Intent-Aware Diverse Social Image Retrieval

Wang Bo

INTENT-AWARE DIVERSE SOCIAL IMAGE RETRIEVAL

A thesis submitted to the Delft University of Technology in partial fulfillment
of the requirements for the degree of

Master of Science in Computer Science

by

Wang Bo

November 2018

The work in this thesis was made in the:

Multimedia Commputing Group
Department of Intelligent Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Supervisor:	Prof.dr. Martha Larson
Thesis committee:	Prof.dr. Martha Larson
	Prof.dr. Alan Hanjalic
	Dr. Mohammad Soleymani

ABSTRACT

Behind each photographic act is a rationale that impacts the visual appearance of the resulting photo. Better understanding of this rationale has great potential to support image retrieval systems in serving user needs. However, at present, surprisingly little is known about the connection between what a picture shows (the literally depicted conceptual content) and why that picture was taken (the photographer intent). In the thesis, we investigate photographer intent in a large Flickr data set. First, an expert annotator carries out a large number of iterative intent judgments to create a taxonomy of intent classes. Next, analysis of the distribution of concepts and intent classes reveals patterns of independence both at a global and user level. Finally, we report the results of experiments showing that a deep neural network classifier is capable of learning to differentiate between these intent classes, and that these classes support the diversification of image search results.

Keywords Multimedia retrieval; user intent; multimedia indexing; diversification;

PREFACE

The thesis paper is the final work of my Master study in Computer Science at Delft University of Technology.

The thesis is partially based on papers published at ACMMultimedia 2017 MUSA2 workshop and the MediaEval workshop, which benefits the whole thesis with a comprehensive dataset and the advanced evaluation tool. I was lucky to have a chance to go to Trinity College Dublin for MediaEval 2017 and present my working notes on the Retrieving Diverse Social Images Task. Attending such an academic conference and turned out to be an amazing and unforgettable experience for me.

I would like to thank my supervisor, Prof. Martha Larson for her excellent guidance and continuous support during this process. If I ever lost interest, you always kept me motivated. Thanks Prof. Alan Hanjalic and Dr. Mohammad Soleymani for being my thesis committee. I also wish to thank all of the respondents, without whose cooperation I would not have been able to conduct this analysis.

Additionally I would like to thanks Dr. Irek Karkowski, your wise counsel and kind words have, helped me a lot. I would also like to give my special thanks to academic counsellors at TU Delft, especially Ms. Susanne van Aardenne, who always helped me back on the track.

Last but not least, I would like to express my deepest thanks to my family members, including my parents and girlfriend, who were always standing by my side and supporting me with their best wishes.

Bo Wang
Delft, November 2018

CONTENTS

1 INTRODUCTION

1.1 ON SOCIAL IMAGE RETRIEVAL

In today's web, end users create, organize and search documents through social tagging and collaborate activities, this is particularly the case for multimedia documents uploaded on social media websites, such as Flickr[1] and Instgram[2]. These technologies, have greatly boosted the amount of web data, including not only multimedia documents, but also the text associated with these documents. This situation, on one hand, make it possible for information seekers to find what they need. On the other hand, if results in an excess of information processing capacity, which is often referred to "information overload" [16].

In the context of image retrieval, two major approaches have been adopted to better satisfy user's information needs, namely *text(tag)-based image retrieval* [65] (TBIR) and *content-based image retrieval* [68] (CBIR). The text-based approach employs associated text or tags as information cues to convert a multimedia document retrieval problem into a conventional text retrieval scheme [37]. In such systems, tags or even surrounding words were indexed and fed into a well-defined ranking function, such as BM-25 [52]. In the latter approach, images are indexed by their visual content, such as color, texture, and shapes. However, it is more desirable and practical for a user to retrieve photos from the database using textual queries. The retrieval models deployed on the web and by social photo sharing sites rely heavily on search paradigms developed within the field of text retrieval. This way, image retrieval can benefit from years of research experience from TBIR.

