Investigating Methods Of Annotating Lifelogs For Use In Search

A THESIS SUBMITTED TO
THE SCIENCE AND ENGINEERING FACULTY
OF QUEENSLAND UNIVERSITY OF TECHNOLOGY
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
HONOURS IN INFORMATION TECHNOLOGY

Harrisen Scells

Supervisor: Guido Zuccon

School of Electrical Engineering and Computer Science
Science and Engineering Faculty
Queensland University of Technology

November 2016

Acknowledgements

Thank you to my friends and family for the massive amount of support and encouragement over the past year – I could not have done this without it!

Most importantly I would like to thank my amazing supervisor Guido for the incredible opportunities and support he has provided me with – and for motivating me to pursue honours in the first place. It has been a challenging but amazing year of learning, and I am very grateful.

Abstract

The recent technological advances in and widespread marketing of quantified self devices is allowing more and more people to capture data about themselves. Lifelogging is an umbrella term which encompasses many personalised data gathering forms which primarily involve wearing a device that continuously captures images of every day events. These data capturing processes create vast amounts of personalised data about the user, however there is a lack of effective search methods that can accurately retrieve moments of interest for a user. The primary form of lifelogging is through the capture of images from wearable cameras. While image search has been the subject of extended research in the past and great advances have been made in both text-based image retrieval and content-based image retrieval, only limited work has considered searching lifelog image data, and the solutions currently available have not demonstrated effective performance. In this thesis four annotation methodologies are investigated to discover their performance in a text-based image retrieval system. These methodologies include textual descriptions, tags, relevance assessment, and queries. Annotations are investigated to understand which methodology is the best to enable effective text-based image retrieval on lifelog data and then automatic image captioning is investigated to determine if textual annotations can automatically be derived. Finally, each methodology is compared in terms of the cost to manually and automatically collect annotations.

The results of this research indicate that under ideal conditions, annotations which represent a query are the most effective in a retrieval task. The query annotation represents the text a user may enter into a typical search engine to retrieve an image they are looking for. This is compounded by the fact that on average they are the most cost effective annotation to collect of the four under investigation.

Copyright in Relation to This Thesis

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature:

Date:

Table of Contents

Chapter 1

Introduction

What if you had hundreds of thousands of images of your everyday activities collected through a lifelogging device and you wanted to search for meaningful events in this collection? Searching through large data sets of lifelogging images for insights about daily activities is a fast emerging area of research within the information retrieval domain [11]. A range of consumer products allow anyone to capture their daily activities and it is becoming increasingly popular [11][35][2]. These devices are typically sold under the umbrella term of 'lifelogging cameras'. Like precursors blogging and vlogging, lifelogging is the next step in this series of life event recording devices. Lifelogging not only benefits consumers who wish to document their lives, but also has applications for security and policing services where there is a need to search the images recorded through body cameras that are already being worn.

Preliminary research has provided an insight into how difficult searching lifelogging images is [28]. The text-based search of images is most commonly referred to as text-based image retrieval (TBIR). This research considers text based search solutions for lifelogging search. The alternative, content-based image retrieval (CBIR) does away with textual features and relies on the visual features of an image; for example colour, shape or texture, but generally assumes querying is through an example query image. In a study by Hartvedt [12], it was found that the benefit of TBIR over CIBR is that is supports image retrieval based on high-level textual semantic concepts. This claim is supported by Shao et.al [29] who notes that TBIR exploits semantic information in the text associated with images. CBIR is not applicable as images are searched with textual queries only and, as stated before, CBIR assumes querying occurs through example images. TBIR is the predominant approach for image retrieval and most commercial

search engines rely on this scheme [8]. TBIR intuitively relies heavily on quality annotations; if annotations have grammatical and spelling errors or do not correctly describe the content of the image (no semantic context) then the retrieval effectiveness is degraded. The requirements of a 'good' textual document in, for example, web search, also applies to annotations of images.

Currently, no machine learning approach can consistently and reliably generate concepts (or annotations) from lifelog images due to the difference between the training data and the data contained in lifelogs. There are multiple methods of annotating images automatically, however this research is primarily focused on investigating the performance of manual annotations. A number of annotation methodologies are tested to find the most appropriate type of annotation for text based image retrieval of lifelog images.

Collecting annotations is important, but another important challenge is to determine which annotation methodology is the best for text-based image retrieval. Evaluation of images generally relies on gold standard annotations, none of which exist for lifelog images. The problem becomes clear - the standard ways of evaluation are not applicable to lifelog annotations unless a reference exists. A framework involving TREC-style[1] runs is used for evaluating arbitrary types of annotations associated with lifelog images. The quality of the annotations is evaluated by means of using annotations to inform the image representation in a TBIR system for lifelog search. The pipeline for evaluating lifelog images and their annotations and the data produced is an important and meaningful contribution of the thesis.

Due to the sheer volume of annotations required to collect in the short time frame allowed by this thesis project and by the limited resources available for manual annotation, automatic image captioning is also investigated as part of this research. In this thesis, an image captioning system is trained on the manually collected annotations in order to annotate the entire collection of lifelog images. The collection of image is obtained form the the NTCIR-12 lifelog semantic access task (LSAT). The NTCIR-12 LSAT is a pilot task which has research teams retrieve and rank lifelog images using textual queries. The results from this research can be compared with the results from the NTCIR-12 LSAT to determine if the methods for image annotations investigated here improve over the performance of previous attempts.

[1] http://trec.nist.gov/

1.1 Aims and Objectives

Four annotation methodologies are chosen as a basis for this research. Each methodology involves the collection of annotations through some interface and evaluation. The aim of the project is to determine the best methodology for annotating lifelog images. The four methodologies under investigation are:

1. Tags — Sets of keywords that describe objects and semantics of an image. Tags are chosen from a user-defined, non persistent vocabulary. These are similar to what one would expect to see on Flickr and other similar sites.

2. Textual Descriptions — Descriptive long-form annotations that contain semantic meaning. These are similar to the contents of documents such as web pages or news papers.

3. Relevance Assessments — Images are scored on a 0-10 scale by how relevant the image relates to a given concept. These are more purposely obtained, similar to editorial efforts by commercial search engine companies.

4. Reverse Queries — Queries are formulated for a given search result listing. Annotators are asked to provide a query that they think would result in the image being returned by a typical search engine. These are more akin to what one would expect to see in query logs.

To collect these annotations, interfaces are developed to facilitate the entering and recording of the annotations. Once a suitable number of annotations have been collected, each methodology is evaluated. The evaluation strategies available are explored as part of this research.

This research also aims to produce two meaningful contributions to the lifelog research community. Recent results from the NTCIR-12 conference indicate lifelog search engines which incorporate image processing in some way perform much better than simply using annotations alone [26]. A new higher quality collection of annotations for lifelog images as well as training data for image classification systems is expected to boost the performance for everyone researching lifelog search engines: the annotations are released to the research community. This directs future efforts for collecting and evaluating annotations and can act as a comparison for future lifelog annotations.

Finally, automatic image captioning is investigated. Here, a state-of-the-art image captioning pipeline generates captions for images using the manually collected annotations as training data. The end result is a fully annotated collection of lifelog images. Previous attempts at using textual descriptions as annotations [28] resulted in poor, but optimistic results.

Of the four annotation methodologies selected for investigation, a recommendation is made on the most effective of these and the results of the automatically generated captions are compared to the results of the manually collected annotations.

1.2 Research Gaps

Two gaps are identified in the research through reviewing the literature. Annotation of lifelog images and their evaluation is not a standard process. Through preliminary research, a number of techniques have been investigated which are applicable to annotating images. A number of textual summarisation evaluation methods have also been considered, however none of these are capable of evaluation without a ground truth or gold standard. According to an examination of evaluation in information retrieval [27, p. 24], there has been very little work done to evaluate how good a test collection is. Furthermore, finding all relevant documents to build topics is still the accepted approach for creating a test collection [4].

It is currently unknown which annotation methodology most suits a lifelog images for a given TBIR system. This also implies that given a collection of lifelog images, it is unknown which form of annotation methodology for the images results in the best performance. Annotations (in general [31]) are very expensive to collect in both time and financially, thus determining the most appropriate methodology of annotation prior to collecting annotations is highly important.

1.3 Research Questions

This thesis aims to answer the following research questions:

Research Question 1: How can annotations for a collection of lifelog images be evaluated?

It is important that the annotations of images are accurate and of a high quality. Low quality annotations lead to poor evaluation results. Evaluation is performed without using a gold

standard annotation, since this does not exist. In this way, this research uses the ad-hoc TREC-style approach used in the NTCIR-12 lifelog semantic access task. This evaluation methodology is also referred to as extrinsic evaluation, whereby annotations are evaluated by determining the performance of a larger system as a whole with respect to each annotation methodology.

Research Question 2: What methods could be used to annotate lifelog images and what is the retrieval effectiveness of annotations collected with these methods?

There are many state of the art solutions for automatically summarising the contents of images [17][15][23], however this project involves measuring the effectiveness of manual annotations. This is due to the fact that training data specifically for lifelog images does not exit. This makes it challenging to train a model that can identify the contents of a lifelog image. In manually annotating images, a well formed collection of annotated lifelog images is produced and machine learning and computer vision algorithms can exploit this.

Research Question 3: RQ3: How do annotations generated by current state-of-the-art automatic image captioning techniques compare to the effectiveness of manual annotations?

The captions generated by machine learning are able to be evaluated in the same way the manual annotations are; the format of the annotations will not differ. The effectiveness of these automatic annotations provides a good comparison to the manually collected annotations and offers insight into how good the manual annotations are as training examples. Unlike the manual annotations, the automatically generated annotations are for the entire collection of lifelog images.

Chapter 2

Background

This chapter presents a background on the current state of searching lifelog images. This is done through a review of the literature and findings from the NTCIR-12 lifelog semantic access task.

2.1 Literature Review

2.1.1 Representing the Data

One way of viewing lifelogging is the process of creating a surrogate memory for a person. Organising and presenting are the key challenges for lifelog search engines. Gurrin et al. [11] proposes that it is possible to segment the raw, unprocessed lifelog data into meaningful units, or events, which he defines as: ”a temporally related sequence of lifelog data over a period of time with a defined beginning and end”. In order to perform information retrieval, the events need to be annotated with meaningful semantics. Annotations can either be manually created by humans, or generated through machine learning algorithms. These annotations must also be evaluated for effectiveness, as poor annotations will lead to poor performance when performing retrieval on the images.

There are five aspects of human memory access, as proposed by Gurrin et al. [11]:

1. Recollecting: Concerned with re-living and accessing past experiences of episodic memories.

2. Reminiscing: A form of recollecting, concerned with reliving past experiences for emotional or sentimental reasons.

3. Retrieving: A more specific form of recollecting in which specific information needs are to be retrieved such as an address, a document, a location, or any atomic piece of information.

4. Reflecting: A form of quantified-self analysis, performed in order to discover knowledge and insights that may not be immediately obvious.

5. Remembering: Concerned with prospective memory more than episodic memory. A form of planning for future activities or to act as a reminder or prompt for tasks that a person would like to do.

Gurrin et al. [11] argues that an information retrieval system targeted at lifelogging should focus on the Five R's as information needs for the user.

In the context of searching images on the web, L. Vuurpij, et al. [38] states that image retrieval systems are restricted to the domain they cover and require a lot of domain knowledge in order to fulfil the information needs of a user. Furthermore, L. Vuurpij, et al. [38] notes that there has been a shift from computer vision and pattern recognition to psychology and cognitive science in the domain of image retrieval, where models like the Five R's [11], are becoming more prevalent.

2.1.2 Annotating Lifelog Images

The current state of the art models for image retrieval use tag-based or textual description annotations [1]. This is typically due to the fact that retrieval models can use this text as a bag of words, or use the text to attribute some form of semantic meaning.

