despite the latter being hand-picked by users, is likely to be due to the greater visual cohesiveness between the Flickr topic images and the images in the CLEF collection. The more general purpose Flickr images are likely to be more disparate than those in CLEF.

Based on the results from these two offline experiments, we developed an interface which allowed a user to interactively select which Flickr images to use. By allowing interaction, the aim is to investigate if users, while interacting with the two collections, can both choose good queries to be submitted to CLEF, and also make up for the relatively poor performance of the image retrieval system.

## 6.4 User Interface Evaluation using a Specialised Interface

---

The user study used the same underlying collections and retrieval system as described in earlier sections. Similar to the earlier systems, it was decided to compare intermediary search with a text “upper bound” – from the previous results, we expect users to perform less well when compared to directly searching CLEF using text, but we hypothesise that the difference between the two conditions will be narrower than in the previous experiments. However, we also expect, given the previous results, that the degree of user effort will increase, i.e. that users will have to work harder using the intermediary technique to find relevant images.

Two research questions were therefore defined: the first concerning performance: that the performance of users interactively using the intermediary would narrow the performance gap with direct text search, when compared to the automatic. To measure performance we use the standard Mean Average Precision (MAP) and precision at 20 (P20) measures. Additionally, we also look at user perceptions of their own performance after performing the task, via a post-task questionnaire. The second research question concerned effort expended: the effort required to find an image in the Intermediary will be greater than direct text search. The following measures are used to measure effort: the number of searches executed; the number of times the user has clicked to view more results; and the number of images viewed. Again, we also consider user perceptions of effort as measured in post-task questionnaires.

The use of the four CLEF topics (shown in Figure 9) allows us to consider each of the above two questions for the topic types: Visual/Easy, Visual/Difficult, Semantic/Easy and Semantic/Difficult. We expect that users are more likely to perform well with visual topics in the Intermediary condition, while the semantic topics are more likely to favour the Text condition. We use only four topics to limit the time required to carry out the experiment.

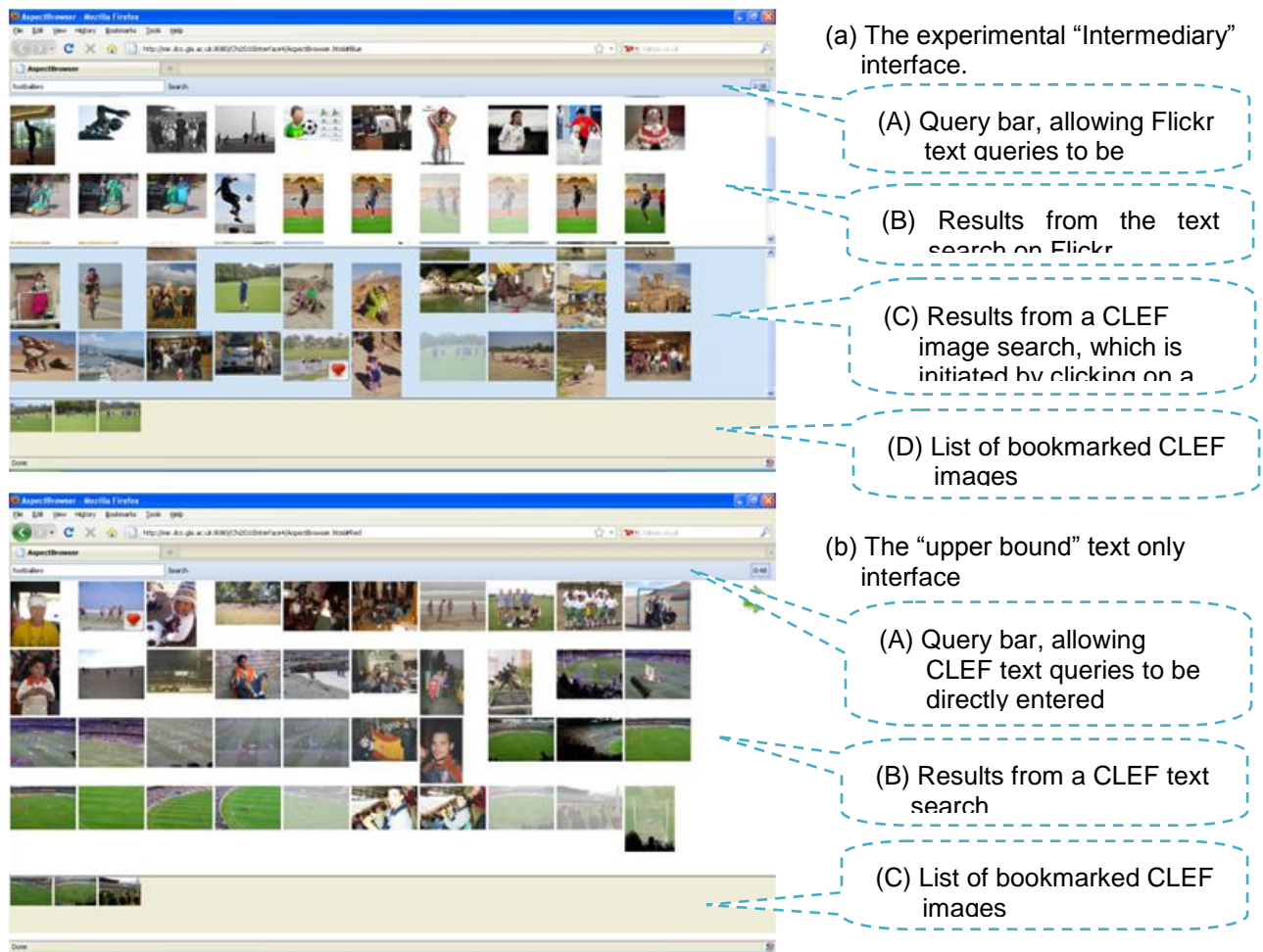


Figure 11: The two interfaces used in the evaluation, supporting the two scenarios illustrated in Figure 1

#### 6.4.1 The Search Interfaces

To support the use of an intermediary collection for image search, the search interface was required to support the following two features: the ability to search multiple different databases (Flickr and CLEF); and the ability to use a previously found search result as an example in a new search, on a different database (i.e. use a Flickr result to search CLEF)

This latter requirement is particularly important and rarely supported by image search systems directly. Ideally, this process should be flexibly supported, to allow the user to easily interact with the two sets of search results which will be manipulated. The final design used a split screen approach, where the screen was split into two, results from Flickr being displayed on top, and CLEF results below.

The intermediary interface is shown in Figure 11, and is intended to support search scenario (a) in Figure 6. It is split into four main parts, A-D. At the top of the screen is the search bar, which contains the text query box, search button, and at the far right the timer which indicates to the user how much time has elapsed for the task. Text queries entered here are executed against Flickr, the results being shown below in the area labelled B.

When the mouse is hovered over a Flickr result in area B, an icon appears on the bottom right of the thumbnail. When clicked, this Flickr image will be used as a content-based image query into the CLEF collection. The results of this content based search are shown in the lower half of the screen, labelled C. Each Flickr result therefore acts similar to a button, carrying out a CLEF search and updating the bottom half of the screen, and allowing the user a simple way of utilising the results of the first search in order to execute queries on the target collection.

When the mouse is hovered over a CLEF thumbnail in area C, a “bookmark” icon is displayed, which when pressed will add the result to the bookmarked image list (displayed at the bottom of the screen labelled D). Similarly, hovering over a bookmarked image will display a delete icon, which allows the image to be deleted from the list.

In all areas of the interface, if a thumbnail image is clicked, the image itself is displayed on screen. In the case of CLEF images, text annotations as used in CLEF-2007 are also displayed, in order to enable users to better determine the relevance of an image to a topic. However, it should be emphasised that this text is not indexed or used by the system in any way beyond its display to the user. The length of the result lists for both Flickr and CLEF searches is fixed at 40 images, to minimise the scrolling required. Next and previous buttons are provided in both parts of the interface to go forwards or backwards in both result lists.

The baseline interface (Figure 11b, supporting search scenario (b) in Figure 6) operates in a similar way to the experimental system, but is simplified, so that text searches directly search the text annotations of CLEF. Results can be viewed and bookmarked using the same buttons, and the result set size is also set to 40 images.

#### **6.4.2 Procedure**

Twenty four users were recruited through an email campaign at our organization, and were split into two groups. One group of twelve users performed the four tasks with collections which could be searched by text; the other twelve users searched the Flickr intermediary by text, and the other target CLEF collections by visual features only.

After arriving at the office where the study took place, users were welcomed before being presented with an information and consent form. After these preliminaries, an entry questionnaire was administered, before the experimenter then demonstrated the search interface, either intermediary or text. This demonstration took approximately 10 minutes, and was followed by a training task, where the user was allowed up to 10 minutes to interact and use the system.

After training, each of the four tasks were administered in a order counter-balanced for learning effects using a Latin square. Before each task, the task description was presented to the user who was then able to read it, before the evaluation was started, and the subject commenced searching. The interface contained an automatic timer, always present to the user at the top right hand corner of the screen, showing the length of time elapsed. After ten minutes, an “end task” dialog box appeared indicating the end of the task. Users were informed that they could end the task before the ten minutes if they were satisfied with their search results. This was repeated four times, one per task. At the end of the experiment, an exit questionnaire was presented to the user.

#### **6.4.3 Results**

With regards to the first research question, we wish to compare the search performance. Table 17 presents the results for MAP and P20. It can be seen that while the search performance is still poor when compared to text, it is considerably improved when compared to the automatic or manual approaches, the gap between text and the intermediary being narrowed considerably. For the two visual topics, in particular, the performance of the users using the intermediary system is approximately 40% that of the direct text system, which compares favourably when compared to the previous experiments.

The user's perception of their own task performance, as determined by a post-topic questionnaire, is shown in Table 18. Likert scales were used, where 5 = agree, and 1 = disagree. To the question “In general it was easy to find relevant images”, there was a trend for the Text interface be considered easier, especially on Topics 57 and 22 where a significant difference was found (Wilcoxon rank-sum test,  $P \leq 0.05$ ). Likewise, for question “I believe I have succeeded in my performance on the task” a similar pattern occurs, with there being a trend for users to perceive their performance as being better in Text (although a significant difference was found for Topic 57 only).