According to the *Probability Ranking Principle* [53]: 1. The user is likely to browse retrieved results sequentially, which means a document with a higher probability of relevance should always be ranked ahead of a document with a lower probability of relevance. 2. The utility of a document to a user is independent of the utility of other documents to the same user. As a result, the vast majority of information retrieval models, such as probabilistic ranking models, vector space models or learning to rank models, are designed to maximize individual relevance.

In the real world social image retrieval scenario, these assumptions do not always hold. First of all, previous studies indicate that, through images are visual information sources with little or no associated text, users mainly use text to formulate their queries [55], and such web search queries are typically short, ranging from two to three terms in average [30], often leverage a certain amount of ambiguous. Moreover, even if the query submitted by the user is clearly specified, it is still non-trivial for retrieval system to determine

1 https://www.flickr.com/
2 https://www.instagram.com/

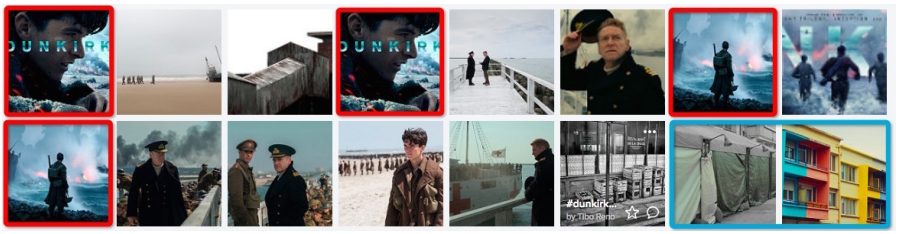

Figure 1.1: Pure relevance oriented retrieval results might return duplicate results (red box), inconsistency between visual information and user assigned tags might bring unrelated documents (blue box). Moreover, though the query is clearly specified, retrieved results do not cover sub-aspects.

which particular aspect the user is interested in. Secondly, by fully ignoring the internal relationship between retrieved documents, relevance-oriented ranking algorithms often return redundant results [29]. Thirdly, since user tagging is known to be uncontrolled, ambiguous, and personalized, thus, it is non-trivial to interpret the relevance of a tag with respect to the visual content it is describing.

1.2 FROM RELEVANCE TO DIVERSITY

When both semantically and visually related query results appear simultaneously as top results, such relevance-oriented approaches, however, can not fully address the usefulness issue [21]. The evidence can be clearly seen in figure 1.1. We collected top ranked results from Flickr gave the query term "Dunkirk movie". Though retrieved results are highly correlated, drawbacks can still be observed. For this reason, "search result diversification" has attracted enormous research efforts from both academia and industry.

Search result diversification was originally proposed in the text retrieval domain. For example, motivated by the need for "relevant novelty", Carbonell et al. [7] proposed to use "maximal marginal relevance" (MRR) for document re-ranking. Zhai et al. [75] introduced "sub-topic retrieval" to find documents that cover many different subtopics of a query topic. Similarly, [2] proposed a systematic method named "IA-Select" (intent-aware) to diversifying results that aim to minimize the risk of dissatisfaction of the average user. These approaches has greatly changed the query results built on the assumption of "independent relevance".

Meanwhile, image search result diversification approaches mainly employ multi-modal features such as visual, text and meta-data to reordering query result [67; 28]. In general, for each image, these approaches first build a feature vector, then apply various unsupervised learning algorithms to create image clusters and aggregate search results based on these clusters.

However, image search result diversification, especially the context of social platform, is even more challenging since it needs to tackle multiple problems such as query ambiguity, visual redundancy and latent query intents listed in section 1.1. For this reason, a well-defined image retrieval ranking/ranking aggregation algorithm should consider multiple dimensions, that is: "coverage", "novelty" and "broad latent aspects".