2.1.3 Evaluating Annotations

Without high quality annotations, semantic search would not work, since semantic search exploits the meaning and context of a sentence rather than the keywords in it [1]. The semantic and contextual data associated with images is important for an effective retrieval model that uses textual features rather than pixel data in images. This textual data can either be generated

using machine learning algorithms, as demonstrated by A. Karpathy et al. [17] and I. Sutskeve et al. [33], or generated manually by humans. The machine learning algorithms do however start from test data, typically of a specific domain or a range of domains, which they learn from. It is important to note that this can have undesirable consequences when trying to apply a model which has been trained on one domain to one which it has no knowledge of. In both situations, it is essential that the annotations themselves are evaluated such that they describe the image with enough detail and are convincing to humans, since queries will be formulated by humans. Three widely used models for this exist: BLEU[1] [24] which is precision based, ROUGE[2] [19] which is recall based, and METEOR[3] [7], used for judging the overall quality of annotations.

All of the metrics above were initially proposed with respect to the evaluation of automatic summarisation and natural language processing. Furthermore, they all use a reference annotation in order to score annotations. ROUGE compares the number of overlapping n-grams, word sequences, and word pairs of annotations with ideal annotations created by humans [19]. BLEU counts the maximum number of times a word appears in any reference annotation, followed by ”clipping” the total count of each candidate word by its maximum reference count, adding these ”clips” up, and dividing by the total unclipped number of candidate words in the annotation [24]. The notion of ”clipping” in BLEU is a variation of precision whereby words are only accepted for the maximum number of times they appear the reference text, for instance if a word appears in an annotation five times but is in the reference annotation twice, the ”clipped” value would be 2/5. Finally, METEOR generally operates by unigram matching (bag of words) between a reference annotation, typically created by a machine, and a human produced annotation. Both METEOR and ROUGE take multiple approaches to comparing annotations, for reference, ROUGE:

1. ROUGE-N N-gram Co-occurence Statistics

2. ROUGE-L Longest Common Subsequence

3. ROUGE-W Weighted Longest Common Subsequence

4. ROUGE-S Skip-Bigram Co-Occurence Statistics

[1] Bilingual Evaluation Understudy
[2] Recall-Oriented Understudy for Gisting Evaluation
[3] Metric for Evaluation of Translation with Explicit ORdering

5. ROUGE-SU Skip-Bigram Co-Occurence Statistics with Unigram Counting Unit

The unigrams matched in METEOR can be based on surface forms, stemmed forms, and meanings, with the option to be extended [7].

R. Vedantam et al. [36] argue through the results of their experimentation, however, that there exists a more effective model for evaluation which is rooted in human consensus. Their method, CIDEr (Concensus-based Image Description Evaluation) is a model which outperforms all other models of evaluating descriptive annotations of images. CIDEr performs so well due to high correlation with human judgement and consensus. According to R. Vedantam et al. [36], the CIDEr metric inherently captures sentence similarity, the notions of grammatically, salience, importance (precision), and accuracy (recall). CIDEr appears to improve upon the other three models and takes into account the weaknesses the other models may have, however it still relies on reference annotations.

The evaluation methods above rely on the availability of a ground truth or reference annotation that can be used to compare with the automatically generated annotation. This approach, however, is ill-suited to generating annotations for lifelog images as it is unclear what these annotations should "look like", because it is unknown what makes an annotation of a lifelog image "good". A better suited alternative for this problem is to embed the evaluation of different annotation methods within a task and thus evaluate the methods with respect to the effectiveness the different methods induce on the task. Specifically, in this research project, the aim is to embed the evaluation of lifelog annotation within a search task. Thus the effectiveness of a system would be evaluated with respect to the search task. None of the system properties would vary apart from the method that is used to annotate images.This evaluation methodology is akin to, for example, previous work that has examined the effectiveness and quality of different topic modelling techniques and semantic models via evaluating the effect they have on search engine result effectiveness [40, 47, 16, 45].

2.1.4 Types of Annotations

As suggested by R. Yan et al. [44], the most common image annotation approaches can be categorised into two types. The first is *tagging*, where annotators choose a set of keywords from

a vocabulary for each image. The second most common approach is described as *browsing*, where a group of images are judged against the relevance of a predefined keyword. There are, however, less commonly used annotation approaches, for example, *descriptive natural language annotations* which are generated in a model by A. Karpathy et al. [17]. This model outperforms the previous work done in this area of research for both image retrieval and image annotation on the Flickr8K, Flickr300K and MSCOCO. B. Hu et al. [13] clarifies why high quality textual descriptions generally perform better than systems that employ keyword or tag based annotation models, in that these models suffer from several limitations:

1. A keyword in a document does not necessarily mean that the document is relevant
2. A relevant document may not contain the explicit word
3. Synonyms of the query keywords lower the recall rate (ratio of retrieved images which are relevant to the total number of relevant images, see Appendix A for details)
4. Homonyms of the query keywords lower the precision rate (ratio of relevant images that are successfully retrieved to the total number of relevant and irrelevant images retrieved) see Appendix A for details)
5. Semantic relations such as hyponymy, meronymy, antonymy are not exploited

Recent work in the consumer health search domain by Zuccon et al. [48] and Stanton et al. [32] focused on generating queries from images. The aim of their research was to understand how the general public would search for information if they had a medical condition as that in the image presented to them. This new methodology used by these previous works could be adapted to the context of gathering annotations for lifelogging, thus leading to an *annotation by querying* method. This method would consist of showing annotators an image from a lifelogger and ask them to provide the queries they would issue to a (standard) search engine to attempt to retrieve the image itself.

2.1.5 Searching for Lifelog Images

Lifelog information retrieval systems typically have very poor performance due to there not being any formal models made specifically for the field, as reported by Gurrin et al. [11]. Until very recently, there have been no large, distributable test collections such as the TREC collection

for text [11]. The NTCIR collection is a set of tagged lifelog images which have been collected by researchers who wore a lifelogging camera for a short amount of time [10]. The tags were automatically generated by using a pre-trained image tagging algorithm.

While there is limited applied methodology to retrieval models in lifelogging, there has been much discussion about what the models should try and solve. Both H. W. Bristow et al. [3] and A. R. Doherty et al. [6] corroborate that detecting and interpreting the implicit semantics and context of lifelogging data from heterogeneous sources would be advantageous in explaining the Who?, What?, Where? and When? questions which occur in every day events. It was also noted that these questions are common among image searchers and that they are not capable of being answered by normal indexing like that in traditional search engines [1].

While there has been some research into tagging and annotating images, there has not been as much work in developing a model for searching these images within the context of lifelogging [11]. Typical image search engines for web pages treat the surrounding text, captions, alternate text and HTML titles [9] as a bag of words for retrieval. The success of these search engines rely on a sufficient amount of surrounding text, something which is not provided by current automatic image annotation models for lifelogging. The longer and more detailed the text is within the context of the image, the better the performance of the search engine. This is perhaps why other research has involved novel search techniques [38], since the current models for generating captions of images are not yet detailed or accurate enough for current textual information retrieval models to work.

Generally, image based retrieval methods can be classified into two categories: text-based image retrieval (TBIR) and content-based image retrieval (CBIR). A CBIR system utilises image features such as grid colour movements, edge direction histogram, Gabor textual features, and Local binary pattern histograms. as described in work by Wu et al.[41]. These features (colour, texture, shape, SIFT keypoints) become a query to the search engine which match visually similar images. CBIR systems, although extensively studied for over a decade, are still limited in comparison to TBIR systems. Zue et al.[46] provide three points for why this is:

1. The semantic gap that exists between low-level visual features and high-level semantic concepts

2. The low efficiency due to high dimensionality of feature vectors

3. The query form is unnatural for image searching (appropriate example images may be absent)

The efficiency of TBIR can be explained when one considers that it can be formulated as a document retrieval problem and can be implemented using the inverted index technique. The downside to TBIR is that is highly expensive: experimental evidence by Wu et al. [42] shows that the performance of TBIR is highly dependent on the availability and quality of manual annotations. If this process can be automated and images can be automatically captioned, it would solve a fundamental issue that exists with TBIR systems.

2.1.6 Automatically Captioning Images

There have been some recent advances in machine learning which combine convolutional neural networks and recurrent neural networks that enable images to be automatically captioned. In this thesis neuraltalk2 [17] is used for automatically captioning images. Neuraltalk2 works by feeding the last hidden layer of the convolutional neural network (CNN) as input into the recurrent neural network (RNN). This recipe is followed by other current state of the art automatic image such as Jia et al. [15] (CAFFE[4]), and Vinyals et al. [37].

These neural network approaches have reported to perform better than previous systems. Older systems attempt to manually select features and perform some clustering or probabilistic technique. Some examples of these previous works include a subspace clustering algorithm by Wang et al. [39] and a probabilistic approach that involves graphs by Pan et al. [23]. Wang et al. identifies the core problem with manually selecting features: that image data is highly dimensional, and many dimensions are irrelevant which confuse these older systems (i.e. hiding clusters in noisy data). This is why neural networks and deep learning excel at captioning — they learn the features and how to optimally segment images.

Generating captions for images reduces the time cost of TBIR systems, although manually annotating or labelling a test collection for training data still takes time and is prone to human error. This process may be able to be alleviated by automatically generating images from text. Recent machine learning architectures like that of Reed et al. [25], while limited to specific domains, can produce images from textual descriptions. In time this could allow hybrid

[4]Convolutional Architecture for Fast Feature Embedding

TBIR/CBIR systems that outperform the current state of the art image retrieval systems.

Attempts have been made at image retrieval that exploits both TBIR and CBIR methodologies. One such attempt by Escalante et al. [8] introduces two novel formulations: annotation based expansion (ABE) and late fusion of heterogeneous methods (LFHM). In ABE, segmented regions in images are labelled. An annotation is formed by expanding the labels into a textual representation. The annotation associated with an image is treated like a document and typical text-based retrieval is then performed. LFHM consists of manually building several retrieval models based on different information from the same collection of documents. Each retrieval model returns a list of relevant documents to a query and the output of each is combined to obtain a single list of ranked documents.

2.2 Findings From NTCIR

The NTCIR-12 lifelogging latent semantic access pilot task consists of four research teams contributing to the automatic retrieval component and one participant in the interactive retrieval component. The highest performing automatic team, LIG-MRIM, uses computer vision to classify images and does not rely on the visual concepts distributed with the task. The interactive team (LEMoRe) outperformed all other automatic teams, but this is generally the case with tasks that contain both automatic and interactive components. The three other automatic teams are consisted of VTIR, III&CYUT and QUT (the preliminary work done for this research).

LIG-MRIM [26] uses dynamic convolutional neural networks and a multi-class support vector machine (MSVM) in order to classify images. Visual indexing is composed of two parts: three deep convolutional neural network models (AlexNet, GoogleNet and Visual Geometry Grouping (VGG)) process each image. The output is normalised and has principal component analysis (PCA) performed on it. These outputs are then concatenated together. The same normalisation and PCA process is repeated and fed into the MSVM. The output of this is concatenated with the VGG data. The second part involves temporally naming times of the day in order to attempt to extract semantic meaning from times of the day. While this team submitted runs that fit the definition of automatic for the task, queries were generated manually from the topics by an expert.

The VTIR team [43] attempts to exploit location meta data associated with the images. To

this end 3,000 random images are labelled against a rich semantic location ontology. More concepts are utilised by applying the WordNet database to find cognitive synonyms. Despite the additional annotations, this system failed to provide good retrieval effectiveness.

III&CYUT [20] uses a traditional textual based approach to lifelog retrieval. A skipgram word embedding obtained with word2vec[22] is computed for the visual concepts distributed with the data set. These embeddings are then used in an attempt to add more semantic meaning to images. Specifically, the embeddings are used within a document expansion process, resulting in a translation language model [47]. Query expansion is also used on every keyword.

The QUT team [28] manually annotates a subset of the images in the collection with long textual descriptions. To select images for annotation, they follow an approach based on temporal and visual clustering of the images. They then further extend the annotation process by propagating the annotations to other images contained within the same cluster of already annotated images.