	Topic 57	Topic 26	Topic 22	Topic 14
Type	Semantic/ Easy	Semantic/ Difficult	Visual/ Easy	Visual/ Difficult
MAP				
Text	0.720 (0.082)	0.309 (0.358)	0.386 (0.313)	0.35 (0.305)
Intermediary	0.129 (0.113)	0.094 (0.071)	0.168 (0.109)	0.142 (0.121)
P20				
Text	0.396 (0.045)	0.467 (0.478)	0.792 (0.367)	0.538 (0.446)
Intermediary	0.071 (0.062)	0.371 (0.282)	0.667 (0.347)	0.242 (0.207)

**Table 17: Search performance of Text and Intermediary interfaces on the four topics, Mean (SD)**

	Topic 57	Topic 26	Topic 22	Topic 14
In general it was easy to find relevant images				
Text	<b>4 (2)</b>	4 (1.25)	<b>4.5 (1)</b>	3.5 (2.25)
Intermediary	<b>1 (2)</b>	3 (2.00)	<b>3.5 (2)</b>	3.0 (2.00)
The relevant images that I chose in the end match what I had in mind before starting the topic				
Text	5 (1.25)	4 (2.25)	5 (1)	<b>4 (3)</b>
Intermediary	3.5 (3.25)	4 (1.50)	5 (1)	<b>5 (0)</b>
I believe I have seen all possible images satisfying the topic				
Text	3 (2.5)	3 (3.25)	4 (1.25)	4 (3.25)
Intermediary	1 (3.0)	3 (3.00)	3 (2.00)	1 (3.25)
I believe I have succeeded in my performance of the task				
Text	<b>4 (1.25)</b>	4 (1.25)	5 (1.0)	5 (2.25)
Intermediary	<b>2 (1.50)</b>	4 (1.25)	4 (1.5)	3 (3.00)

**Table 18: User perceptions of their task performance, where agree = 5 and disagree = 1, and bold indicates a significant difference between Text and Intermediary; Median (IQR)**

The other two questions show less of a trend, although a significant difference was found between Text and Intermediary on Topic 14 for the question “The relevant images that I chose in the end match what I had in mind before starting the topic”. In this case users perceived that their final bookmarked images better match those they had in mind at the start of the search.

The second research question is concerned with the degree of searcher effort a user must put in to find relevant images using the interfaces. In Table 19 the results for the number of searches, number of “next” result pages requested (i.e. number of times a user clicked the “next 40 results” button to go further down the ranking), and number of images viewed are shown. Results for the “previous 40 results” button are not shown due to its very rare use (median and IQR values being zero or near zero for all topics). For all three of these measures, we split the results from the intermediary interface into two, recording actions on Flickr separately from CLEF.

	Topic57	Topic 26	Topic22	Topic 14
Number of searches				
Text	20.00 (5.50)	9.00 (8.75)	9.50 (6.25)	13.00 (8.00)
Inter (Flickr)	1.00 (2.25)	6.50 (5.50)	3.50 (3.25)	3.50 (3.00)
Inter (CLEF)	12 (10)	5.5 (2.25)	8.5 (7.5)	10 (11.25)
Next page of results requested				
Text	3 (11.5)	11 (24.5)	5 (9.5)	13 (17)
Inter (Flickr)	1 (1.25)	1 (2.5)	2 (2)	1.5 (2.5)
Inter (CLEF)	30.5(17.5)	16 (13.5)	33.5(20.5)	25.5 (15)
Number of images viewed				
Text	6.50 (19.50)	39.00 (32.50)	59.50 (64.25)	30.00 (18.00)
Inter (Flickr)	1.00 (2.75)	1.00 (2.75)	0.00 (3.25)	1.00 (3.25)
Inter (CLEF)	1.50 (4.00)	4.50 (6.50)	7.50 (18.75)	10.00 (10.75)

**Table 19: Effort measures of Text and Intermediary (Inter) on the four topics, Median (IQR).**

It can be seen immediately from the first rows of Table 19 that users carry out more text searches on CLEF in the Text condition than Flickr or CLEF searches in Intermediary. On the other hand, the use of the “next results” button is much larger in the intermediary condition, indicating that users are being forced to browse down the ranking to find relevant images. For the number of images viewed, it can be seen that users were more likely to view CLEF images than Flickr images in the Intermediary condition. In addition, it was found that users generally viewed more CLEF images in Text.

The user perceptions of task effort from the post-task questionnaire are shown in Table 20. No significant differences were found between the two systems for stress, frustration, and uncertainty in action. There is a trend for users to be more frustrated on Topic 57 in the Intermediary condition, although this is not statistically significant using a Wilcoxon rank-sum, with significance level  $P \leq 0.05$ .

	Topic57	Topic 26	Topic22	Topic14
I was stressed when carrying out the topic				
Text	1 (1.0)	1.5 (1)	1 (0.00)	1.0 (2.25)
Intermediary	2 (1.5)	1.0 (1)	1 (0.25)	1.5 (1.00)
I was frustrated when carrying out the topic				
Text	1.5(1.25)	2 (1.25)	1 (1.00)	2 (2.0)
Intermediary	3.5(2.00)	1 (1.50)	1 (1.25)	2 (1.5)
I was unsure of what action to take next when using the system				
Text	1.5(1.00)	1 (1.25)	1 (0.25)	1 (2.00)
Intermediary	1.0(0.25)	1 (1.00)	1 (0.25)	1 (0.25)

**Table 20: User perceptions of topic effort, agree = 5 and disagree = 1; median (IQR)**

#### 6.4.4 Discussion

The results presented in the previous sections show that the text interface, as is expected, outperforms the intermediary interface in all topics, although there is a large variation in the performance of users in both conditions. Comparing the performance in the interactive experiment to the previous automatic and manual methods, there is a closing of the gap between the intermediary technique and direct text search, albeit with an associated increase in the effort users must put in to reach that performance. There is a trend for the two visual topics used in the interactive experiment to have a performance closer to the text when compared to the semantic topics.

Given the fact that text retrieval does perform better than content-based search, the performance of users on the visual topics especially is encouraging – users were able to use the intermediary collection to find relevant images. Out of the 48 sessions with the intermediary interface, only on four occasions were users unable to bookmark any images in the ten minutes. In all four of these occasions, the topic being search was 57 (easy, semantic). This result reinforces the results in Table 18, user perceptions of task performance, where users found the Text system easier to find images for Topic 57. In both visual topics, all users bookmarked at least one image in each session. In comparison, on the Text condition zero images were bookmarked in two sessions (26 and 14).

Concerning effort, from Table 19 it is clear that users are forced to look deeper into result lists when compared to text. Indeed, for some CLEF 2007 topics there may be a relative lack of text annotations, the ranked lists returned by text searches often being less than the result list size of 40 used in the interface. On the other hand, the content based searching with image intermediaries would always return 40 results, until the collection itself was exhausted, due to the matching system. One surprising difference between the two conditions was the number of images viewed. In Text users viewed more images than in Intermediary, where the number of images viewed was relatively low.

Levels of perceived stress and frustration are no greater in the Intermediary condition when compared to Text, although for topic 57 there is a trend to more frustration. Again, this is encouraging – while the intermediary approach does take more effort, it does not appear to be more frustrating than text.

### 6.5 Conclusions

---

The aim of this work was to investigate the potential utility of using an intermediary to carry out content based searches on an image database. Importantly, the aim is not to replace text search, but rather provide a complementary technique which can be used in situations where conventional text and metadata search is not possible.

Using text retrieval as an upper bound (rather than a baseline), we have shown that the use of an intermediary image collection, while lower in performance, does enable users to find relevant material, albeit with more effort being extended. In particular, users must look through many more search results to find relevant images when using the intermediary approach. As may be expected, searches on semantic topics are less likely to be successful. For situations where no image annotations are available, these results are encouraging, and show that such a technique can be used to search un-annotated collections, such as those produced by PGP and other SALERO partners.

## 7 Vigor: Evaluation of a Grouping Interface

---

In this chapter, we present the evaluation of ViGOR (Video Grouping, Organisation and Retrieval) a video retrieval system that allows users to group videos in order to facilitate video retrieval tasks. In this way users are able to visualise and conceptualise many aspects of their search tasks and carry out a localised search in order to solve a more global search problem. The main objective of this work is to aid users while carrying out explorative video retrieval tasks; these tasks can be often ambiguous and multi-faceted.

Two user evaluations were carried out in order to evaluate the usefulness of this grouping paradigm for assisting users. The first evaluation involved users carrying out broad tasks on YouTube, and gave insights into the application of our interface to a vast online video collection. The second evaluation involved users carrying out focused tasks on the TRECVID 2007 video collection, allowing a comparison over a local collection, on which we could extract a number of content-based features. The results of our evaluations show that the use of the ViGOR system results in an increase in user performance and user satisfaction, showing the potential of a grouping paradigm for video search for various tasks in a variety of diverse video collections.

### 7.1 Motivation

---

Interactive video retrieval has the goal of alleviating some of the problems associated with video search [de Rooij 2008, Hauptmann 2006, Villa 2008], by offering the user specialised tools to more extensively and effectively explore a video collection. In this context, we have developed ViGOR, a video retrieval system that allows users to create semantic groups of results to help conceptualise and organise their results for complex video search tasks. This interactive grouping is a flexible means for the user to express short and long-term goals, as well as specific and multi-faceted information needs. The localised grouping also allows the user to focus on one particular aspect of a global task. We also believe that the semantic gap, [Jaimes 2005], which is the difference between low level features that machines use to represent multimedia and the high level concepts that humans associate with the same video, is shrunk by the construction of high-level semantic groupings rather than focusing on the search process, and also by a flexible user interaction with the video collection. We also believe that the use of this system can result in a number of advantageous outcomes for users: improved user performance in terms of task completion and task exploration, and increased user satisfaction with their search and their search results.