1.3 COVERAGE, NOVELTY AND BROAD LATENT ASPECTS

Having discussed the importance of search result diversification, it is now necessary to explain the course and solutions. In a previous study [47], diversity was categorized into two manners, "extrinsic diversity" and "intrinsic diversity".

Extrinsic diversity was defined as *diversity as uncertainty about the information need*, and it can be further grouped into two sub-groups, that is "ambiguous queries" and "underspecified queries". In the first case, a query might refer to different interpretations. For example, query "Dunkirk" might represent for the name of a place or the name of a movie. To address the ambiguous query issue, the returned results should cover as many sub-topics as possible. In the latter case, we are uncertain about which particular aspect does the user interested in. An example of this is the query "Dunkirk movie". We know exactly that the user has a movie named Dunkirk in his/her mind, but we do not know which type of the image the user is searching for: the movie poster? director or scenes? This situation can be solved by covering as many latent aspects as possible.

Intrinsic diversity can be defined as *diversity as part of the information need*. More concretely, diversity should avoid redundant documents (images) and presenting a novel and a useful set of retrieved result.

Problem	type	Solution
ambiguous queries	extrinsic	cover different interpretations
underspecified queries	extrinsic	cover latent aspects
visual redundancy	intrinsic	produce novel results list

Table 1.1: Components of search result diversification.

The effectiveness of various social image search result diversification has been exemplified in several studies. For example, the MediaEval Retrieving Diverse Social Images task [27], is an annual benchmark in which participants develop algorithms for refining the results by providing a set of images that are relevant to the query and, at the same time, offering a visually diversified summary of it was taken. However, since researchers have focused nearly exclusively on visual diversity (novelty), and this might result in intrinsic drawbacks.

Being aware of this, a re-ranking method based on topic richness analysis is proposed to enrich topic coverage in retrieve diverse social images was proposed by Song et al. [64]. The authors defined a quantitative measurement named "topic richness" and re-ranked retrieved images based on topic richness score. It is also interesting to observe that when it comes to extrinsic diversity problem, the author started to employ textual feature for diversification.

However, the studies would have been far more convincing if the researchers had analyzed such issues:

1. **Enrichment issue**: How can we return search results that cover different latent aspects when facing underspecified queries?

2. **Transfer issue**: Can we adopt search result diversification algorithms developed for text retrieval to social image search result diversification?

3. **Modality correlation issue**: Does there exist a global correlation between diversity type (i.e. extrinsic & intrinsic) and different modalities?

Before we dive into the transfer issue and modality correlation issue, it is necessary to investigate a strategy to refine a result list that can cover multiple latent aspects. In the next section, we will first present our motivation for intent-based search result diversification that serve to address enrichment issue.

1.4 INCORPORATE INTENT FOR DIVERSIFICATION

The area of multimedia information retrieval has seen exceptions to the tendency towards "what" at the cost of "why", which include recent work on user search intent [40; 51; 34]. The purpose of intent-based search result diversification is to build on previous work and to provide concrete, large-scale substantiation of the importance of moving multimedia analysis beyond concept detection, and the potential of techniques that automatically detect visual classes related to user intent to improve image retrieval. Contrasts in user intent are illustrated in Fig. 1.2, which shows four images uploaded to Flickr relevant to "sailing". Despite the fact that all four are relevant to the same topic, clear differences can be observed.

We argue that user intent should be exploited in image retrieval since the goals of the photographer provides a simple, easily understandable explanation for the differences observed between photos. If we were to restrict the descriptions of photos used for indexing to containing only visual concepts, then both pictures in the top row could be described as "Photo depicting a sailboat in water". We could then capture the fact that users see a clear difference between the two by describing the first as "sailboat in a lot of water" and "sailboat near a dock". However, such descriptions are not entirely satisfying since the exact amount of water or the presence of the dock is not the key characteristic differentiating these photos. Instead, the difference can be simply and naturally explained by user intent, i.e., the goal of the photographer in capturing the image.