Finally, the LEMoRe [5] team, which use an interactive approach, combine existing technologies and methodologies in order to develop a search engine. Colour correlogram, edge histogram, joint composite descriptor and pyramid histogram of oriented graphics are used by the image retrieval system as features to retrieve images. Both a novice and an expert used the system to produce runs.

The results from this pilot task offer a promising glimpse into the future of searching lifelog images. Figure 2.1 presents the best run from each of the teams that participated in NTCIR-12 LSAT. LIG-MRIM shows that automatic methods for annotating lifelog images can result in decent text-based image retrieval effectiveness. The teams that do not perform well (including the QUT team) offer insight into areas of research to avoid.

The difference in retrieval effectiveness between LIG-MRIM and the other three teams is highly likely due to the annotations of the images. The task provides teams with automatically generated annotations for the images distributed with the task. These annotations are generated using a previously state-of-the-art captioning framework, CAFFE [15]. The problem lies within the fact that a CAFFE model is trained on a data set that does not align with the lifelog images. Three of the teams use these annotations in their systems; however LIG-MRIM generate their own annotations. This indicates that no matter how well tuned and suited a text-based image retrieval system is to lifelog images, poor textual representations for images can significantly

impact the retrieval effectiveness.

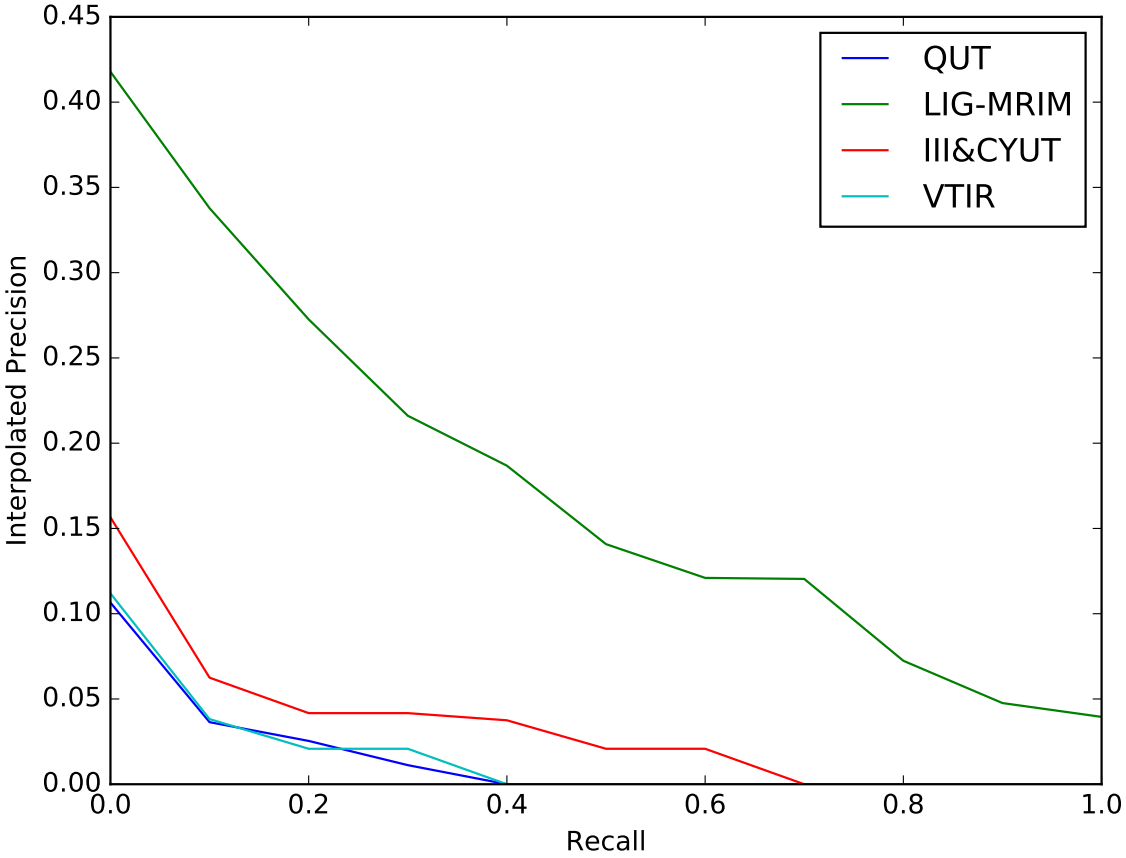


Figure 2.1: Precision-recall curves for the four LSAT teams

Chapter 3

Methods

The methodology used in this thesis involves three steps: sampling images to annotate, annotating images (manually or automatically), and evaluating annotations. This chapter describes each of these steps in detail.

3.1 Sampling Images

It is not feasible for every image in the NTCIR-12 LSAT data set to be annotated manually. The collection is very large, containing 88,125 images. Moreover it is even more unfeasible to annotate every image four times (for each annotation methodology), thus sampling images to reduce the number of images to annotate is necessary. There are two methodologies employed to sample images: The first builds upon previous work [28] which identifies a way of clustering lifelog images using image histograms for features; and aligning the images temporally to determine cluster segmentation boundaries[1]. After clustering images, the sampling technique involves selecting one image at random from each cluster. The cosine similarity measure is used to determine if an image should be added to an existing cluster. The threshold value of the cosine between two images is set to 0.86[2]. Clusters are then combined based on visual similarity using the aforementioned image histograms and a representative image from each of these clusters is chosen at random. This process results in about 16,000 images for annotation. The second sampling method entails processing a file[3] containing known relevant images distributed after

[1] `https://github.com/hscells/lifelog-sampling`

[2] This value was empirically found to provide a range of sufficiently different clusters, each containing similar images

[3] The qrels file, this was not distributed as part of the task, but several months later

the NTCIR-12 LSAT to extract only the relevant images. A maximum of 30 images are selected from each of the 48 topics which results in just over 1,000 images required for annotation. Some topics have less than 30 relevant images which is why the total number of images is lower than what one would expect.

There is some overlap between the images chosen from the clustering process and the images that are known to be relevant, however it reduces the overall number of images to annotate by around 80%. In reality, it is not expected that all of these images are annotated, rather, the smaller sample set should provide a reasonable interpolation of the larger data set. Annotating both relevant and non-relevant images ensures that retrieval is working correctly: both relevant and non-relevant images should be retrieved, however the images annotated with relevant annotations should rank higher.

The sampled images are uploaded to a database to be annotated. Annotations are then collected through specially designed interfaces.

3.2 Collecting Annotations

Once the images have been sampled and uploaded into a database, they are ready to be annotated. Four web interfaces are used to collect of annotations[4]. The interfaces allow experts to annotate images selected randomly from a list of unannotated images. It is important to note that each image is annotated only once and an attempt is made to ensure that there is an overlap between the images and the annotations such that most images annotated should have four annotations.

The architecture of the interfaces consists of:

1. A database to store the annotations and the sampled images. The database also records statistical data about each user performing the annotations: who annotated each image, and how long it takes to perform an annotation. This ensures there is a record of who annotated each image, and allows a statistical analysis to be performed at a later stage.

2. A web server and that handles the ‘business logic’. This server exposes some password protected RESTful services that applications can hook into.

[4] `https://github.com/hscells/lifelog-ia`

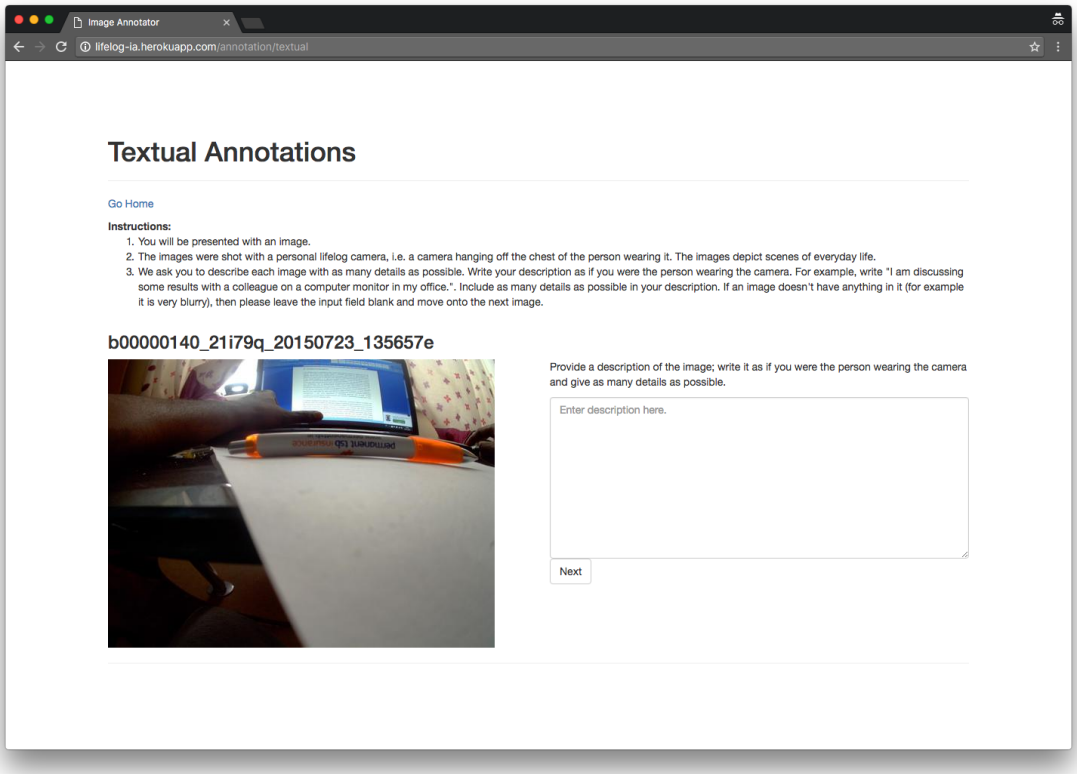
3. A website which consumes the API provided by the web server and handles 'view logic'. This is the layer that annotators interact with directly. Each interface is one of these views.

The interfaces are designed with great care to ensure collecting annotations is as painless as possible. It is important that the total time it takes to annotate an image (including the time between annotating images, i.e. loading an image) is as small as possible; If it takes a minute to annotate one image, it will take an hour to annotate only 60 images.