In order to evaluate the usefulness of ViGOR for assisting users in carrying out video search tasks a number of user evaluations were carried out. The first evaluation involved users carrying out broad tasks on YouTube and demonstrated the benefit of using this grouping functionality for broad, multi-faceted tasks on a large online collection. The second evaluation involved users carrying out focused TRECVID tasks on the TRECVID 2007 video collection, with the benefit of having relevance judgments on each task, this allowed us to provide a more rigorous evaluation of ViGOR, and further illustrate the benefits of the retrieval system.

### 7.2 The Vigor Search Interface

---

ViGOR (see Figure 12) comprises of a search panel (A), results display area (B), workspace (C) and playback panel (D). These facilities enable the user to both search and organise results effectively. The users enter a text based query in the search panel to begin their search. The result panel is where users can view the search results (a). Additional information about each video shot can be easily retrieved by placing the mouse cursor over a video keyframe for longer than 1.5 seconds, which will result in any text associated with that video being displayed to the user (we will hence forth refer to this action as tooltip) (e). If a user clicks on the play button the highlighted video shot will play in the playback panel. Users can play, pause, stop and navigate through the video as they can on a normal media player (D). Similar to the ImageGrouper [Nakazato 2003] and EGO [Urban 2006] systems, the main component of ViGOR is the provision of a workspace (C). Groups can be created by clicking on the create group button. Users must then select a textual label for the group and can potentially add any number of annotations to the group, but each group must have at least one annotation. Drag-and-drop techniques allow the user to drag videos into a group or reposition the group in the workspace (b). Groups can be deleted, minimised and moved around the workspace using a number of buttons (f). It should be noted that any

video can belong to multiple groups simultaneously. The workspace is designed to accommodate a large number of groups. Each group can also be used as a starting point for further search queries. Users can select particular videos and can choose to view an expansion of the group that contains similar videos based a number of different features (c, d). The description above describes the basic functionality of ViGOR; two slightly different versions of ViGOR were used for evaluation on two different datasets i.e. YouTube and TRECVID. These different interfaces as well as the baseline systems used for evaluation will be described in the experimental methodology. The following subsection presents a scenario describing how a user can potentially use the ViGOR system, with a discussion of the key advantages the system is intended to provide.

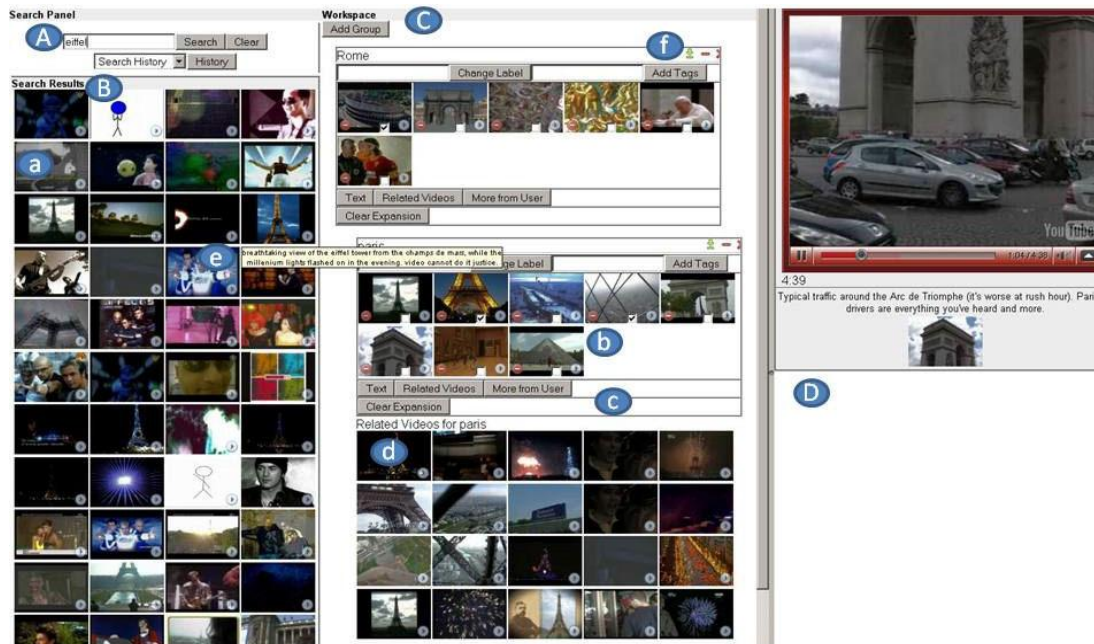


Figure 12: A screenshot of the ViGOR search interface

### 7.3 User Studies

In order to measure the effectiveness of ViGOR we conducted two user-centred evaluations. The goal of both evaluations was to investigate the effect of using the grouping paradigm available in ViGOR to help users complete video search tasks. There are a number of research questions that we wanted to address.

1. Does user performance improve with the grouping functionality available in ViGOR?
2. Do the workspace and the grouping functionality available in ViGOR allow users to explore more aspects of their task?
3. Will the ViGOR system increase user satisfaction with their search and the search process?

Two user evaluations were carried out in order to answer the research questions above. Both user evaluations were slightly different and allowed us to use different measures. The first evaluation was a between subjects evaluation that involved users carrying out broad tasks on YouTube, this provided us with a large and dynamic data collection, and facilitated the analysis of ViGOR in an online situation. The second evaluation was a within subjects evaluation involving users carrying out focused TRECVID tasks on the TRECVID 2007 video collection, this evaluation provided us with relevance judgments for task completion and also allowed users to compare ViGOR and a related baseline system directly.

For research question 1, a number of different measures of performance were used. For both evaluations we measured the number of videos marked as relevant for ViGOR and the baseline system, thus we can see if ViGOR is of benefit for both broad and specific tasks. In addition the relevance judgments supplied with the TRECVID 2007 collection allowed the measurement of precision and recall values for the second evaluation. Our hypothesis for research question 1 is:

- Hypothesis 1: Despite the overhead involved in the extra grouping functionality, that user's performance will improve using the grouping functionality in the ViGOR system in comparison

with an appropriate baseline system. (Mark more videos as relevant, find better videos, i.e. similar or improved precision and recall)

Research question 2 is slightly more difficult to quantify. It has been argued that for complex search tasks that an increase in user interactions is indicative of a good task performance, and is preferable for these kinds of tasks [Byström 2005]. Thus for the YouTube evaluation we will be looking at the user interactions with both systems to see if there is any change. In addition, for the YouTube evaluation we had independent judges make judgments on the results retrieved by users of the baseline system to discern how many unique aspects of the task the participants had investigated; this was compared with the number of groups that users of ViGOR created the complete the same tasks. The same measure was not appropriate for the TRECVID hypothesis, as these are very direct and unambiguous tasks and do not encompass as many aspects. Hypothesis for research question 2 is:

- Hypothesis 2: Users will explore more aspects of their task using ViGOR and that the workspace will help the users explore and see more options in large and unfamiliar datasets. (More interactions, explore more aspects for each task)

In order to address research question 3 we asked the users to complete a number of questionnaires at different stages of both evaluations. In both evaluations users were asked about the videos they were returned by the search system, their interaction with the search system, their search process, the task they had carried out and the search interface itself. As the TRECVID evaluation was a within subjects evaluation we also asked the participants in this evaluation to directly compare ViGOR with the baseline system. Using the results of all of these questionnaires we measured the user reactions to a number of aspects of the searches that they had carried out. Hypothesis 3 for research question 3 is:

- Hypothesis 3: Users will be more satisfied with their search results and the search process using ViGOR and in a direct comparison users will have a preference for the ViGOR system. (Satisfaction, user questionnaires)

In the following sections we will describe both evaluations in full detail and will also outline the results obtained for both evaluations.

## 7.4 Exploratory Task Evaluation

---

### Task and Collection

For the purposes of this evaluation we used the YouTube API to provide access to YouTube to provide a collection. Four simulated work task situations were created in order to provide broad, ambiguous, open ended tasks for the users [Borlund 2003]. These tasks were related to different topics and multiple aspects. Each task prompted the user to address at least two fixed aspects of the task, in order to aid our possible posterior analysis and comparisons. Users were encouraged to address at least one additional and open aspect of the task, there was no limit on the number of aspects the user could search for. The four evaluated simulated tasks were:

- A task of finding videos of political figures of 2008: George Bush and Barack Obama and at least one other political leader.
- A task of finding video clips about Paris, Rome and at least other Europe location.
- A task of finding videos that illustrate Scottish culture, in particular Scottish dancing and food, among other aspects.
- A task of finding the major sport news stories of 2008: Beijing Olympics, the Euro 2008 Football Championship and at least another event.

The ultimate goal of these simulated tasks was to write a short essay (e.g. a class project, a short description for a friend, etc.). In this way users were encouraged to carry out a deep exploration of the information addressed in the tasks and think thoroughly about a possible structure of the retrieved information. We thus avoid a “berry picking” effect, as users were encouraged to store only those videos that were potentially relevant for each task’s goal.

### Experimental Setup

A between subjects design was adopted for this evaluation. Two interfaces were evaluated; the first was ViGOR, which implemented the extra within group search functionality with the YouTube API. The interface offers three expansion options for each group (see Figure 12 (d)): 1) related videos; 2) videos from the same user 3) and text expansion which is the result of a new search using text extracted from the selected videos. All of the videos returned by these expansion options are retrieved using the

YouTube API. The second interface, which we will refer to as YouTube Interface (YI), mimicked the functionality of YouTube. Users could search via text and when a video was playing users were presented with lists of related videos and videos from the same user, in the same way that YouTube does, this also mimicked the functionality available through the group expansions explained above. In addition, users of the YI were provided with a panel where they could drag and drop relevant videos that they had found. We made the supposition that users of ViGOR would store relevant results in each group panel. Each participant carried out all four tasks either using the YI or ViGOR. The order of tasks was varied; this was to avoid any order or learning effect associated with the tasks. Using this experimental model we can evaluate the usefulness of the grouping functionality in ViGOR's workspace for helping users to complete open ended and broad tasks. Each participant was given five minutes training on their search system and was allowed to carry out training tasks. Users had a maximum of 20 minutes to complete each of these tasks. For each participant their interaction with the system was logged, the videos they marked as relevant were stored and they also filled out a number of questionnaires at different stages of the experiment.