We build on the assumption that user intent, i.e., the goals that users are pursuing when they take photos or search for images, has visual reflexes that can be captured by automatic visual classifiers. Our motivation for assuming a connection between the reason why a user takes a photo, and the visual content of the photo, is the phenomenon of *intentional framing*. Riegler et al. [51] define intentional framing as, 'the sum of the choices made by photographers on exactly how to portray the subject matter that they have decided to photograph.' This definition implies that intentional framing is observable in the photo, and also captured the fact that intent cannot be reduced to the concepts that are literally depicted in the image. Our notion of intent is related to the idea of *broad latent aspects* from conventional web search [72]. This type of query aspect is described by [72] as having two

properties: broad latent aspects applies to a broad set of queries, and, users queries frequently leave these aspects unspecified.

Figure 1.2: Example search results for the query "sailing". The intent (goal) of the photographer is different for each picture. Plausible characterizations of these goals are: provide an overall impression of the space in the wider world where sailing takes place (top left), depict a sailboat as an object (top right), capture a portrait of someone sailing (bottom left), capture information on sailing from another media source (bottom right).

Note that the user intent behind a query does not reduce to a sub-topic of a query. To illustrate this point, we return to figure 1.2 to discuss the two pictures in the bottom row. If we considered these pictures sub-topics of sailing, then we would lose connection with the intent of capturing a portrait and the intent of capturing media information that apply to a large range of queries beyond "sailing". In other words, intent cross-cuts the topic (conceptual content) of a photo and provides us with an additional, easily understandable dimension with respect to which images can be indexed for image retrieval.

1.5 RESEARCH QUESTIONS

In this work, we will first investigate the ability of automatic intent inference to improve image retrieval, with a focus on retrieval algorithms that diversify image results lists. Given that intentional framing is noticeable to users as visible in photos, but cannot be directly reduced to topical or conceptual categories, it is surprising that user intent has yet to be fully exploited in image retrieval systems. Because the intent of the user is actually the goal of the user, we anticipate that user intent potentially has an important contribution to make to image retrieval. We believe that there are three major roadblocks that explain the relative lack of research on intentional framing in image search, which we now discuss in turn. The three major contributions of this paper address each of these three roadblocks.

The first roadblock is the lack of intent taxonomies (definitions of intent classes) and data sets annotated with intent labels. With the exception of early work by Lux et al. [40], we do not know of work that has asked

photographers about their intent. However, asking photographers might not always be necessary. As mentioned above, people also perceive intent classes when they look at images. However, there is no standard set of intent classes that has universal applicability, and inner-annotator agreement on intent judgments is difficult to achieve. We address this roadblock by moving away from the idea that we should assume that a single authoritative set of intent classes is necessary before progress can be made on intent-based image retrieval. Instead, we pursue the idea that any well-informed intent taxonomy that has been applied consistently in order to label an image data set will be able to drive forward the state of the art in applying user-intent to improve image retrieval. To this end, we build an intent taxonomy on the basis of the analysis of a large collection of social images by a single expert annotator. The first contribution of this paper is this taxonomy and a large set of images labeled with the intent classes that it contains. This resource will allow reproduction of our research results and serves as a point of reference for future extension or modification of the intent taxonomy.

The second roadblock standing in the way of applying intent-based approaches to image retrieval is general assumptions about the nature of intent. We believe that the currently dominant assumption is that the visual variability among photos associated with a single intent class is too high for a classifier to be able to generalize. Likewise, the visual similarity among photos associated with different intent classes is also assumed to be too high. The second contribution of this paper is to show that these assumptions should not be made: with enough labeled data it is possible to train a classifier capable of automatically inferring the intent classes of images.

The third roadblock is the assumption that the usefulness of intent in image retrieval is likely to be limited. The third contribution of this paper is a set of experiments on multimodal social image retrieval that show that our intent classifier is able to contribute productively to the diversification of image search results.