3.3 Annotation Methodologies

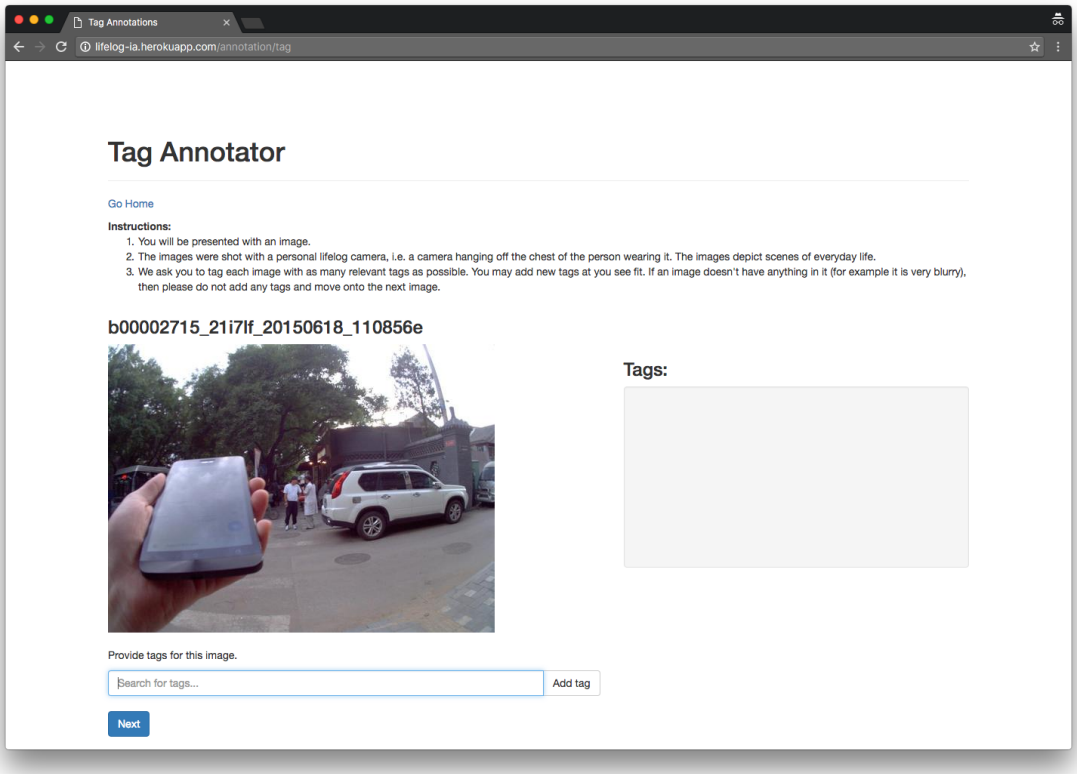
Four annotation methodologies are selected for investigation: **textual**, **tags**, **relevance assessment** and **reverse query**. While the annotation types are wildly different to each other, they are all collected in a very similar way. An expert annotator is shown a randomly chosen image from the sampled set of images and is asked to provide an annotation (or in the case of relevance assessment, multiple annotations) for the image. The interfaces used for collecting annotations are pictured and described as follows:

Textual



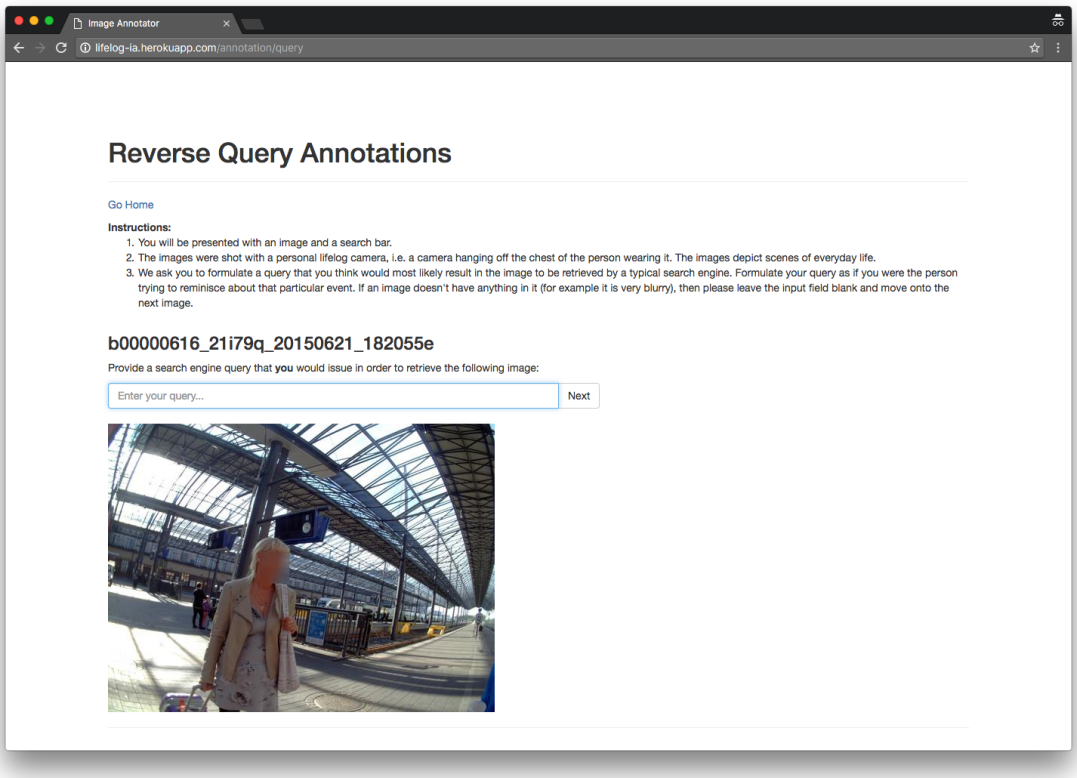
Here, annotations are collected using free form text through a text box on the page. Each textual annotation should contain semantic information about an image, and describe the image with a high level of detail. These annotations are very similar to a textual document in a typical web search engine, which is why they are selected as one of the methodologies to investigate. Information retrieval is commonly associated with textual documents which contain many sentences with varied semantic and contextual information. This type of annotation highly resembles typical document retrieval and can almost be seen as a baseline, whereas the other annotation methodologies are somewhat novel.

Tags



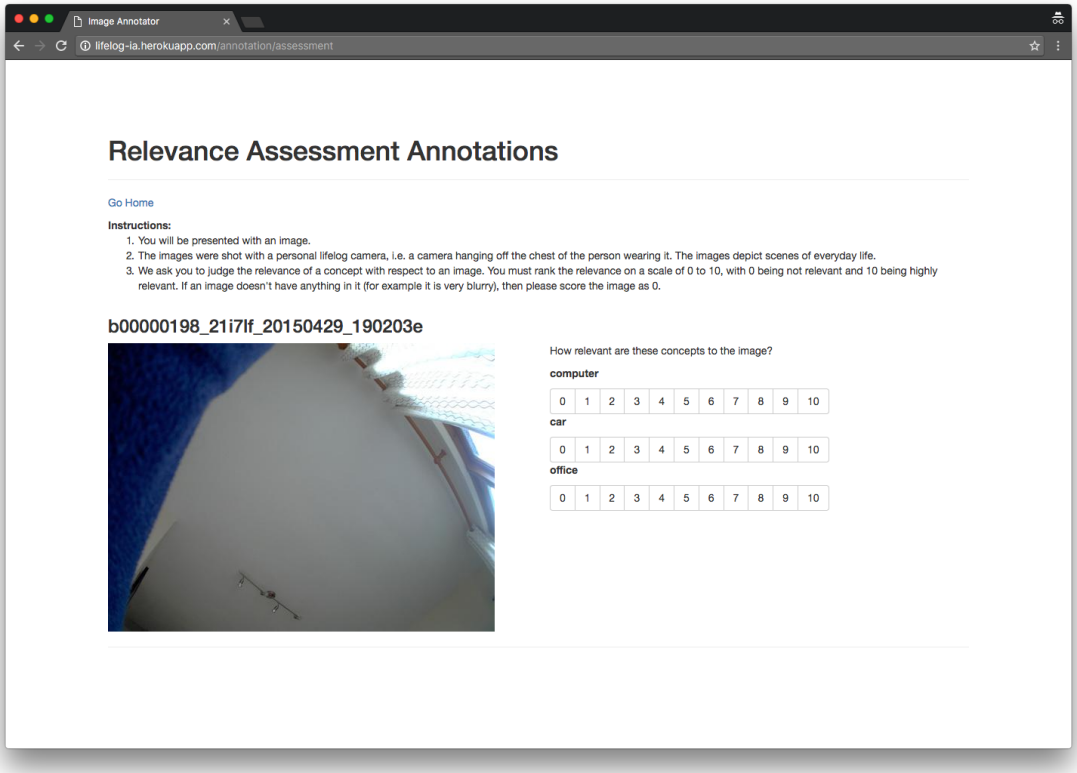
Tags are collected through this specifically designed interface. The vocabulary of tags is created from previously added tags, the list of tags available is arbitrary and can be expanded. This annotation methodology and style of collection is similar to that seen in other online image tagging scenarios such as Flickr. Similarly, the vocabulary is not restricted to a preset group of terms. This unrestricted vocabulary can result in human error (spelling mistakes), but allows for precise observations about key objects or events in an image. The tags that annotators are allowed to input are allowed to contain more than one word, to cover concepts like ‘train station’ or ‘shopping mall’.

Reverse Query



In this interface, user queries are collected by presenting an image taken from the lifelog camera and asking the annotator to provide a query with what they expect to be returned by a typical search engine. This is a novel way of annotating *any* type of document or image [48]. This form of annotation represents the query logs from a search engine; search engine companies have this data but generally do not make this available due to privacy concerns [30]. These annotations are less focused on detail and more focused on one to two key objects or events in the image.

Relevance Assessment



Relevance assessment involves presenting an annotator with an image, and asking them to judge how relevant a concept is to the image. Assessors are asked to choose concepts from a list to assess, and from those chosen concepts are asked to assess the relevance of that concept to the target image. Concepts are ranked on a scale of zero to ten, where zero is not relevant at all and ten is highly relevant.

These annotations are similar to the tag annotations in that annotators are annotating with some notion of a 'concept'. The major differences between the two annotation methodologies is that the vocabulary is finite and the relevance of a concept to the target image is not binary.

Relevance assessment annotations are collected last, the list of concepts is formed from analysing terms in the other annotations. Concepts are chosen by creating a list of terms from the existing textual, tag and query annotations, which are then filtered to the terms that occur in the NTCIR-12 LSAT topic titles and descriptions. Each term is scored using IDF[5] and then ranked using an algorithm similar to discounted cumulative gain [14]; in which higher scoring terms are selected less frequently, and more emphasis is placed on selecting lower scoring terms.

[5]Inverse Document Frequency

Automatic Image Annotation

Finally, an attempt to automatically caption images for each annotation methodology is done by utilising a recent state-of-the-art machine learning image captioning approach [17] which has been open sourced[6]. The architecture of neuraltalk2 consists of (1) a convolutional neural network (CNN) which learns features of image regions, and (2) A recurrent neural network (RNN) that generates textual descriptions using a long short-term memory (LSTM) network. The last layer of the CNN is fed as input into the RNN.

A model is trained in the system described above using the $adam$ optimiser [18] with α set to 0.8 and β set to 0.999. The $adam$ optimiser algorithm is for first-order gradient-based optimisation of stochastic objective functions (i.e convergence in a neural network). The parameters α and β represent the step size and exponential decay rates for the moment estimates respectively. The learning rate of the language model is set to 0.0004. Two training attempts are performed: one where the model is trained for a maximum of 70,000 iterations, and another where the model is trained for a maximum of 300,000 iterations. More iterations are performed afterwards to fine tune the deep learning architecture, but in all cases the changes to values of the optimisation function are insignificant.

3.4 Evaluating Annotations

Annotations are evaluated with an ad-hoc, TREC style methodology. The topics and qrels from the NTCIR-12 Lifelog semantic access pilot task are used to perform evaluation. In total eight runs are produced: five consisting of each of the manually annotated annotations plus the combination of these, and another four consisting of automatically generated captions for textual, tag, and query annotations as well as the combination of these. This is to investigate if any correlations between the manually annotated annotations and the automatic annotations exist, and to determine if an automatic system can effectively generate suitable annotations for lifelog images. The relevance assessment annotations cannot be learnt by the neural network framework used, due to the format of the annotations.

Document rankings are generated by submitting queries to ElasticSearch using porter stemmer and a default English stoplist. Queries are formulated by using the title and description

[6] https://github.com/karpathy/neuraltalk2

fields from the NTCIR-12 Lifelog topics. The stemming and stopping are also applied to the queries. ElasticSearch is also set to only retrieve a maximum of 1,000 images. The runs produced by this system are evaluated using `trec_eval` and the NTCIR-12 Lifelog qrels. The retrieval model for producing runs is $tf*idf$ and the method is the default `simple query string`[7] query.

The document rankings and evaluation is performed through a custom-built framework[8]. A Java RESTful application wraps Elasticsearch. This is so a query issued to the Java application can return a properly formatted TREC run, and so many queries can be issued with one request. Rather than the system producing a search listing, it outputs in a format ready for `trec_eval` to process.

[7] `https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html`

[8] `https://github.com/hscells/lifelog-eval`

Chapter 4

Results

The findings of the research is presented in this chapter in the form of the annotation statistics and the retrieval effectiveness of the annotations. The annotation statistics cover the time taken to annotate using each annotation methodology, and details about the number of annotations collected. The retrieval effectiveness section provides a breakdown of the performance of both the manually and automatically collected annotations.