16 participants took part in our evaluation, they were randomly divided into two groups of 8 and each group used one of the systems. The participants were mostly postgraduate students and researchers at a university. The participants consisted of 12 males and 4 females with an average age of 29 years (median: 27.5) and an advanced proficiency with English. The participants indicated that they regularly interacted with and searched for multimedia. They were paid a sum of £12 for their participation in the evaluation, which took approximately 2 hours.

#### 7.4.1 Exploratory Task Evaluation Results

##### System Performance

In a direct comparison between the two interfaces it was found that on average users of ViGOR marked 35.09 videos as being relevant (by assigning them to a group) in comparison with 23.16 videos for users of the YI. This was also achieved in less time, with users of ViGOR completing their task in 18.6 minutes in comparison with 19.06 minutes for users of the YI. While the difference in time may not be that noteworthy the increase of over 50% in the number of retrieved videos is.

In order to gain a further insight into the difference in the performance between the two interfaces, a further analysis of the logs was carried out to investigate the user interactions, the results of this analysis is shown in Table 21. It can be seen in Table 21 that the ViGOR's users have more user interactions with the system overall in comparison with users of the YI. This is an encouraging result, as more interactions for complex tasks are a preferable result [Byström 2005]. In addition, most of this difference is due to the increased use of the tooltip functionality of the ViGOR users; this is a lightweight functionality which is of low cost for the user to carry out. In comparison, actions that require more effort and time decrease on ViGOR in comparison with the users of YI. Notably, YI users viewed 19.9% more videos and issue 8.41% more queries. In conclusion, it can be seen quite clearly that users of ViGOR find more videos in less time. Also the overall number of interactions is higher which is a positive result, however the number of high cost interactions in terms of user effort are reduced by using ViGOR freeing the user to explore the task and collection.

Users of the ViGOR systems created an average of 4.5 group panels, which shows that they went well beyond the 3 mandatory aspects and were comfortable using the interface to investigate the number of aspects. We evaluated the results and essays created by the users of the YI interface to determine the number of aspects that were investigated. Users of the YI interface explored slightly less aspects than the ViGOR interface, 4.1. This indicates that the users that were using ViGOR investigated more aspects of the tasks. Furthermore, this fact together with the higher number of relevant documents retrieved indicates that ViGOR users created more complex and more detailed aspects than those of the YI.

Interface	YouTube Interface		ViGOR	
	Number	%	Number	%
Tooltip	240.25	60.74%	317.50	61.16%
View	25.13	5.43%	20.13	3.88%
Query	56.50	13.86%	51.75	9.97%
Relevant	88.50	19.48%	126.63	24.39%
Irrelevant	2.38	0.49%	3.13	0.6%
Total	412.15	100%	654.88	100%

**Table 21: Total number of different interactions for each user for each interface. Interactions are for unique videos e.g. if a user plays the same video twice we only record it once**

## User Feedback

In post search task questionnaires we solicited subjects' opinions on their assigned system and their reaction to the retrieved videos. The following 5-point Likert scales and semantic differentials were used. "The videos that I have received through the searches were" "Relevant / Irrelevant" (Relevant), "Appropriate / Inappropriate" (Appropriate), "Complete / Incomplete" (Complete) and "Familiar / Strange" (Familiar). "I had an idea of which kind of videos were relevant for the topic before starting the search" (Prior). "I found it easy to formulate queries on this topic" (Formulate). "During the search I have discovered more aspects of the topic than initially anticipated" (Discover). "The video(s) I chose in the end match what I had in mind before starting the search" (Match). "The tools provided allowed me to find videos that matched the topic" (Tools). "My idea of what videos and terms were relevant changed throughout the task" (Change). "I am satisfied with my search results" (Satisfy). Table 22 presents the average responses for each of these scales using the labels after each of the Likert scales in the list above. The most positive response for each user type is shown in bold.

<i>Differential</i>	<i>YouTube Interface</i>	<i>ViGOR</i>
<i>Relevant</i>	4.03125	<b>4.09375</b>
<i>Appropriate</i>	3.875	<b>4.09375</b>
<i>Complete</i>	3.12875	<b>3.375</b>
<i>Familiar</i>	<b>3.90625</b>	3.5
<i>Prior</i>	<b>4</b>	3.875
<i>Formulate</i>	3.8125	<b>4.15625</b>
<i>Discover</i>	2.875	<b>3.3125</b>
<i>Match</i>	<b>3.7185</b>	3.65625
<i>Tools</i>	3.84375	<b>4.1875</b>
<i>Change</i>	2.6875	<b>2.875</b>
<i>Satisfy</i>	3.65625	<b>3.75</b>

**Table 22: Perceptions of Retrieved Videos (Higher = Better)**

From the results in Table 22 it appears that participants have a better perception of the retrieved videos while interacting with ViGOR. The trend indicates that the users believe that they found more relevant, appropriate and diverse videos while using ViGOR in comparison with the YouTube interface. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the 4 tasks to test these assertions. None of the differences for the findings in Table 22 were found to be statistically significant. However a number of findings in other parts of the questionnaires were found to be significant. When asked about their task performance, users of YI had a much stronger perception that the video collection didn't contain video(s) they wanted ( $F=6.5$ ,  $p=0.0134$ ), that the YI system didn't return relevant videos ( $F=10.64$ ,  $p=0.0018$ ) and that they did not have enough time to complete the task ( $F=5.18$ ,  $p=0.0265$ ). These findings are consistent with the findings of the questionnaires relating to the retrieved videos (see Table 22). Also, when asked for their reaction to the system via semantic differentials a significant difference was found between ViGOR and the YI on how novel ( $F=4.93$ ,  $p=0.0434$ ) and flexible ( $F=8.14$ ,  $p=0.0128$ ) the users perceived both systems to be. In summary, the first evaluation has shown that ViGOR users retrieved more videos in less time and with less expensive interactions in comparison with a comparable baseline interface. The users of ViGOR also perceived that they retrieved more relevant, appropriate and diverse videos through ViGOR even though the retrieval mechanisms were the same for both systems. In addition, these users found ViGOR to be more novel and flexible than the YI.

## 7.5 TRECVID Evaluation

In order to provide further validation for these findings a second evaluation was carried out. This evaluation used the TRECVID collection and tasks. This allowed us to calculate precision and recall values for the retrieved results. However, the TRECVID tasks are much more focused tasks, ViGOR was not designed with these types of task in mind, but it is hoped that ViGOR can still aid user performance. Furthermore, this evaluation had a within user Latin Square design, which allowed us to directly compare user reactions to the two interfaces, as users used ViGOR and a baseline system to complete search tasks.

### Tasks and Collection

There are two main reasons for using the TRECVID 2007 collection. First, the TRECVID collection is a large, well known and commonly used video collection. Secondly, the TRECVID collection has a number of tasks for which the relevant and irrelevant video shots in the collection are known. In 2007

the TRECVID collection contained 18,142 shots (over 100 hours) of Dutch magazine television. For the TRECVID 2007 interactive search evaluations there were a total of 24 tasks. For our evaluation we limited the number of tasks that the users carry out to 8. This allowed us to carry out more evaluations, as 24 individual search topics did not have to be carried out for each participant. In order to examine user interactions on different types of tasks we choose the 8 tasks which had the highest number of shots marked as being relevant during TRECVID runs. The 8 tasks were

1. Find shots of a person walking or riding a bicycle (1175 relevant shots)
2. Find shots of a woman talking toward the camera in an interview - no other people visible (400 relevant shots)
3. Find shots of one or more people playing musical instruments such as drums, guitar, flute, keyboard, piano, etc. (376 relevant shots)
4. Find shots with hills or mountains visible (343 relevant shots)
5. Find shots with 3 or more people sitting at a table (332 relevant shots)
6. Find shots of waterfront with water and buildings (265 relevant shots)
7. Find shots of a very large crowd of people (fills more than half of field of view) (264 relevant shots)
8. Find gray scale shots of a street with one or more buildings and one or more people (210 relevant shots)

### **Experimental Setup**

For our evaluation we adopted a 2-searcher-by-2-topic within subject's Latin Square design. Two interfaces were evaluated; the first was ViGOR. The interface offers three expansion options for each group (see Figure 12(d)): 1) similar colour; 2) similar shapes, this was retrieved using edge histograms 3) and similar homogenous texture. This functionality allowed users to use videos as examples and retrieved the most similar videos based on the low level feature selected from the collection. The second interface, which we will refer to as the Search Interface (SI), allowed the users to query the collection via query by text and query by example. The main difference between both interfaces was the lack of grouping functionality in the SI. Also by including query by text and query by example in both interfaces we are including the two most widely used search functionalities at TRECVID [Christel 2007]. Users of both systems were also provided with a panel where they could drag and drop relevant videos for the task. This design encouraged users to add videos to groups that may help their search task, but may not be relevant for the very specific topics that are part of TRECVID. Otherwise users may have been discouraged from adding irrelevant shots to their groups. This also gave the users the option of completely bypassing the grouping functionality if they wished, as they could just interact in the same way as they would with the SI, giving the grouping functionality more of an add on feel, rather than it being the complete focus of the evaluation, which may bias some users. Each participant carried out two tasks using the SI, and two tasks using ViGOR. The order of system usage was varied as was the order of the tasks; this was to avoid any order effect associated with the tasks or with the systems. Each participant was given five minutes training on each system and was allowed to carry out training tasks. These training tasks were also tasks from TRECVID 2007, which we found appropriate as they also had relatively high numbers of relevant documents. Each actual task had a fifteen minute maximum time limit. For each participant their interaction with the system was logged, the videos they marked as relevant were stored and they also filled out a number of questionnaires at different stages of the experiment.