Thus we formalize related research questions as follows:

1. **RQ1**: Are there noticeable patterns in social images that go beyond the conceptual content of the image and reflect the goals of the photographer?
    a) Can we build a taxonomy of classes that reflect the goals of the photographer?
    b) Do images with the same subject matter reveal different perspectives and goals of users (user intent)?
    c) Do categories that express a certain perspective/goal (i.e., user intent categories) appear across different types of subject matter (concept categories)?
2. **RQ2**:Does user intent, i.e., the goals that users are pursuing when they take photos or search for images, have visual reflexes that can be captured by automatic visual classifier?
    a) Can we build a visual classifier to interfer the rationale behind a photographic act?
    b) How does the visual classifier performance in the face of high visual variability and high visual similarity?

3. **RQ3**: Is the intent classifier is able to contribute productively to social image diversification?

As was pointed out in section 1.3, images on the Web can be searched indirectly via the accompanying textual information based on ranking function developed for text retrieval. Consequently, search result diversification algorithm originally proposed for text retrieval probably can be applied to image search result diversification as well. For this reason, we will also investigate:

4. **RQ4**: Can we directly employ text-based search result diversification, such as maximal marginal relevance for social image search result diversification?

And finally, we analyze the correlation between feature modality and diversification type by conducting error analysis:

5. **RQ5**: Is diversity type, i.e. extrinsic or intrinsic diversity, affected by different feature modalities?

1.6 THESIS STRUCTURE

The remainder of the thesis is organized as follows: In Chapter 2, we cover related work, focusing on the relevance-oriented ranking scheme, the diversity-oriented ranking scheme and multimedia intent taxonomies that have been proposed in the literature. Then, in Chapter 3 we describe the iterative process by which we create our intent taxonomy and label our image data set. Next, we train a classifier using the labeled data, and report the results of experiments demonstrating its effectiveness. Following this, in Chapter 4, we apply our intent classes to social image search result diversification. In Chapter 5, we conduct an experiment using search result diversification originally proposed for text retrieval on social image retrieval. The next Chapter 6 describes the correlation between diversity type and feature modality by performing error analysis. Chapter 7 of the thesis provides conclusions and an outlook.

2 PRELIMINARY AND RELATED WORK

Thus far, the thesis has introduced the motivation and research questions. In this chapter, we will briefly present related work on purely relevance-oriented ranking and analyze its drawbacks (Section) 2.1 . In Section 2.2, we cover the most prominent work in the domain of search result diversification. Lastly, in Section 2.3, we will introduce the work that has been done in the area of intent in multimedia research. More specifically, we put our focus on intent in image retrieval and intent taxonomy.

2.1 RELEVANCE ORIENTED RANKING

Information Retrieval can be defined as *finding material (usually documents) of an unstructured nature that satisfies an information need from within large collections stored on computers* [57]. Here, *Ranking* is the central problem. Through years of work, many different ranking models have been proposed and the majority of these models were designed to maximize independent relevance.

Roughly, The most commonly adopted image ranking algorithms[1] can be categorized to text-based approaches (such as *similarity-based models* and *probabilistic models*) and learning-based approaches (such as learning to re-rank). In order to motivate diversity-oriented ranking, we will first discuss each of these ranking schemes blow.

2.1.1 Text-Based Approaches

In a tag-based (keyword) image retrieval scenario, the associated textual information such as tags, description, and title were indexed and aggregated using ranking schemes developed for text retrieval problem. In which, we further selected three representative text-based ranking functions, including pivoted document length vector space model, okapi-bm25 and language modeling as an example, to see how text-based ranking schemes were designed, and their drawbacks.

As a similarity-based retrieval model, pivoted document length normalization vector space model was originally proposed in [62]. In this case, the relevance status is measured based on the correlation between user query and documents. A query and each document are represented as two high-dimensional term vectors, each value of the vector reflects how important the term is. The score between the query and document is then calculated by employing a distance measure, such as cosine similarity.