4.1 Annotation Statistics

In total, five annotators managed to annotate a total of 10,982 images across all annotation methodologies. These annotators consist of students and staff from an information retrieval group at QUT. Figure 4.1 illustrates the number of annotations completed by each annotator. Two annotators account for the majority of the annotations, while three others provide an additional 904. The exact number of each annotation type, the total time it took to annotate each annotation type and the average time it took to annotate is displayed in Table 4.1.

A statistical analysis of the collected annotations reveals that they are appropriate for a

Name	Count	Average Time	Total Time
Text	3172	1 minute	2 days, 23 hours
Tag	2897	31 seconds	23 hours, 40 minutes
Query	3616	16 seconds	15 hours, 10 minutes
Assessment	1327	58 seconds	21 hours, 36 minutes

Table 4.1: Annotation statistics obtained by taking the average across all five annotators

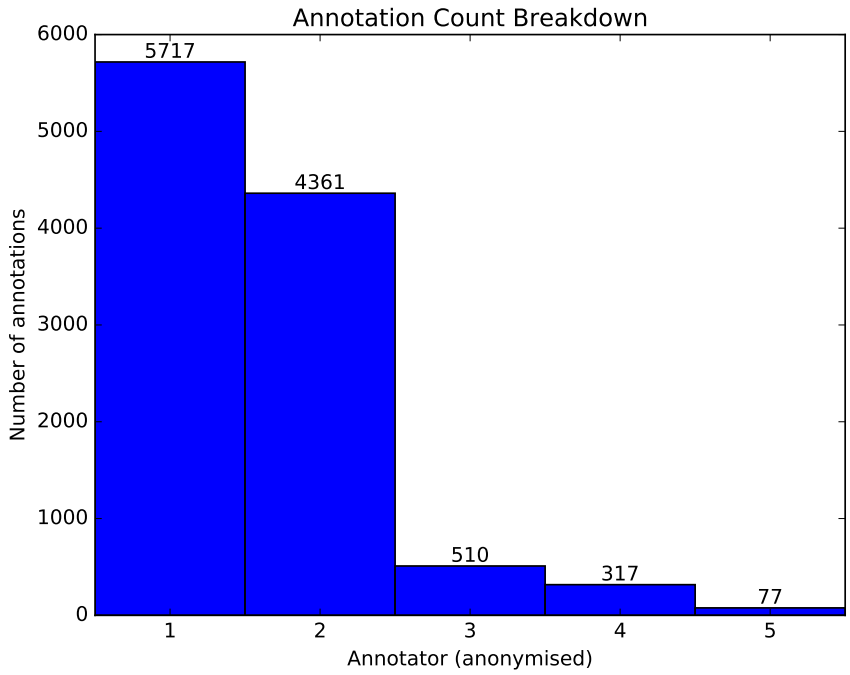


Figure 4.1: Total number of annotations by annotator

typical textual corpus. Figures 4.2 and 4.3 are what one would expect to see in a Zipfian distribution [34]; that is the frequency of each concept is inversely proportional to it's rank in the frequency distribution (the most commonly used concept appears twice as often as the second most frequent concept and three times as often as the third most frequent concept). The notion of concepts are different to terms since a concept can contain more than one word (this is possible through tags, where a tag can contain multiple words such as 'shopping mall' and 'street sign').

Textual annotations accounted for the highest amount of time taken on average in the collection process. The largest number of annotations collected for a methodology are the query annotations. Qualitative feedback from annotators note that the relevance assessments are the most tedious to collect. Intuitively, formulating a query for a typical search engine does not take a very long time, which can account for the marginal average time for this annotation type. On the other hand, composing (in the annotators mind and physically typing on a keyboard) a descriptive paragraph filled with context and semantics leads to the conclusion that textual annotations do, in fact, take a significant amount of time to annotate. In the same manner, the process of completing a relevance assessment involves scrolling and clicking a multitude of times; this time adds up and is evident in the reported average time.

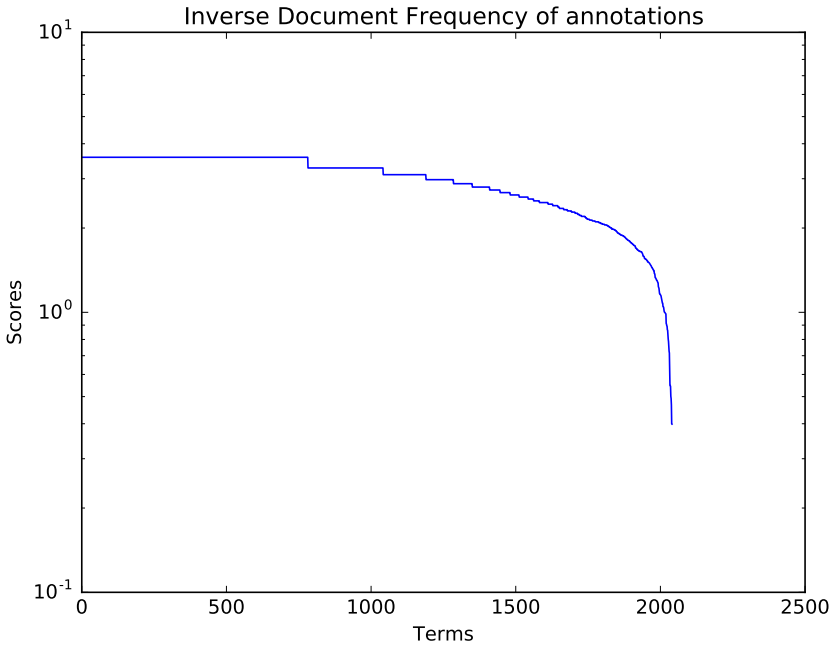


Figure 4.2: IDF scores for concepts in the annotations

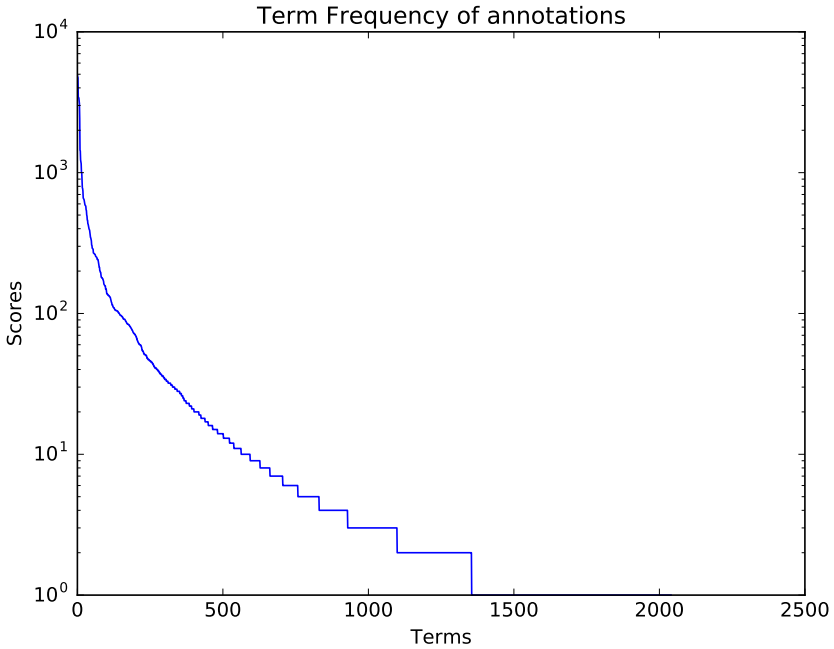


Figure 4.3: Term Frequency scores for concepts in the annotations

4.2 Retrieval Effectiveness

Results of each experiment are reported as a table which provides the results from `trec_eval` and as a precision-recall graph. The results from only the manually annotated images are displayed first. Table 4.2 presents the `trec_eval` results for the four annotation methodologies *and* the result of combining all four of the methodologies. The NTCIR-12 LSAT topics are used for experiments. Among other fields, each topic contains a title and a description; these fields are read as input queries. The title field is more representative of what a typical query looks like when issued by a user. The results of running these experiments are visualised as precision-recall graphs in Figures 4.4 and 4.5.

In experiments that use the title field as an input the query annotations perform the best of the four methodologies, however when combining all four collections of annotations together the effectiveness increases slightly, and the precision is higher than the query annotations at high recall (more relevant results). Combining annotations for the experiments that use the description field outperform all four of the methodologies, and the query annotations perform significantly worse.

Topics Titles

Methodology	MAP	RR	P@10	Relevant Retrieved
Text	0.3442^{qc}	0.9223^{c}	0.8333^{rc}	1062
Tag	0.5468^{qcd}	0.9578^{d}	0.8396^{rcd}	1040
Query	0.6400^{tgad}	0.9653^{rd}	0.8750^{rd}	1559
Relevance Assessment	0.4815^{qc}	0.8406^{qc}	0.6917^{tgqrd}	831
Combined	0.6495^{tgrc}	1.000^{ta}	0.9062^{tgr}	1612

Topic Descriptions

Methodology	MAP	RR	P@10	Relevant Retrieved
Text	0.5285^{gqc}	0.9792^{gq}	0.8521^{gqc}	1088
Tag	0.4627^{trcd}	0.8928^{tqcd}	0.7687^{tqcd}	1055
Query	0.4457^{tcd}	0.7683^{tgrcd}	0.5958^{tgrcd}	1566
Relevance Assessment	0.5219^{c}	0.9271^{q}	0.7938^{tqcd}	972
Combined	0.5855^{tgqrc}	0.9815^{gq}	0.8875^{tgqr}	1613

Table 4.2: MAP, Reciprocal Rank, Precision at 10 scores, and number of relevant retrieved images for the manual annotations. Two tails t-test statistical significance ($p < 0.05$) is indicated through the following labels for inter-measurement: text ^{t}, tags ^{g}, query ^{q}, relevance assessment ^{r}, combined ^{c}. Statistical significance between the same methodology is represented as ^{d}.

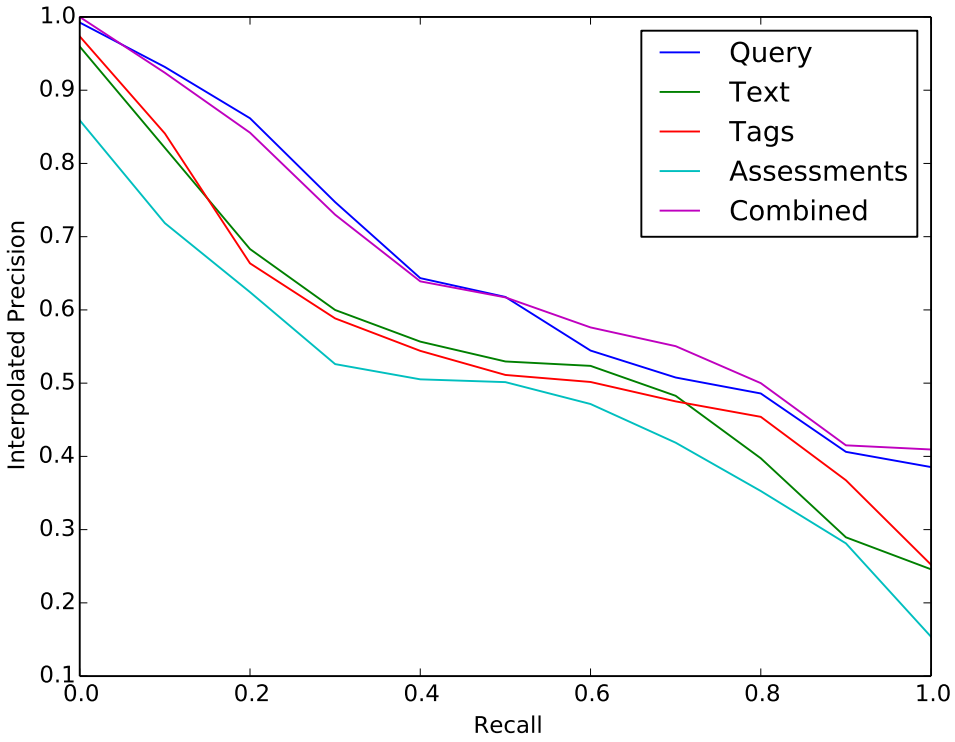


Figure 4.4: Precision-recall curves for the manual annotations using topic titles