### **Users**

16 participants took part in our evaluation. The participants were mostly postgraduate students and researchers at a university. The participants consisted of 13 males and 3 females with an average age of 25.6 years (median: 26) and an advanced proficiency with English. The participants indicated that they regularly interacted with and searched for multimedia. They were paid a sum of £15 for their participation in the experiment, which took approximately 2 hours.

### **7.5.1 TRECVID Evaluation Results**

#### **System Performance**

As we were using the TRECVID collection and tasks, we were able to calculate precision and recall values. The results of our evaluation show that ViGOR outperforms the SI on a number of measures for the majority of the tasks. When using ViGOR users marked more shots as relevant and indeed found more shots in the collection which are actually relevant for 6 of the 8 tasks (see Figure 13).

For the other 2 tasks (Task 6 and Task 7) the SI performs significantly better in terms of the number of videos marked as relevant and the number of relevant videos found. A further analysis of the user logs was carried out in order to find an explanation for this. It was found that for these tasks the user success was down to navigation rather than using the search tools. Some of the users found a sequence of video shots which contained large numbers of relevant shots; the users would navigate through the video marking each shot as relevant, this is not normally a realistic search solution for large video collections. For example in YouTube where navigation facilities are available, most of the video views can be attributed to searching rather than navigation Figure 13. Also, for the same tasks using ViGOR, users persisted with using the search tools. The relative success of the SI for these tasks can then be attributed to the nature of the collection and tasks. In order to gauge the actual success of ViGOR we also analysed the user performance in terms of precision, recall and mean average precision (MAP). MAP is the average of the precision values calculated at every position where a relevant document appears in the result list, and is normally used for a simple and convenient system performance comparison. MAP combines both recall and precision oriented measures and gives an overall measure of the performance of the system, so for this reason we concentrate on the MAP of the results obtained by the users. We see that for 4 of the 8 tasks the MAP when using ViGOR is higher (see Figure 14). This is somewhat skewed by the fact that the users of ViGOR retrieved more videos and also the performance of SI users for tasks 6 and 7. Overall it was found that when using ViGOR users retrieved more videos, more relevant videos and despite the fact that ViGOR was not designed for direct TRECVID type tasks, it is comparable with and in some cases outperforms a more traditional video search interface.

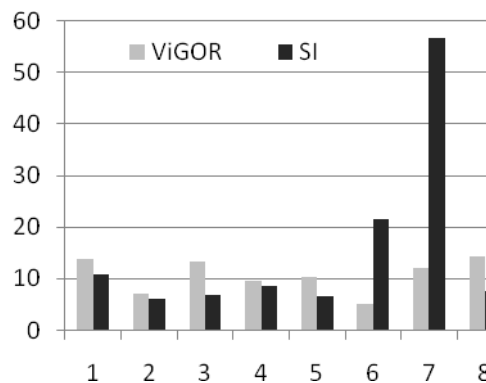


Figure 13: Number of relevant shots found for each task and system combination

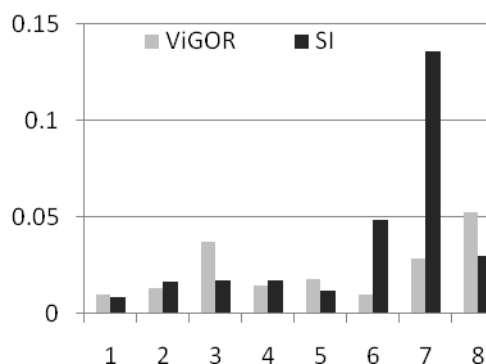


Figure 14: MAP for each task and system combination

### 7.5.2 User Feedback

With the intention of providing further validation for our findings and to gauge user perceptions, we analysed the post task and post experiment questionnaires that our participants filled out.

#### System and Interaction

In post search task questionnaires we solicited subjects' opinions on and reaction to the system. The following semantic differentials were used to solicit user's reaction to the system; "terrible/wonderful", "satisfying/frustrating", "dull/stimulating", "easy/difficult", "rigid/flexible", "efficient/inefficient",

“novel/standard” and “effective/ineffective”. For each of these differentials we assign a value of 5 to the most positive response and 1 to the most negative. Table 23 presents the average responses for each of these differentials. The most positive response across for each differential is shown in bold.

<i>Differential</i>	<i>ViGOR</i>	<i>SI</i>
<i>Terrible/Wonderful</i>	<b>3.6451</b>	3.2222
<i>Satisfying/Frustrating</i>	<b>3.5806</b>	3.3056
<i>Dull/Stimulating</i>	<b>3.4193</b>	2.9722
<i>Easy/Difficult</i>	<b>3.9356</b>	3.8056
<i>Rigid/Flexible</i>	<b>3.7097</b>	2.9444
<i>Efficient/Inefficient</i>	<b>3.4838</b>	2.9444
<i>Novel/Standard</i>	<b>3.9555</b>	2.8056
<i>Effective/Ineffective</i>	<b>3.6129</b>	3.0556

**Table 23: Perceptions of System and Interaction (Higher = Better)**

From the results in Table 23 it appears that participants have a better perception of the system while interacting with ViGOR. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the 8 tasks to test these assertions. It was found that the differences in how wonderful ( $F=5.21$ ,  $p=0.0263$ ), flexible ( $F=6.58$ ,  $p=0.0131$ ), efficient ( $F=5.83$ ,  $p=0.0191$ ), novel ( $F=15.07$ ,  $p=0.0003$ ) and effective ( $F=4.61$ ,  $p=0.0362$ ) the users found the system, was system dependent. This demonstrates that the users had a strong preference for ViGOR, and that they felt it was more flexible and effective for their search tasks, thus providing a better user experience. This finding shows that ViGOR is providing a better search experience for the user, and that they are more at ease and confident using the this system. These results also validate the results from user questionnaires that were found in the first evaluation.

### System Support

In post search task questionnaires we also solicited subjects' opinions on their interaction with the systems and the support for their search tasks that each system provided. The following 5-point Likert scales were used. “The system was effective for solving the task” (effective) and “the system helped me to” ... “explore the collection” (explore), “find relevant videos” (relevant), “express different aspects of the task” (aspect), “focus my search” (focus), “find videos that I would not have otherwise considered” (consider). Table 24 presents the average responses for each of these scales using the labels after each of the Likert scales in the list above. The most positive response for each user type is shown in bold.

Once again it can be seen from the results in Table 24 that participants have a better perception of ViGOR. The users found ViGOR to be superior to the SI system with regards to a number of aspects. We applied two-way ANOVA to each scale across both systems and the 8 tasks to test these assertions. It was found that the differences in the scales effective ( $F=6.14$ ,  $p=0.0146$ ), aspect ( $F=6.9$ ,  $p=0.0111$ ) and focus ( $F=5.07$ ,  $p=0.0284$ ) were statistically significant with regards to the system used. Once again this demonstrates that the users had a strong preference for ViGOR, and that they felt while they were able to express different aspects of the task in hand, the system also helped them to focus on solving the task. As the second evaluation is within subjects, as opposed to the first evaluation which was between subjects, we were able to solicit user's opinions about both interfaces. In the exit questionnaires while the users stated that they found the SI easier to use (ViGOR = 2, SI = 10, Undecided = 4) and easier to learn to use (ViGOR = 1, SI = 10, Undecided = 5), that they still had a preference for the ViGOR (ViGOR = 12, SI = 3, Undecided = 1) and found it to be better overall (ViGOR = 11, SI = 3, Undecided = 2), highlighting the potential of ViGOR as far as the users were concerned. The following section will provide a discussion of all of the findings of both evaluations and the ViGOR system overall.

<i>Scale</i>	<i>ViGOR</i>	<i>Search</i>
<i>Effective</i>	<b>3.8065</b>	3.25
<i>Explore</i>	<b>3.3226</b>	2.8333
<i>Relevant</i>	<b>3.3548</b>	2.8889
<i>Aspect</i>	<b>3.4516</b>	2.8889
<i>Focus</i>	<b>3.5161</b>	3.0556
<i>Consider</i>	<b>3.2258</b>	2.8611

**Table 24: Perceptions of System Support (Higher = Better)**

## 7.6 Conclusions

---

In this chapter we have introduced the ViGOR system, a video search and retrieval system that allow users to create groups of results to help conceptualise and organise their results for complex video search tasks. It was hoped that grouping search results on the workspace would motivate the user to organise results for their search/work task. This should enable the users to break up their global search task into a small set of individual search tasks. Although the concept of grouping has been investigated in a number of retrieval scenarios [Nakazato 2003, Urban 2006], its application and value for searching video collections and archives has not been formally evaluated. As has been discussed previously, video provides a number of unique problems for search and retrieval that are not present in other search scenarios. Thus there are a number of important contributions that this chapter makes. First, we present the first system that allows explicit grouping as part of the video search process. Second, we show how this search paradigm can be applied to a number of video search scenarios and video libraries, i.e. YouTube and TRECVID. Our goal in this chapter was to investigate three hypotheses relating to the use of ViGOR: 1) that user performance would improve through the use of ViGOR, 2) that ViGOR can aid user exploration of the task at hand and 3) that the use of ViGOR can also increase user satisfaction with their search and their search results. To that end we have conducted two user evaluations, involving in total 32 participants, on a variety of very different video search tasks that incorporate different user goals. In the first part of the evaluation users searched over YouTube and carried out broad, multi-faceted search tasks. In the second part of the evaluation the users searched over the TRECVID 2007 collection and carried focused search tasks.