As one of the best performing vector space model, pivoted document length vector space model was defined as:

1 hereby, we mainly consider query dependent models

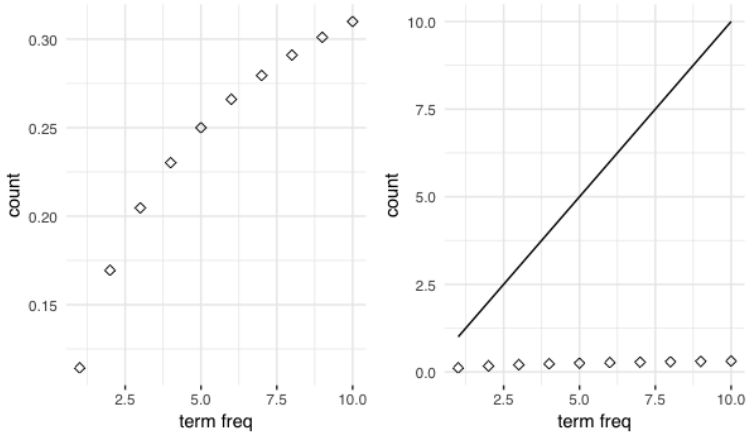

Figure 2.1: Term frequency transformation in the pivoted document normalization vector space model. Frequency count decays with the increase of term frequency (left figure). Frequency count comparison between with transformation (left figure) and without transformation (right figure)

$$f(q,d) = \sum_{w \in q \cap d} \frac{ln(1 + ln(1 + tf(w,d)))}{1 - b + b * \frac{|d|}{avgdl}} * ln \frac{N+1}{df(w)}$$

By decomposing this formula, three major components can be clearly observed. First of all, by applying a certain level of transformation on term frequency (numerator), the terms occur more frequently were more favored (see Figure 2.1). Secondly, incorporating inverse document frequency (i.e. $ln \frac{N+1}{df(w)}$, where N is the number of documents in collection), more distinctive terms will be weighted higher. Thirdly, to ensure that the difference between documents length will not post a significant impact on the final score, a free parameter b was used to control the degree of document length normalization.

As compared with vector space models, the ranking functions built based on a probabilistic ranking principle such as Okapi-BM25 [54] and language models [19] attracted more attention in the past decades. By ranking documents based on their log-odds relevance, BM-25 (one instantiation) can be defined as:

$$f(q,d) = \sum_{w \in q \cap d} \frac{(k+1) * tf(w,d)}{tf(w,d) + k * (1 - b + b * \frac{|d|}{avgdl})} * ln \frac{N+1}{df(w)}$$

In addition to the original BM-25 being derived from the probabilistic model, one major difference between BM-25 and pivoted document length vector space model is the term transformation function. Like the pivoted function, BM-25 favors frequent occurring terms, while it penalize term frequency with $\frac{(k+1)*tf(w,d)}{k+tf(w,d)}$, where k is a free parameter to control the degree of term frequency penalization. It is also interesting to observe that BM-25 also assign higher weight using inverse document frequency while penalizing longer documents.

Another representative retrieval scheme is language models. In which, we usually differentiate these ranking function based on the type of smoothing

methods they use, such as the Jelinek-Mercer method and Dirichlet Prior Method [77]. These methods use the smoothing method to smooth a document language model and then ranks documents according to the likelihood of the query according to the estimated language model of each document. For instance, the Jelinek-Mercer language model can be defined as:

$$f_{JM}(q,d) = \sum_{w \in q \cap d} log[1 + \frac{1-\lambda}{\lambda} * \frac{tf(w,d)}{|d| * p(w|C)}]$$

where $p(w|C)$ is the probability of a term w given by the collection language model that indicates its popularity, and λ is the smoothing parameter ranges from $[0,1]$. In this formula, it's clear to see that $tf(w,d)$ is used to assign higher weights to more frequently occurred terms, $|d|$ is used for document length normalization, and $p(w|C)$ has the same functionality as inverse document frequency, which assign higher probability to distinctive words.