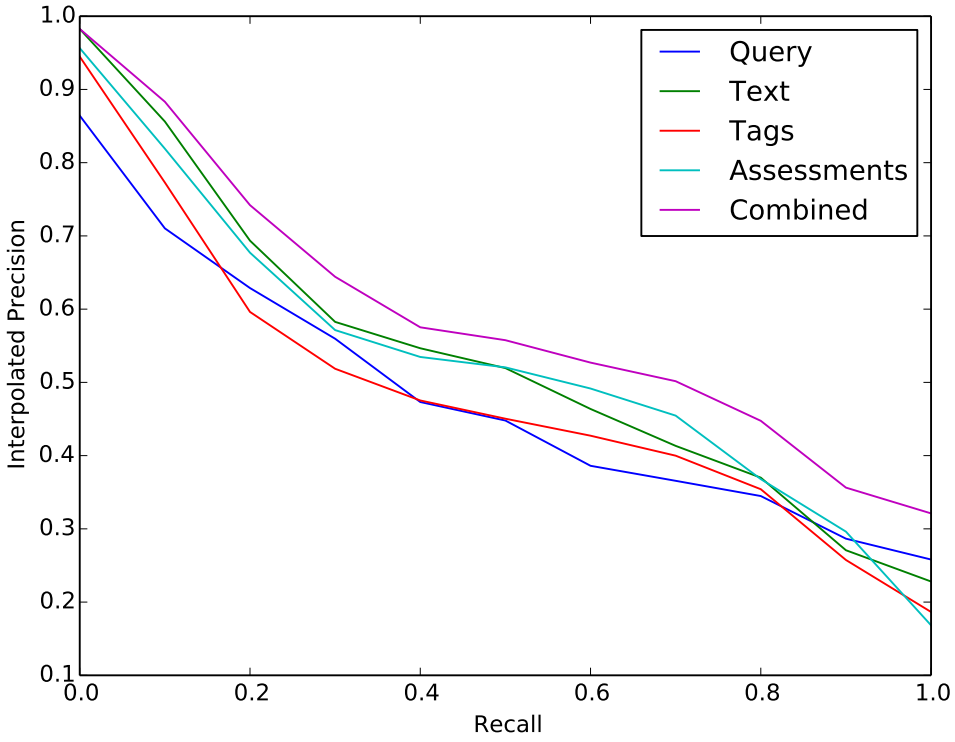


Figure 4.5: Precision-recall curves for the manual annotations using topic descriptions

A neural network framework (neuraltalk2) is trained on the manual textual, tag, and query annotations (i.e. after the manual annotations are collected). Neuraltalk2 is able to produce captions for every image in the collection. The quality of these captions is summarised in Table 4.3 – an unfortunate result which could be attributed to the amount of training data. The automatic captions that are generated are evaluated in the same way as the manual annotations. The output of the neural network architecture is formatted to be used in the evaluation pipeline as described in Section 3.4.

There is no clear individual annotation methodology that outperforms the others, the scores are too low to indicate this. Combining the three automatic annotations together does seem to increase the overall precision. The learnt queries do retrieve the most number of images, in a similar result to the manual annotation results.

The number of iterations the neural network architecture covered for each annotation methodology is visualised in Figures 4.8 and 4.9. The results of the automatic captioning do not get better over time – in fact they get worse. The gap between topics with a large number of relevant images, topics with a low number of images, and the low number of training examples is presumably the contributing factor to the results in these figures.

Topic Titles

Methodology	MAP	RR	P@10	Relevant Retrieved
Text	0.0048	0.0248	0.0196	187
Tag	0.0083^{c}	0.0184	0.0022	136
Query	0.0076	0.0174	0.0021	246
Combined	0.0164^{g}	0.0393	0.0169	337

Topic Descriptions

Methodology	MAP	RR	P@10	Relevant Retrieved
Text	0.0048	0.0232	0.0167	222
Tag	0.0050	0.0184	0.0063	145
Query	0.0051	0.0062	0.0062	199
Combined	0.0096	0.0470	0.0187	325

Table 4.3: MAP, Reciprocal Rank, Precision at 10 scores, and number of relevant retrieved images for the automatically generated annotations. Two tails t-test statistical significance ($p < 0.05$) is indicated through the following labels for inter-measurement: text ^{t}, tags ^{g}, query ^{q}, relevance assessment ^{r}, combined ^{c}. Statistical significance between the same methodology is represented as ^{d}.

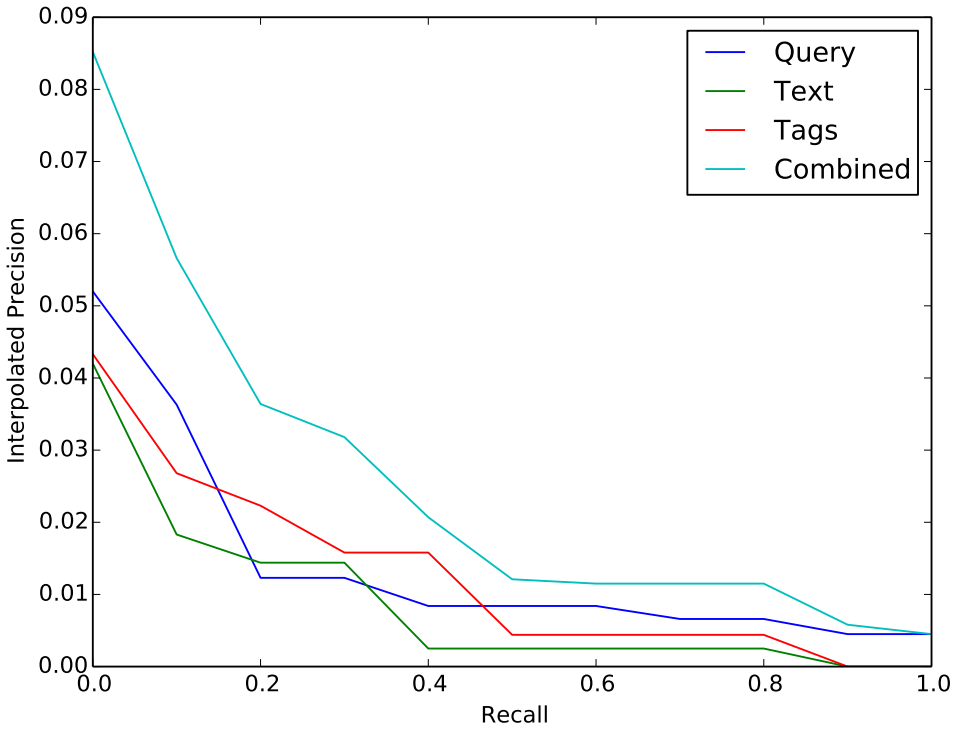


Figure 4.6: Precision-recall curves for the learnt annotations using titles

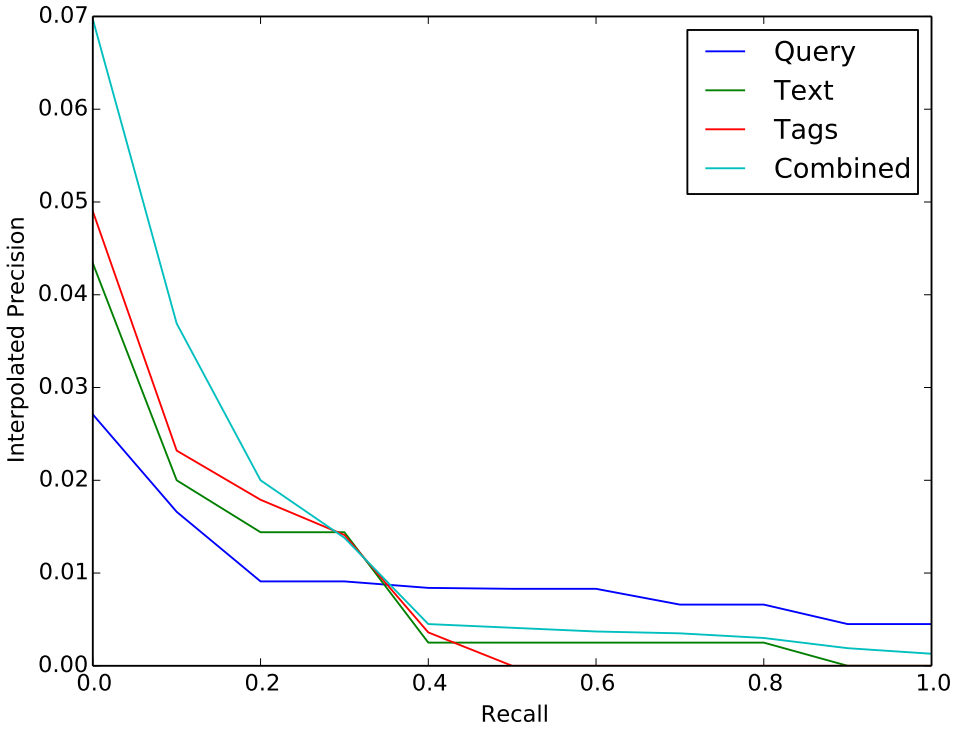


Figure 4.7: Precision-recall curves for the learnt annotations using descriptions

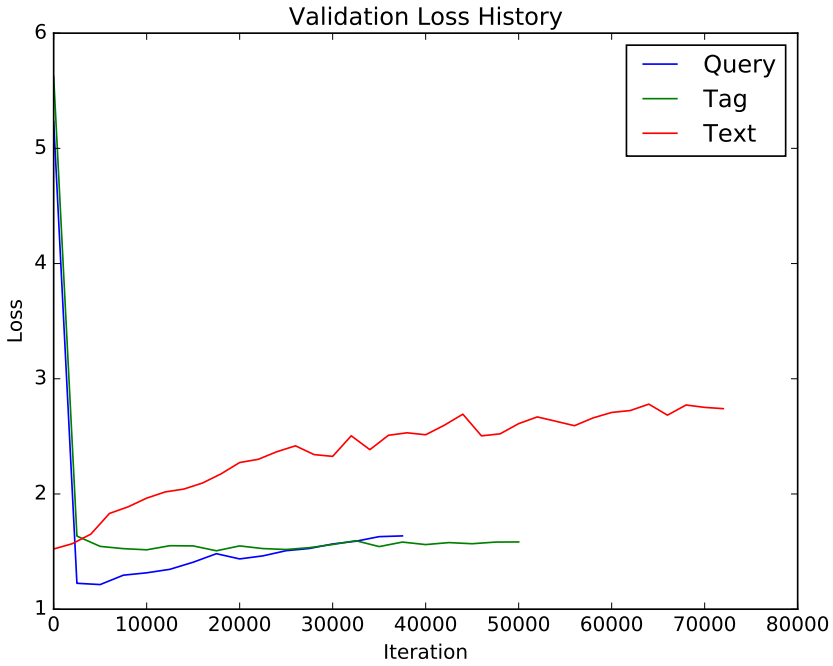


Figure 4.8: Validation loss history (<100,000 iterations)

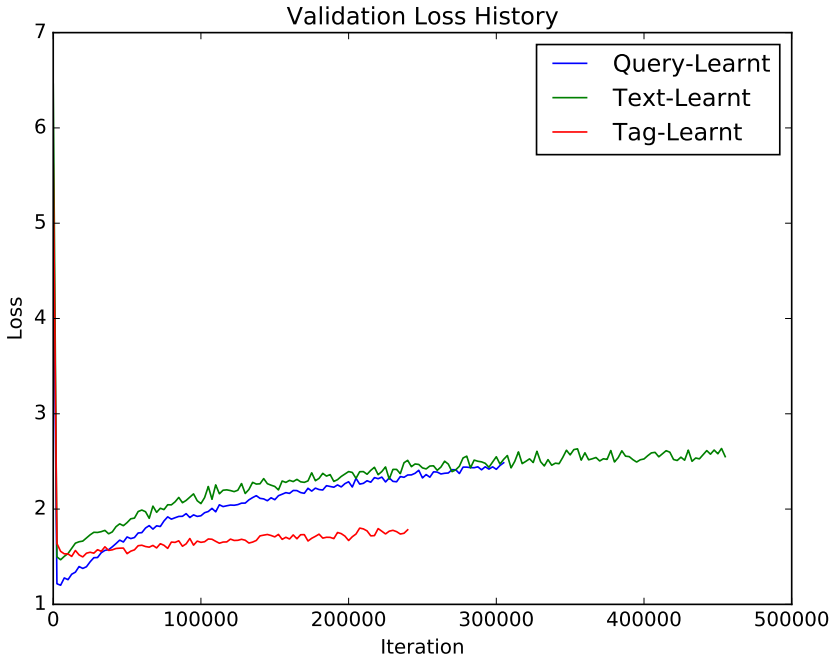


Figure 4.9: Validation loss history (>200,000 iterations)

Chapter 5

Discussion

This chapter discusses the effectiveness of the annotation methodologies, the automatic image annotation results, and the annotation interfaces. Each section aims to explain and summarise the results and evaluate the significance of the results.