There are a number of interesting points that can be made about the results of these evaluations. For both search scenarios it was found that the use of the grouping functionality resulted in users retrieving more search results in comparison with a baseline system. In the YouTube evaluation users retrieved approximately 50% more videos when using the grouping interface. This increase in the number of retrieved videos was also coupled with an increase in user interactions. However most of this can be attributed to non-expensive lightweight functionalities, while more expensive heavyweight functionality decreases, in comparison with the baseline interface, users of the grouping interface viewed 18% less videos and carried out 5% less queries. These videos were also retrieved by these users in less time than users of the baseline system. For the TRECVID collection the difference in the number of retrieved videos is not as large as in the YouTube evaluation. However, the ViGOR system was designed with vague and multi-faceted search tasks in mind. The fact that there is any increase for these focused search tasks is a positive result. As we were using the TRECVID collection and tasks, we have relevance judgements which allowed us to carry out some analysis that we could not carry out for the YouTube evaluation. It was found that as well as the users of ViGOR were retrieving more video shots, that they were also in fact retrieving more relevant shots overall. In terms of MAP the grouping interface is more than comparable with the baseline, indeed outperforming it for half of the tasks. Once again this result is encouraging as this type of grouping interface has been designed for broad, exploratory search tasks and not the focused tasks that are part of TRECVID. Overall it can be seen that the availability of the grouping functionality improves user performance when searching digital video libraries. These results provide validation for our first two hypotheses.

There were also a number of interesting findings in terms of user perceptions. In both user evaluations the differences in how novel and flexible the users found the system were both system dependent and statistically significant. In addition, for the TRECVID evaluation where the users were exposed to both a baseline system and a system with grouping functionality it was found that the differences in how wonderful, efficient and effective the users found the system, was system dependent. Also for this evaluation the users stated that the grouping functionality helped them to focus on the task in hand while being able to express different aspects of the search tasks. For the TRECVID evaluation we also asked users for a direct comparison between the grouping and baseline interfaces. Whereas the users stated that they found the baseline easier to use and easier to learn to use, they still had a preference for the grouping interface and found it to be better overall. Although users could not directly compare the interfaces in the YouTube evaluation the users did not highlight any statistically significant difference between how easy to use and how easy it was to learn how to use the grouping functionality. These results demonstrate that the users had a strong preference for the interfaces that provide the grouping functionality. They found it more flexible, effective and interesting than the baseline, and also found themselves able to express more aspects of the topics while still focusing on the task in hand, thus providing a better user experience and validating our third hypothesis.

Overall it can be seen that the addition of grouping functionality for video search tasks can lead to a number of favourable outcomes. In terms of task performance users retrieve more search results in less time with less search interactions. With respect to user perceptions the users find the grouping systems to be more flexible and novel, in a direct comparison between a grouping interface and a baseline system the grouping functionality helped users to focus on the task in hand while being able to express different aspects of the search tasks. While these findings illustrate the benefit of the grouping metaphor for video search, there are also a number of benefits that occur as a result of using a grouping search metaphor. The interactive grouping is a supple means of communicating a multitude of information needs e.g. short-term vs. long-term, specific vs. multi-faceted. The semantic gap is narrowed by the abstraction to high-level semantic groupings, reflecting an individual's task-specific mental model of the data. Finally, the user leaves a trail of their interactions, which can not only be exploited by the system for adaption but by which can be used by other users for collaboration.

## 8 Conclusions

---

In this report, a range of different evaluations have been presented, both automatic test-collection evaluations and end-user evaluations. In the development of any system, both kinds of evaluation technique are inevitably required. In the development of algorithms and underlying technology elements, such as the backend retrieval engine, automatic evaluation is preferred, allowing developers to quickly test a large range of different technological options with relative ease.

In Chapter 3 we presented a summary of the evaluation work which has gone hand in hand with the development of the underlying content-based search system, designed for the SALERO project. During the project, a range of different techniques have been tested, including different types of visual features (e.g. global MPEG-7 features, SIFT features), high-level concepts when available in collections, and retrieval techniques (e.g. clustering and LDA based methods as used in TRECVID 2009). This automatic evaluation has allowed us to try many different techniques, providing a proving ground for methods, the more successful techniques being rolled into the main SALERO content-based search system. The experience of carrying out this work has also proved invaluable in the design of the retrieval system for the Alan Online experimental production, developed by TAIK, where similar evaluation techniques were used to create a new system with an improved performance.

The results from the automatic annotation work reported in Chapter 4 are inevitably more preliminary, given the ongoing nature of this research. Image annotation is potentially a powerful method of bridging the semantic gap, but is still very much in its infancy when considering the performance of state of the art systems. The work reported in Chapter 4 considers the evaluation of automatic annotation using more realistic evaluation collections and topics than often the case in current research, and provides a way of judging the current state of such systems in practice.

Chapter 5 describes recent work which was undertaken to improve the retrieval performance of the content-based search system for the Alan online experimental production, as requested by the SALERO reviewers in the IBC review meeting in 2009. An improved retrieval algorithm which included the use of a wider range of data, plus a new fusion techniques using three of the most effective features for the TAIK data, allow us to significantly improve the performance on a specially created test set. Using the new technique and the created test queries, we are able to retrieve the correct Alan online symbol 82% of the time, a considerable improvement on the previous technique.

While the problem of retrieving on the Alan online collection of symbols is an important one, and one which was specifically requested, the more general problem of retrieval from un-annotated collections of images remains an important one for SALERO. For example, the data set of image assets provided by PGP, and used in the My Tiny Planets experimental production, consists of a large number of images without descriptive names or other textual information. Methods for searching such collections automatically, without the expensive of manual annotation as required in semantic search, is of considerable importance. A longer term solution to this problem is automatic annotation, as described in Chapter 4. Unfortunately, such systems are still in their infancy. As an alternative, the technique of searching via an intermediary collection has been investigated, as supported in SALERO's AspectBrowser interface. This techniques involves a user search an intermediary collection which *does* contain text to find image examples, which can then be used to search a target collection which *does not* contain textual annotations.

Chapter 6 describes this technique, and an evaluation which was carried out on a specially designed interface to investigate the potential utility of the method compared to direct text search. As was expected, text annotations work better when they exist, although the performance of users using the intermediary technique was extremely encouraging: for visual topics, users were able to perform roughly half as well as text search. For the more semantically oriented topics, performance was lower, but users were still able to find material relevant to the search topics. While this technique does require more effort to be expended on search tasks, it has the huge advantage of being available now, and supported by the AspectBrowser interface.

The final chapter of the report presents and discusses the ViGOR grouping interface, a search interface which was designed as complimentary to the AspectBrowser interface. This interface provides a different method of search result organisation, providing users with a workspace in which images or videos can be organised. The results of user testing are encouraging, and we are hopeful that further work comparing ViGOR and the AspectBrowser will bring to light the advantages and disadvantages of both interfaces and their respective organisation features.

This report, therefore aims to present the past and current state of the search evaluation which has been carried out during the SALERO project: from the underlying content-based techniques developed and described in Chapter 3, through the specific problem of retrieval for the Alan online experimental production, the current performance of the intermediary retrieval technique supported by the AspectBrowser interface and described in Chapter 6, and finally outlining the more long-term research problems, such as automatic annotation. Future work aims to consolidate the gains made during SALERO, through the basic research carried out and the development of the SALERO tools and software, by continuing to develop promising techniques (e.g. the LDA retrieval technique used in TRECVID 2009, Section 3.6), and to further evaluate the trade-offs inherent in the systems developed, such as the pros and cons of the AspectBrowser and ViGOR methods of interaction.

## 9 References

---

- [Agichtein 2006] Agichtein, E., Brill, E., Dumais, S.: Improving Web Search Ranking by Incorporating user behavior information. In: Proc. SIGIR'06 (2006) 19-26
- [Andre 2006] Andre, D., Pelletier, R., Farrington, J., Safier, S., Talbott, W., Stone, R., Vyas, N., Trimble, J., Wolf, D., Vishnubhatla, S., Boehmke, S., Stivoric, J., Teller, A.: The Development of the SenseWear Armband, A revolutionary energy assessment device to assess physical activity and lifestyle. Technical Report (2006)
- [Arapakis 2008] Arapakis, I., Jose, J.M., Gray, P.D.: Affective feedback: An investigation into the role of emotions in the information seeking process. In: SIGIR'08 (2008) 395-402
- [Arapakis 2009] Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., Jose, J.M.: Enriching user profiling with affective features for the improvement of a multimodal recommender system. In Conf. on Image and Video Retrieval (2009)
- [Arjan 2009] Arjan T. Setz and Cees G. M. Snoek. (2009) Can social tagged images aid concept-based video search? ICME 2009
- [Attias 2000] Attias, H.: A variational bayesian framework for graphical models. In: In Advances in Neural Information Processing Systems 12, MIT Press (2000)
- [Ayache 2007] Ayache, S., Quénot, G.: Trecvid 2007 collaborative annotation using active learning, TRECVID'2007 Workshop (November 2007)
- [Badi 2006] Badi, R., Bae, S., Moore, J.M., Meintanis, K., Zacchi, A., Hsieh, H., Shipman, F., Marshall, C.C.: Recognizing user interest and document value from reading and organizing activities in document triage. In Proc. 11th International Conference on User Interfaces (2006) 218-225
- [Bagga 1998] Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proc. Comput. Ling. (1998) 79-85
- [Bailey 2003] Bailey, P., Craswell, N., Hawking, D.: Engineering a multi-purpose test collection for Web Retrieval experiments. Inf. Process. Manage. 39(6) (2003) 853-871
- [Belkin 1980] Belkin, N.J. Anomalous state of knowledge for information retrieval. In: Canadian Journal of Information Science 5 (1980) 133-143
- [Bilal 2002] bilal, D., Kirby, J.: Differences and Similarities in Information Seeking: Children and Adults as Web Users. Information Processing and Management: An International Journal 38(5) (2002) 649-670
- [Bishop 2006] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
- [Blei 2003] Blei, D.M., Jordan, M.I.: Modeling annotated data. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, USA, ACM Press (2003) 127-134
- [Borlund 2003] Borlund, P. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Inf. Res. 8(3). (2003).
- [Bucsein 1992] Bucsein, W. Electrodermal activity. Plenum Press (1992)
- [Byström 2005] Byström, K. and Järvelin, K. 1995. Task complexity affects information seeking and use. Inf. Process. Manage. 31, 2, pp 191-213. (2005)
- [Campbell 1996] Campbell, I., van Rijsbergen, C.J.: The Ostensive Model of developing information needs. In: Proc. Library Science (1996) 251-268
- [Campbell 2000] Campbell, I. Interactive evaluation of the Ostensive Model, using a new test-collection of images with multiple relevance assessments. Inf Retr 2(1):89-114. (2000)
- [Carneiro 2007] Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. TPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3) (2007) 394-410
- [Chan 1997] Chan, Y and Kung, S Y (1997) A hierarchical algorithm for image retrieval by sketch. First IEEE Workshop on Multimedia Signal Processing, 564-569