Constraints	Intuitions
TFC1	to favor a document with more occurrence of a query term
TFC2	to favor document matching more distinct query terms
TFC2	to penalize a larger TF (term frequency transformation)
TDC	to regulate the impact of TF and IDF
LNC1	to penalize a long document(assuming equal TF)
LNC2, TF-LNC	to avoid over-penalizing a long document
TF-LNC	to regulate the interaction of TF and document length

Table 2.1: Information Retrieval Constraints (Source: A Formal Study of Information Retrieval Heuristics [17]).

As was pointed out by [17], the majority of conventional information retrieval ranking functions were built based on several heuristics (see Figure 2.2), and these heuristics can be clearly observed by decomposing ranking functions as we did. These heuristics has greatly boosted the "relatedness" between query and document, although there also exist several drawbacks. First of all, none of these raking constraints take the uncertainty of the user into consideration, such as, as we mentioned, the ambiguous queries and underspecified queries. Secondly, none of these heuristics take the internal relationship between ranking results into account, since they are designed to maximize "independent relevance".

2.1.2 Learning-to-rerank Approaches

Under a social image retrieval context, a query is typically monomodal (e.g. a text term), while the relatedness of an image can be spread over multi-modalities. One popular approach for image retrieval is "multimodal reranking". The rationale behind multi-modal ranking is to use the information from modality X to refine the initial ranking list based on information from modality Y. Usually, X refers to textual descriptors and Y refers to visual descriptors.

Learning-to-rerank has a similar underlying rationale as the learning-to-rank paradigm that is known from IR [36]. Though both leaning-to-rank and leaning-to-rerank share several common characteristics such as the collect a set of training data, utilize human efforts to label documents and extracting features for machine learning algorithm, a clear difference can still be observed. As was discussed in [74], there are two likely causes for the differences between leaning-to-rank and learning-to-rerank. First, in learning-to-rerank, the initial ranking result from the text-based search serves as a prior, which needs to be effectively incorporated into the re-ranking process. Second, in learning-to-rerank the query and the documents have different representations, i.e., the query is textual while an image is visual. The typical learning-to-rerank framework can be seen in Figure 2.2

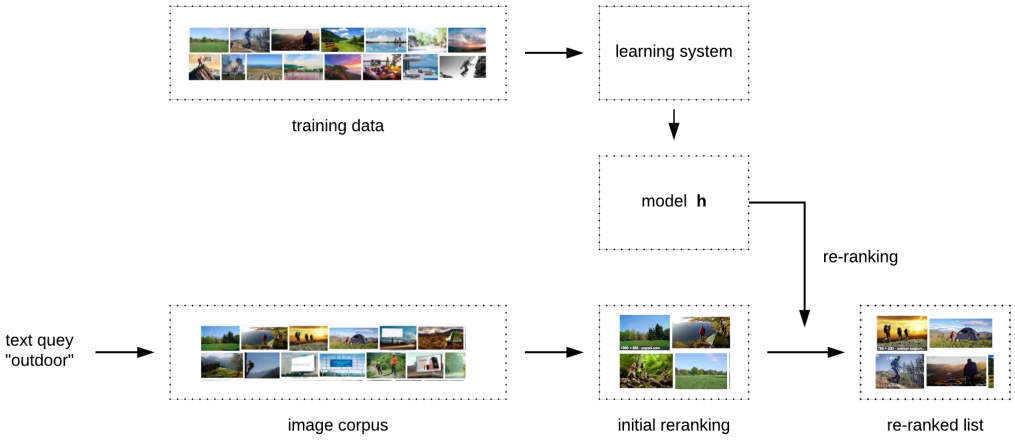

Figure 2.2: Learning to re-rank for image retrieval. In the first step, a model was learned on training data with human-provided labels. Once the initial ranking results were returned based on text-based ranking function, the learned model is then employed to refine the ranking list.