5.1 Annotation Effectiveness

The results from these experiments outlined in Chapter 4 indicate that of all the types of annotation methodologies, the query annotations provide the best performance at the lowest cost (i.e. the time it took to annotate) in ideal circumstances. Utilising multiple annotations can increase the interpolated precision further towards total recall. In fact when retrieving as many images as possible, combining all of the annotations achieves the highest precision overall. The tag annotations perform badly in both experiments, since these annotations are unable to encode semantic meaning like text and queries. For instance, consider images that fall into the topic 'Building a Computer': the tag 'building' can have more than one meaning, a physical structure and the act of assembling an object. The tag annotations did poorly in this topic, whereas the textual annotations did the best – these contain semantics which the search engine can exploit when ranking images. When using the description as an input query, for some topics assessment annotations perform better. This is most likely due to the fact that the weights of annotated concepts allow the search engine to rank these relevant documents higher. The performance of the textual annotations can perhaps be attributed to spelling and grammatical errors; there is difficulty associated with ensuring these errors are corrected during a pre-processing step in

addition to blocking incorrect annotations from being entered in the first place.

In retrospective, the concepts for each query should be chosen manually based on the topics, but this would render a retrieval system unusable behind topics for which relevant concept annotations have not been defined a priori. When choosing concepts for use in relevance assessment, there is a large overlap between the titles and the descriptions of the topics and this was seen as good enough. Prior to analysing the results of the relevance assessment annotations, it was thought that they might perform only as well as the images retrieved using tag annotations. Similar to tags, the concepts of the relevance assessments do not contain semantic meaning; however unlike tags, each concept is assigned a weighting of how relevant it is to an image. In ascribing weights, the concepts assigned to images are contextualised; now 'building' is highly relevant *and* 'computer' is highly relevant, as opposed to another image that may have an equally high weight for 'building' but a high weight for 'architecture'. The image with the high 'computer' concept will be ranked higher because of the weighting; explaining why relevance assessments outperform the tags even though less images are retrieved. The trade-off in the end is that tagging takes half as long for slightly worse retrieval performance. An assumption that can be made for the relevance assessment annotations then is that if more concepts are added that are relevant to the topics, the retrieval effectiveness can be increased.

In the case of the query annotations performing the best out of the four annotation methodologies, one might assume that it is simply due to the fact that more images are annotated. The results of the relevance assessments indicate that this is not the case – the effectiveness of the query methodology can not solely be attributed to the number of images annotated. The relevance assessment annotations perform better (where the input query is taken from the description of the topic) than the query, tags, and in some cases the textual annotations, having the least number of images retrieved. It would seem then that the query annotations are highly suited to the experiments where the input query is composed of very few terms (much like a typical query). It is also possible that, being significantly shorter than the textual annotations, there is less error for spelling and grammatical mistakes. Moreover, the low cost of annotation time further makes the query annotation methodology the most attractive option for annotating a collection of lifelog images for use in text-based image retrieval.

5.2 Automatic Image Annotation

In total, there are 6,657 images that the NTCIR-12 LSAT organisers consider relevant from a collection of 88,125 images (only 7.5%). If images are annotated using the clustering method of sampling images, only 16,014 images (18% of the data set) could be annotated. Of that, 1,176 images are considered relevant. Sampling known relevant images is highly important – the clusters do not provide a good enough distribution of relevant and non-relevant images to annotate if we just pick at random. A distribution that consists of more relevant images than non-relevant images (i.e. relevant to topics) would appear to be more appropriate if one is interested in using annotations for training data for machine learning. Intuitively it would also seem that topics with a high number of relevant images perform badly due to a low number of training examples; and topics with a low number of relevant images achieve above 0 precision at high recall because these topics have all of their images manually annotated (remember that a *maximum* of 30 images are chosen from the known relevant images). This intuition is made visible in Figure 5.1. The query annotations are used as a demonstration as they are the most interesting of all the learnt annotation methodologies in that they are able to retrieve the highest number of images.

Next, the results obtained when automatically annotating images using the neural network architecture examined in this thesis are discussed. The selection of image to be annotated also intuitively impacts the way a neural network captions images. If a topic contains a diverse range of locations and actions but the images sampled for annotation only cover one of these locations then it becomes obvious that many images will get mis-captioned. In hindsight, taking the time to select a variety of perspectives and locations within a moment to cover edge cases could increase the accuracy of the captions; but it is unknown by how much. Another factor that contributes to the poor retrieval effectiveness is the number of training examples required — there simply may not be enough annotations for the neural network to learn. One thing that can be said for certain is that the number of training iterations required is not the limiting factor in generating accurate captions. Over 200,000 iterations were performed for the three methodologies, and all of them decreased in effectiveness over time. An acceptable number of iterations for this particular setup appears to fall between 30,000 and 100,000; depending on the type of annotation.

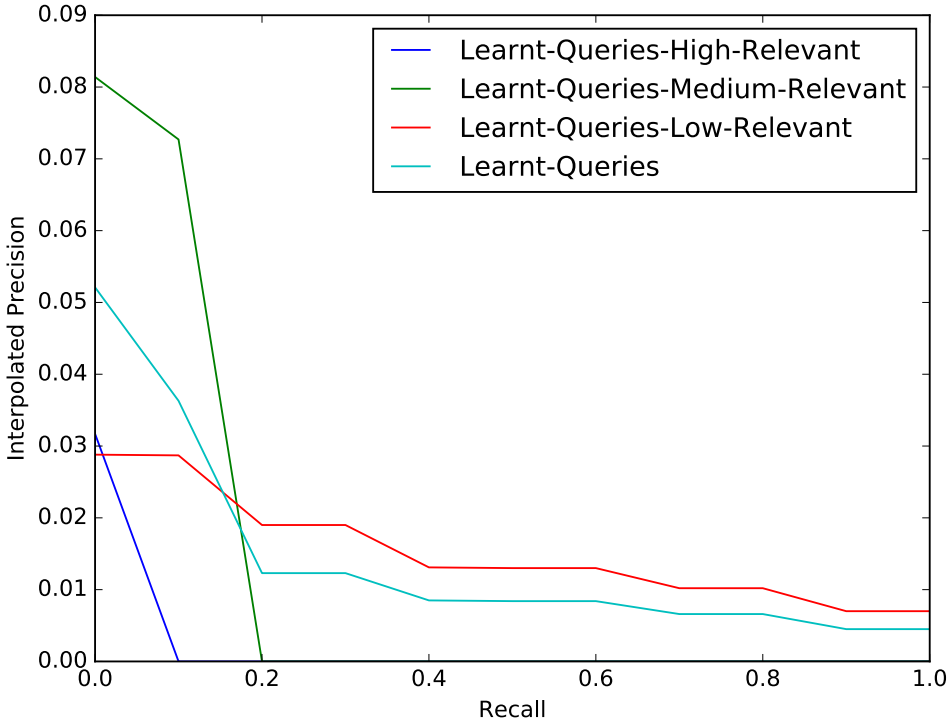


Figure 5.1: Breakdown of running the same experiment on topics with a large proportion of relevant images (>300), the middle number of relevant images (between 100 and 300), and topics with a low number of relevant images (<100)

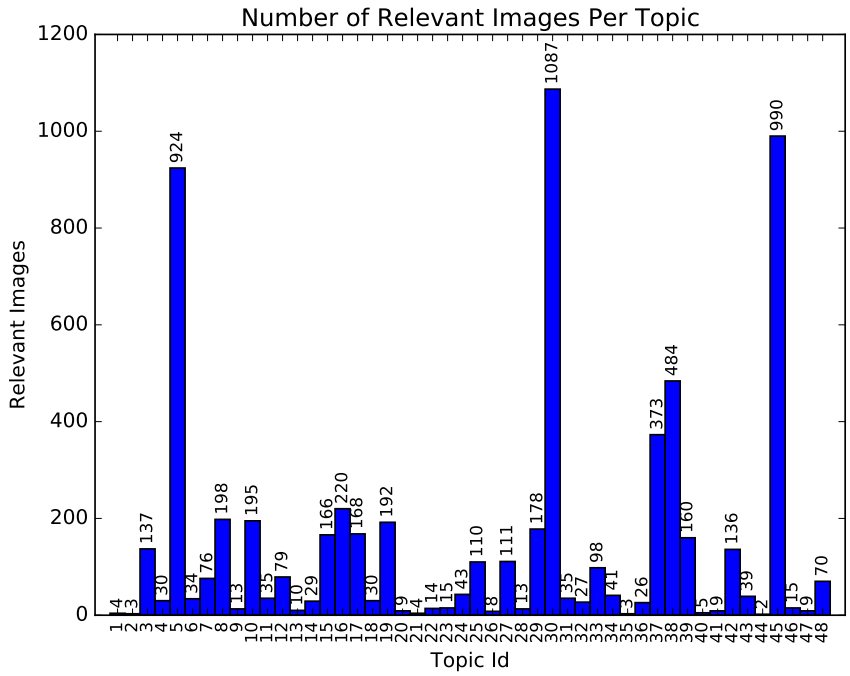


Figure 5.2: Relevant images in each topic

The size of the collection and the number of manual annotations are the most likely reasons for the poor performance when automatically captioning images. Compare for example the number of annotations collected as part of these experiments with the MSCOCO data set [21] which contains more than 300,000 images with five annotations per image. The number of relevant images in each topic may also be a problem: many topics have a less than 200 relevant images associated with them (around 2% of the actual collection). The number of images in each topic is detailed in Figure 5.2

5.3 Interfaces

The original intent of the topics perhaps does not appear to suite annotating individual images sampled at random. Relevant images in a topic are actually a group of images, or a 'moment'. Each topic has several relevant moments, which contain certain images that do not appear to be relevant to the topic at all. As a general rule the images that do not look relevant, even though they are considered to be so by the task organisers are ignored. For instance, in the topic 'Conversation while eating', the lifelogger often wipes his mouth with a serviette blocking the camera. These blurry or obscured images are not annotated.