- [Chan 2005] Chan, C.H., Jones, G.J.F.: Affect-based indexing and retrieval of films. In: MM'05 (2005) 427-430
- [Chanel 2008] Chanel, G., Rebetez, C., Betrancourt, M., Pun, T.: Boredom, engagement and anxiety as indicators for Adaptaion to difficulty in games. In Mindtrek'08 (2008) 13-17
- [Christel 2006] Christel, M.G. and Conescu, R.M. Mining Novice User Activity in TRECVID Interactive Retrieval Tasks. In Proc CIVR 2006, 21-30. (2006)
- [Christel 2007] Christel, M.G. Establishing the Utility of Non-Text Search for News Video Retrieval with Real World Users. In Proc ACM MM 2007, 707-716, (2007).
- [Cleverdon-1967] Cleverdon, C. W. The Cranfield tests on index language devices. In Aslib proceedings, volume 19, pages 173-192, 1967 (Reprinted in Readings in Information Retrieval, K. Spark-Jones and P. Willett, editors, Morgan Kaufmann, 1997)
- [Cleverdon-1991] Cleverdon, C. W. The significance of the Cranfield tests on index languages. In proceedings of the Fourteenth Annual Internation ACM/SIGIR Conference on Research and Development in Information Retrieval, pages 3-12, 1991
- [Cornelius 2000] Cornelius, R.R. Theoretical Approaches to Emotion. Proc. ISCA (2000) 3-11
- [Craswell 2007] Craswell, N. and Szummer, M., Random walks on the click graph. In Proc. SIGIR 2007, ACM Press (2007), 239-246.
- [Daelemans 2005] Daelemans, W., van den Bosch, A., Memory-based Language Processing. Cambridge University Press (2005)
- [Damasio 1994] Damasio, A.R: Descartes Error: Emotion, Reason and the Human Brain (1994)
- [de Rooij 2008] De Rooij, O., Snoek, C.G.M. and Worring, M. MediaMill: fast and effective video search using the forkbrowser. In proceedings of CIVR 2008, pp 561-562. (2008).
- [Dudev 2008] Dudev, M., Elbassuoni, S., Luxenburger, J., Ramanath, M., Weikum, G.: Personalizing the Search for Knowledge. In: Proc. PersDB (2008)
- [Duygulu 2002] Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV'02: Proceedings of the 7th European Conference on Computer Vision-Part IV, London, UK, Springer-Verlag (2002) 97-112
- [Ekman 1979] Ekman, P., Oster, H.: Facial Expressions of Emotion. In: Annual Review of Psychology 30(1) (1979) 527-554
- [Ekman 1999] Ekman, P. Facial Expressions. Handbook of Cognition and Emotion. John Wiley & Sons Ltd. (1999)
- [Ekman 2003] Ekman, P. Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life. Time Books (2003)
- [Ekman 2003b] Ekman, P.: Unmasking the Face. Malor Books (2003)
- [Fass 2000] Fass, A.M., Bier, E.A. and Adar, E. PicturePiper: using a re-configurable pipeline to find images on the Web. In Proceedings of UIST 2000. pp 51-62
- [Fei-Fei 2006] Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Analysis & Machine Intelligence 28(4) (April 2006) 594\_611
- [Feng 2004] Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: CVPR '04 Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (27 June-2 July 2004)
- [Figueiredo 2002] Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3) (2002) 381-396
- [Flickner 1997] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. and Yanker, P. (1997) Query by Image and Video Content: The QBIC System, Intelligent Multimedia Information Retrieval, MIT Press. 7-22.
- [Fogarty 2008] Fogarty, J., Tan, D.S., Kapoor, A. and Winder, S.A.J. CueFlik: interactive concept learning in image search. In Proceedings of CHI 2008. pp 29-38. (2008)

- [Grubinger 2007b] Grubinger, M. Clough, P. (2007) On the Creation of Query Topics for ImageCLEFphoto. Proceedings of the Third Workshop on Image and Video Retrieval Evaluation, pages 50-63, Budapest, Hungary
- [Grubinger 2007] Grubinger, M. (2007) Analysis and Evaluation of Visual Information Systems Performance. PhD Thesis. Victoria University, Melbourne, Australia
- [Guy 2006] Guy, M. and Tonkin, E. Folksonomies Tidying Up Tags, D-Lib Magazine, Volume 12, Number 1, (2006).
- [Halvey 2007] Halvey, M. and Keane, M.T. Analysis of Online Video Search and Sharing. In Proc. ACM HT 2007, ACM Press (2007), 217-226.
- [Hanjalic 2005] Hanjalic, A., Xu, L.Q.: Affective Video Content Representation and Modeling. In: Transactions on Multimedia 7(1) (2005) 143-154
- [Hauptmann 2004] Hauptmann, A.G. and Christel, M.G. Successful approaches in the TREC video retrieval evaluations, ACM Multimedia 2004: 668 – 675.
- [Hauptmann 2006] Hauptmann, A.G., Lin, W-H, Yan, R., Yang, J. and Chen M-Y. Extreme video retrieval: joint maximization of human and computer performance. In proceedings of ACM Multimedia 2006, pp 385-394. (2006)
- [Hopfgartner 2007] Hopfgartner, F.: A news video retrieval framework for the study of implicit relevance feedback. In Proc. of the Second International Workshop on Semantic Media Adaptation and Personalization (2007) 233-236
- [Hopfgartner 2007] Hopfgartner, F. Understanding Video Retrieval. VDM Verlag (2007)
- [Hopfgartner 2008] Hopfgartner, F., Vallet, D., Halvey, M. and Jose, J.M. Search trails using user feedback to improve video search. In Proceedings of ACM Multimedia 2008, pp 339-348. (2008)
- [Jaimes 2005] Jaimes, A., Christel, M., Gilles, S., Ramesh, S., and Ma, W-Y. Multimedia Information Retrieval: What is it, and why isn't anyone using it? In Proc MIR, ACM Press, 3–8, (2005).
- [Jeon 2003] Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, USA, ACM Press (2003) 119-126
- [Joachims 1999] Joachims, T.: Making large-scale support vector machine learning practical. (1999) 169\_184
- [Jordan 1998] Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. In: Proceedings of the NATO Advanced Study Institute on Learning in graphical models, Norwell, MA, USA, Kluwer Academic Publishers (1998) 105-161
- [Järvelin 2001] Järvelin, K., Kekäläinen, J., Niemi, T.: Expansion Tool: Concept-Based Query Expansion and Construction. Information Retrieval, 4(3), (2001) 231-255
- [Kataoka 1998] Kataoka, H., Kano, H., Yoshida, H., Saijo, A., Yasuda, M., Osumi, M. Development of a skin temperature measuring system for non-contract stress evaluation. In Proc. 20th Annual international Conference of the IEEE Engineering in medicine and Biology Society, volume 2 (1998) 940-943
- [Kelly 2003] Kelly, D. Teevan, J.: Implicit feedback for inferring user preference: A bibliography. SIGIR Forum 37(2) (2003), 18-28
- [Koenemann 1996] Koenemann, J., Belkin, N. J.: A Case for interaction: A study of interactive information retrieval behavior and effectiveness. In: CHI'96 (1996) 205-212
- [Kraaij 2009] Kraaij, W. and Awad, G. (2008) TRECVID 2008 High-Level Feature Task: Overview. Available at <http://www-nlpir.nist.gov/projects/tvpubs/tv8.slides/tv8.hlf.slides.pdf> (retrieved 10th Sept 2009)
- [Kracker 2002] Kracker, J.: Research Anxiety and Students' perceptions of research: An Experiment. Part i: Effect of teaching Kuhlthaus ISP Model. Journal of the American Society for Information Science and Technology 53(4) (2002) 282-294

- [Lavrenko 2003] Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In Thrun, S., Saul, L., Schölkopf, B., eds.: *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA (2003)
- [Lopatovska 2008] Lopatovska, I., Mokros, B.: Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. *Information Processing and Management: An International Journal* 44(1) (2008) 92-104
- [Lovasz 1993] L. Lovasz. *Random walks on graphs: A survey*. Combinatorics, Paul Erdos is Eighty, 2:353–398, 1993
- [Lowe 1999] Lowe, D. G., "Object recognition from local scale-invariant features", *International Conference on Computer Vision*, Corfu, Greece, September 1999.
- [Lowe 2004] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60 (2004) 91\_110
- [LSCOM 2006] LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006.
- [Lynne 2007] Lynne, M., Sheldrick, R.C., Oaulette, R. *Affective Dimensions of Information Seeking in the Context of Reading*. *Information Today* (2007)
- [Magalhaes 2007] Magalhaes, J., Røger, S.: Information-theoretic semantic multimedia indexing. In: *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, New York, NY, USA, ACM (2007) 619-626
- [Manjunath 2002] Manjunath, B.S., Salembier, P., Sikora, T.: *Introduction To Mpeg-7: Multimedia Content Description Interface*. John Wiley & Sons (2002)
- [Misra 2010] Misra, H., Hopfgartner, F., Goyal, A., Punitha, P., Jose, J.M.: *TV News Story Segmentation based on Semantic Coherence and Content Similarity*. In: *Proc. Multimedia Modeling* (2010)
- [Mooney 2006] Mooney, C., Scully, M., Jones, G.J., Smeaton, A.F.: *Investigating Biometric Response for Information Retrieval Application*. LNCS (2006)
- [Nahl 1996] Nahl, D., Tenopir, C.: Affective and cognitive searching behaviour of novice end-users of a full-text database. *Journal of the American Society for Information Science* 47(4) (1996) 276-286
- [Nahl 1998] Nahl, D.: Learning the internet and the structure of information behaviour. *Journal of the American Society for Information Science and Technology*, 49(11) (1998) 1017-1023
- [Nahl 2004] Nahl, D.: Measuring the affective information environment of web searchers. In *Proc. American Society for Information Science and Technology*, volume 41 (2004) 191-197
- [Nakazato 2003] Nakazato, M., Manola, L. and Huang, T.S. *ImageGrouper: A Group-Oriented User Interface for Content-Based Image Retrieval and Digital Image Arrangement*. *J. Vis. Lang. Comput.* 14, 363-386, (2003).
- [Naphade 2005] Naphade, M. R., Kennedy, L., Kender, J. R., Chang, S.-F., Smith, J. R., Over, P. and Hauptmann, A. (2005) *A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005*, IBM Research Technical Report, 2005.
- [Naphade 2006] Naphade, M., Smith, J.R., Tesic, J., Chang, J.-S., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J. *Large-Scale Ontology for Multimedia*. In *IEEE MultiMedia* 13(3), 2006, 86-91.
- [Nasios 2006] Nasios, N., Bors, A.: Variational learning for gaussian mixture models. *Systems, Man, and Cybernetics, Part B, IEEE Transactions on* 36(4) (Aug. 2006) 849-862
- [Ounis 2005] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. *Terrier Information Retrieval Platform*. In *Proceedings of the 27th European Conference on Information Retrieval (ECIR 05)*, Santiago de Compostela, Spain, 2005
- [Over 2007] Over, P., Awad, G., Kraaij, W., Smeaton, A. F., (2007) *Trecvid 2007 overview*. In: *TRECVID 2007 - Text REtrieval Conference TRECVID Workshop*.

- [Pantic 2000] Pantic, M., Rothkrantz, L. Expert System for Automatic Analysis of Facial Expression. *Image and Vision Computing Journal* 18(11) (2000) 881-905
- [Pantic 2005] Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective Multimodal Human-Computer Interaction. In *MM'05* (2005) 669-676
- [Pfister 2008] Pfister, H.R., Böhm, G. The multiplicity of emotions: A Framework of emotional functions in decision making. *Judgment and Decision Making* 3 (2008) 5-17
- [Picard 2002] Picard, R.W., Wexelblat, A., Clifford, C.: Future Interfaces: Social and Emotional. In: *CHI02 Extended Abstracts on Human Factors in Computing Systems* (2002) 698-699
- [Polarusa] <http://www.polarusa.com/>
- [Puolamäki 2005] Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., Kaski, S.: Combining Eye Movements and Collaborative Filtering For Proactive Information Retrieval. In *SIGIR'05* (2005) 146-153
- [Reeves 1996] Reeves, B., Nass, C. The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press (1996)
- [Rui 2000] Rui, Y., Huang, S.: Optimizing learning in image retrieval. In *IEEE Proc. of Conf. on Computer Vision*, volume 1 (2000) 236-243
- [Russell 1982] Russell, J.A., Steiger, J.H. The Structure in Person's Implicit Taxonomy of Emotions. *Journal of Research in Personality* (1982)
- [Russell 2003] Russell, A., Bachorowski, J.J., Fernandez-Dols, J. Facial and Vocal Expressions of Emotion. *Annual Review of Psychology* (2003)
- [Salojärvi 2005] Salojärvi, J., Puolamäki, K., Kaski, S.: Implicit Relevance Feedback from Eye Movements. In: *Artificial Neural Networks: Biological Inspirations* (2005)
- [Saracevic 1996] Saracevic, T.: Relevance considered. *Information science: Integration in perspectives. Proc. of the Second international conference on conceptions of Library and Information Science* (1996) 201-218
- [Scherer 2001] Scherer, K.R. Appraisal considered as a process of multi-level sequential checking. *Appraisal processes in Emotion: Theory, Methods, Research*. Oxford University Press (2001)
- [Scherer 2005] Scherer, K.R.: What are emotions? And how can they be measured? *Social Science Information* 44(4) (2005) 695-729
- [Smeaton 2007] Smeaton, A.F., Over, P., Kraaij, W. Evaluation campaigns and TRECVID. In *MIR06* (2006) 321-330
- [Smeaton 2009] Smeaton, A.F., Rothwell, S.: Biometric Responses to Music-rich segments in Films: The CDVPLEX. In: *CBMI* (2009) 162-168
- [Smeulders 2000] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A, and Jain, R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(12): 1349-1380 (2000)
- [Snoek 2006] Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M. (2006) The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *MM'06, ACM*, 421-430.
- [Snoek 2006b] Snoek, C., Worring, M., Koelma, D., and Smeulders, A. Learned Lexicon-Driven Interactive Video Retrieval. In *Proc CIVR 2006*, 11-20.
- [Soleymani 2008] Soleymani, M., Chanel, G., Kierkels, J.J., Pun, T.: Affective Ranking of Movie Scenes using Physiological Signals and Content Analysis. In *MS'08* (2008) 32-39
- [Sparck Jones 1975] Sparck Jones, K., and Van Rijsbergen, C.J., Report on the need for and provision of an ideal' information retrieval test collection, Computer Laboratory, University of Cambridge, 1975. Also available in the *Journal of Documentation*, 32:5975, 1976
- [Stathopoulos 2009] Vassilios Stathopoulos, Joemon M. Jose: Bayesian Mixture Hierarchies for Automatic Image Annotation. *ECIR 2009*: 138-149

- [Tait 2001] Tait, J., McDonald, S., Lai, T. (2001) CHROMA: An Experimental Image Retrieval System. *NDDL*, 141-151
- [Tenopir 2008] Tenopir, C., Wang, P., Zhang, Y., Simmons, B., Pollard, R. Academic users' interactions with ScienDirect in Search Tasks: Affective and cognitive behaviors. *Information Processing and Management: An International Journal* 44(1) (2008) 105-121
- [Urban 2006] J. Urban and J. M. Jose. Adaptive image retrieval using a graph model for semantic feature integration. In *Proc. of the 8th ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR'06)*. ACM, 2006.
- [Urban 2006b] Urban, J and Jose J.M. EGO: A Personalised Multimedia Management and Retrieval Tool. In the *International Journal of Intelligent Systems*, Wiley, Vol 21, Issue 7, 725-745, (2006).
- [Urban 2006c] Urban, Jana, Hilaire, Xavier, Hopfgartner, Frank, Villa, Robert, Jose, Joemon M., Chantamunee, Siripinyo, and Gotoh, Yoshihiko, Glasgow University at TRECVID 2006, TRECVID 2006 - Text REtrieval Conference TRECVID Workshop, NIST, pp. 363--367, 11 2006
- [Urban 2007] Urban, J., Jose, J.M.: Evaluating a workspace's usefulness for image retrieval. In: *Journal of Multimedia Systems* 12(4-5) (2007), 355-373
- [Urruty 2006] Thierry Urruty. Fatima Belkouch. Chabane Djeraba. "Efficient Indexing for High Dimensional Data: Applications to a Video Search Tool." 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM KDD 2006), Philadelphia, USA, 20-23 August 2006
- [Valenti 2007] Valenti, R., Sebe, N., Gevers, T.: Facial Expression recognition: A fully integrated approach. In: *Image Analysis and Processing Workshops (2007)* 125-130
- [van Ahn 2008] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D. And Blum, M. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, September 2008, pp 1465 – 1468. (2008)
- [van Rijsbergen-1979] van Rijsbergen, C. J. *Information Retrieval*. Butterworths, 2nd edition, 1979
- [Vasconcelos 1998] Vasconcelos, N., Lippman, A.: Learning mixture hierarchies. In: *Proceedings of Neural Information Processing Systems 11*, Cambridge, MA, USA, MIT Press (1998) 606-612
- [Vasconcelos 2000] Vasconcelos, N.: *Bayesian Models for Visual Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, June (2000)
- [Villa 2008] Villa, R., Gildea, N. and Jose, J.M. A Faceted Interface for Multimedia Search. In *Proceedings of ACM SIGIR 2009*, 775 – 776, (2008)
- [Vinciarelli 2009] Vinciarelli, A., Suditu, N., Pantic, M. Implicit Human-centred tagging. In: *Proc. Int. Conference on Multimedia and Expo (2009)*
- [von Ahn 2004] von Ahn, L. and Dabbish, L. (2004) Labeling Images with a Computer Game. *ACM Conference on Human Factors in Computing Systems, CHI 2004*. pp 319-326
- [Vrochidis 2008] Vrochidis, S., King, P., Makris, L., Moutzidou, A., Mezaris, V., Kompatsiaris, I.: MKLab interactive video retrieval system. In *CIVR'08 (2008)* 563-564
- [Wang 2000] Wang, P., Hawk, B., Tenopir, C. Users' interaction with world wide web resources: An exploratory study using a holistic approach. *Information Processing and Management: An international Journal*, 36(2) (2000) 229-251
- [Wilson 2000] Wilson, G.M., Sasse, M.A.: Listen to your heart rate: Counting the cost of Media Quality (2000) 9-20
- [Yavlinsky 2005] Yavlinsky, A., Schoeld, E.J., Røuger, S.: Automated image annotation using global features and robust nonparametric density estimation. In: *Proceedings of the 4th International Conference on Image and Video Retrieval (CIVR)*. Volume 3568 of *Lecture Notes in Computer Science (LNCS)*., Singapore, Springer-Verlag (July 2005) 507-517

## 10 Glossary

---

### Partner Acronyms

AM	Activa Multimedia, ES
BLITZ	Blitz Games, UK
DFT	Digital Film Technology, DE
DIT	Dublin Institute of Technology, IE
DLLNI	DTS Licensing Limited (Northern Ireland), UK
FBM-UPF	Fundació Universitat Pompeu Fabra, ES
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT
LFUI	Leopold-Franzenzs Universtät Innsbruck, AT
MTG-UPF	Music Technology Group, Universtat Pompeu Fabra, ES
PGP	Pepper's Ghost Productions Ltd., UK
TAIK	Taideteollinen Korkeakoulu, FI
UG	University of Glasgow, UK
URL – Funitec -	Universitat Ramon Llull, ES