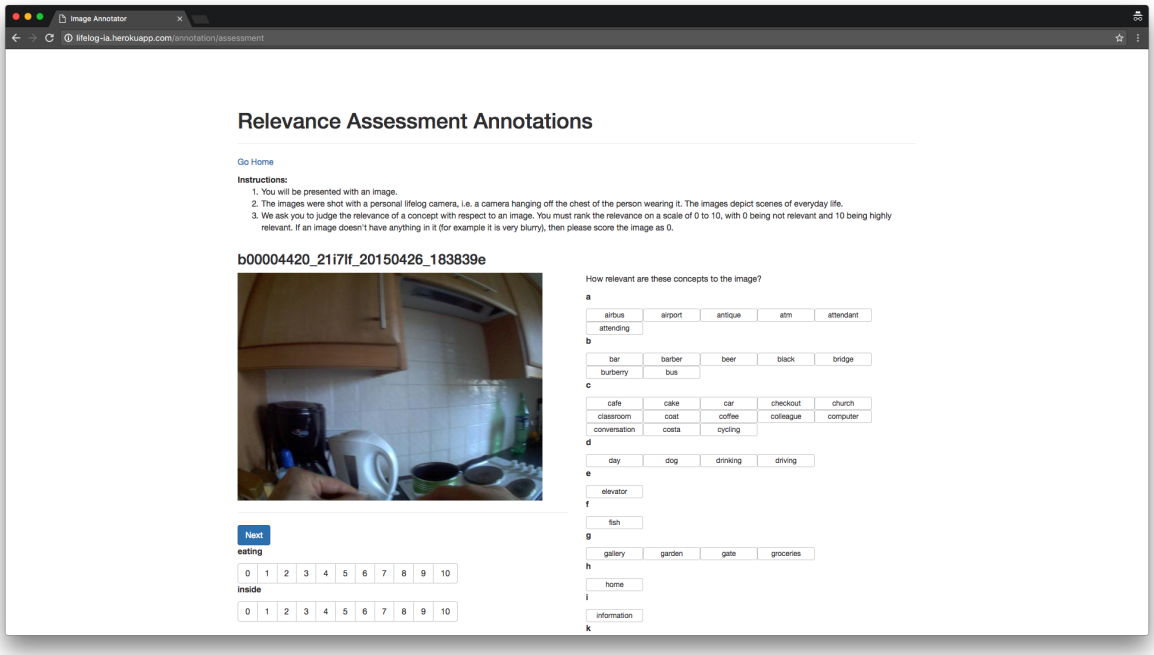


Figure 5.3: Updated query annotation interface

Grouping images into moments may also speed up the annotation process: annotating several images considered to be a moment may not only allow collecting annotations less tedious and time consuming, but could also allow for more images to be annotated. The downside to this, however, may be that these irrelevant images crop up inside each moment. One way to avoid a situation like this is to let the annotator remove images that are covered by objects or too blurry to make anything out.

The interfaces themselves are iterative and dynamic while collecting annotations; they change often in response to feedback from annotators. The biggest change from the initial design is the relevance assessment interface, as seen in Figure 5.3. Now rather than clicking to judge every concept to the image, each concept is grouped alphabetically; when one is clicked, assessments are added underneath the image. Not every caption must be assessed manually: When the next button is clicked, all unassessed captions are automatically considered not relevant.

Chapter 6

Conclusions

This chapter summarises and concludes the research presented in this thesis and discusses possible directions for future work. The contributions of this thesis are summarised by answering each of the research questions.

6.1 Summary of Contribution

How can annotations for a collection of lifelog images be evaluated?

A gold standard annotation does not exist for these lifelog images: individual comparison of annotations and annotation types is not possible. Instead, an alternative extrinsic evaluation methodology is considered. The problem of evaluating image annotations is cast to that of searching for lifelog images based on the collected annotations. In this way image search results are used as a proxy to estimate the quality of the annotations themselves. This extrinsic evaluation method is also the method chosen for the NTCIR-12 lifelog semantic access task. The focus of this research is to investigate the best methodology for annotating images for use in a text-based information retrieval system. The other meta data associated with lifelogs is ignored, in preference of focusing the attention on the images. Supplementary data such as location, time, activity and personal health (heart rate, blood pressure, etc.) are important in pinpointing moments in time, however this research is focused on what types of annotations work best for textual search. Without meaningful annotations the supplementary information becomes useless when performing typical text-based image retrieval. Knowledge of the retrieval effectiveness of each annotation methodology within the context of a text-based image retrieval

system allows the research to concentrate on tuning retrieval models, automatically captioning images, and incorporating the aforementioned meta data.

What methods could be used to annotate lifelog images and what is the retrieval effectiveness of annotations collected with these methods?

Four annotation methodologies are chosen for annotation. Of the selected types of annotations: text, tags, queries, and relevance assessment – the query annotation is the most effective in a text-based image retrieval search task. In fact, not only is the query annotation methodology the most effective, it is the cheapest to collect in that the average time to collect annotations of this type is the lowest of all four methodologies. The second most effective annotation methodology for lifelog images are relevance assessments, however these are time consuming to collect and require prior selection of concepts to judge. The textual and tag annotations have about the same retrieval effectiveness; the difference is that the textual annotations take double the time on average to annotate. The retrieval effectiveness of the annotations is slightly improved when combining all four together, particularly at high recall.

How do annotations generated by current state-of-the-art automatic image captioning techniques compare to the effectiveness of manual annotations?

The effectiveness of annotations generated by one particular state-of-the-art automatic captioning technique is not comparable to the effectiveness of manual annotations. The expectation of providing more training data in combination with some tuning is that the effectiveness will improve. Current results show, however, that the query annotations still retrieve the most number of images overall. In addition, combining the annotations increases the overall effectiveness, especially at low recall.

6.2 Future Work

Moment Annotation

The most important change to make if continuing this research is to segment the collection of images into moments rather than individual images. This would allow many more images to be annotated in a smaller amount of time and cover more edge cases. New techniques are required to sample images into moments rather than individually, and the image annotation interfaces must be updated to reflect this.

Adding and removing images from each moment in the annotation interfaces is necessary (the sampling process is not likely to perfectly capture all relevant images in a moment i.e. there could be outliers on the edges of the moment). One thing that needs further investigation is whether to overlap moments with each other. Is it sensible to annotate images more than once if they appear in moments which overlap?

Automatic Captioning

Automatic captioning tools seem to work very well in other domains and it is unfortunate that this research could not exploit them. The number of annotations used as training data is the most likely reason this research could not generate captions/annotations effectively. Although diverse in moments and objects, there are still less than 100,000 images in total, with less than 20,000 of these images considered 'relevant'. Many topics have less than 100 relevant images so using a data set with more relevant images may improve how images are captioned.

The retrieval system itself can also be expanded to combine TBIR and CBIR systems. There has been recent work done to generate images from textual descriptions by Reed et. al [25]; however it looks limited to a small number of categories and the resolution of the generated images are low. It is not difficult to imagine a system which generates images from a query and retrieves using an image similarity technique.

Image sampling

Another line of future work is to investigate other image similarity measures (such as those used by the LEMoRe team [5] at NTCIR-12) to possibly produce a more uniformly distributed sample set. These similarity measures can be used in combination with a better clustering algorithm. Rather than focus on clustering sets of visually similar images, the result of sampling in future work should produce sequences of moments. Further, annotating moments and the affect sampling moments has on the retrieval effectiveness is yet to be seen.

References

[1] D. A. Ali and S. A. Noah. Semantically indexed and searched of digital images using lexical ontologies and named entity recognition. In *Information Technology (ITSim), 2010 International Symposium in*, volume 3, pages 1308–1314. IEEE, 2010.

[2] I. Askoxylakis, I. Brown, P. Dickman, M. Friedewald, K. Irion, E. Kosta, M. Langheinrich, P. McCarthy, D. Osimo, S. Papiotis, A. Pasic, M. Petkovic, S. Spiekermann, and D. Wright. To log or not to log? - risks and benefits of emerging life-logging applications. Technical report, European Network and Information Security Agency (ENISA), Nov. 2011.

[3] H. W. Bristow, C. Baber, J. Cross, J. F. Knight, and S. I. Woolley. Defining and evaluating context for wearable computing. *International Journal of Human-Computer Studies*, 60(5):798–819, 2004.

[4] W. S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973.

[5] G. de Oliveira Barra, A. C. Ayala, M. Bolaños, M. Dimiccoli, X. Giro-i Nieto, and P. Radeva. Lemore: A lifelog engine for moments retrieval at the ntcir-lifelog lsat task. *Proceedings of NTCIR-12, Tokyo, Japan*, pages 366–371, 2016.

[6] A. R. Doherty and A. F. Smeaton. Automatically augmenting lifelog events using pervasively generated content from millions of people. *Sensors*, 10(3):1423–1446, 2010.

[7] D. Elliott and F. Keller. Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 1292, page 1302, 2013.

[8] H. J. Escalante, C. Hernández, A. López, H. Marín, M. Montes, E. Morales, E. Sucar, and L. Villasenor. Towards annotation-based query and document expansion for image retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 546–553. Springer, 2007.

[9] C. Frankel, M. J. Swain, and V. Athitsos. Webseer: An image search engine for the world wide web. *ONR*, 1996.

[10] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Ntcir lifelog: The first test collection for lifelog research. *Proceedings of NTCIR-12, Tokyo, Japan*, 2016.

[11] C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and trends in information retrieval*, 8(1):1–125, 2014.

[12] C. Hartvedt. Using context to understand user intentions in image retrieval. In *Advances in Multimedia (MMEDIA), 2010 Second International Conferences on*, pages 130–133. IEEE, 2010.

[13] B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt. Ontology-based medical image annotation with description logics. In *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on*, pages 77–82. IEEE, 2003.

[14] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[16] M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 323–330. ACM, 2010.

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, 2004.

[20] H.-L. Lin, T.-C. Chiang, L.-P. Chen, and P.-C. Yang. Image searching by events with deep learning for ntcir-12 lifelog. *Proceedings of NTCIR-12, Tokyo, Japan*, 2016.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[23] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Gcap: Graph-based automatic image captioning. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 146–146. IEEE, 2004.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[25] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *CoRR*, abs/1605.05396, 2016.

[26] B. Safadi, P. Mulhem, G. Quénot, and J.-P. Chevallet. Lig-mrim at ntcir-12 lifelog semantic access task. *Proceedings of NTCIR-12, Tokyo, Japan*, pages 361–365, 2016.

[27] M. Sanderson. Test collection based evaluation of information retrieval systems. *Information Retrieval*, 4(4):247–375, 2010.

[28] H. Scells, G. Zuccon, and K. Kitto. Qut at the ntcir lifelog semantic access task. *Proceedings of NTCIR-12, Tokyo, Japan*, pages 375–377, 2016.

[29] H. Shao, W. cheng Cui, and H. Zhao. Medical image retrieval based on visual contents and text information. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 1098–1103 vol.1, Oct 2004.

[30] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(12):1–174, 2010.

[31] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[32] I. Stanton, S. Ieong, and N. Mishra. Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 133–142. ACM, 2014.

[33] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.

[34] C. Tullo and J. Hurford. Modelling zipfian distributions in language. In *Proceedings of language evolution and computation workshop/course at ESSLLI*, pages 62–75, 2003.

[35] E. van den Hoven. A future-proof past: Designing for remembering experiences. *Memory Studies*, 7(3):370–384, 2014.

[36] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[37] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *ArXiv e-prints*, Sept. 2016.

[38] L. Vuurpij, L. Schomaker, and E. Van den Broek. Vind (x): Using the user through cooperative annotation. In *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, pages 221–226. IEEE, 2002.

[39] L. Wang, L. Liu, and L. Khan. Automatic image annotation and retrieval using subspace clustering algorithm. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 100–108. ACM, 2004.

[40] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[41] L. Wu, S. C. Hoi, R. Jin, J. Zhu, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 135–144. ACM, 2009.

[42] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):716–727, 2013.

[43] L. Xia, Y. Ma, and W. Fan. Vtir at the ntcir-12 2016 lifelog semantic access task. *Proceedings of NTCIR-12, Tokyo, Japan*, 2016.

[44] R. Yan, A. P. Natsev, and M. Campbell. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[45] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval*, pages 29–41. Springer, 2009.

[46] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 461–470. ACM, 2010.

[47] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*. ACM, 2015.

[48] G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *37th European Conference on Information Retrieval (ECIR 2015)*, Vienna University of Technology, Gusshausstrasse, Vienna, March 2015. Springer.