The re-ranking model f can be written as a composition of two functions, h and g:

$$f(q,d) = h \circ g(d,,q)$$

where h is the query independent function learned in a supervised fashion, and g is a query dependent function learned from the initial ranking result in an unsupervised fashion.

In the query-dependent case, function g is learned to model the relevance relation between the images in the initial list and the query in terms of the visual modality, and the function h is set to a constant. For example, Yan et al. [73] proposed a classification-based pseudo-relevance feedback approach that attempts to apply SVM ensembles to refine the initial retrieval result in content-based video retrieval. Similarly, Schroff et al. [59] applied an SVM visual classifier to improve the initial ranking result using top-ranked images. Hsu et al. [25] proposed a novel and generic approach based on the Information Bottleneck Principle which finds the optimal cluster of images.

On the other hand, in 2010, Yang et al. [74] proposed a novel approach that is able to learn a generic, query-independent function for the re-ranking purpose using conventional support vector machine algorithm. However, it's interesting to take a deeper look at its scoring mechanism. The results of the initial text search as a strong prior, thus, were used as the prior reranking features. Besides, two types of features including contextual reranking features and pseudo-relevance feedback based features were taken into consideration. First of all, contextual reranking features were built upon the visual consistency assumption, i.e, visually consistent images should be more relevant. As a result, seven sub-features were created according to a different voting strategy based on visual similarity. Another type of feature was created based on Pseudo Relevance Assumption (i.e. top-ranked images are more likely to be relevant than bottom-ranked images), where PRF was calculated based on probability density estimation and duplicate detection function.

Feature	Description
IR	Initial Ranking
HV_N	Hard Voting of Neighbors
RSV_N	Initial Rank based Soft Voting of Neighbors
$NRSV_N$	Neighbor Rank Weighted Initial Rank based Soft
HV_R	Hard Voting of Reciprocal Neighbors
RSV_R	Initial Rank based Soft Voting of Reciprocal Neighbors
NSV_R	Neighbor Rank based Soft Voting of Reciprocal Neighbors
$NRSV_R$	Neighbor Rank Weighted Initial Rank based Soft Voting of Reciprocal Neighbors
PRF_d	Local Density Estimation for PRF
PRF_{dv}	Duplicate Voting for PRF
PRF_{sdv}	Soft Duplicate Voting for PRF

Table 2.2: An overview of the proposed reranking features, Source: [74]).

In this case, each reranking feature can be considered as a query dependent meta-ranker, and the final query independent reranking function h takes these meta-rankers' output as input to produce final reranking list.

To conclude, models of learning-to-rerank can be treated as an extreme case of conventional text retrieval functions, that starts to employ visual features to reinforce the relatedness between query and document. Both query dependent and independent learning to re-rank models still cannot address the extrinsic issue, more concretely, none of these features is designed to capture the uncertainty of the user's information need. However, as introduced before, by incorporating duplicate detection feature into a supervised learning-to-rerank framework, Yang's work has the potential to produce a more "useful" ranking list.

2.2 DIVERSITY ORIENTED RANKING

The previous section has shown that how relevance oriented ranking schemes were designed and their drawbacks. In the section that follows, It is now

necessary to explain the work that has been done in the field of search result diversification. This topic can best be treated under two headings: text-based approaches and visual-textual joint approaches.

2.2.1 Text-based Approaches

Diversity-oriented ranking has been proposed as a means to overcome ambiguity and redundancy during the search process. various approaches have been proposed in the literature for the search result diversification problem. As stated in [58], the vast majority of these approaches differ by how they implement the objective function. The authors defined a taxonomy of search result diversification algorithms based on two complementary dimensions: aspect representation and diversification strategy. Aspect representation determines how the information need underlying a query is represented as multiple aspects of this query, i.e. the ambiguous query and underspecified query introduced in Chapter 1. The diversification strategy determines how to achieve the goal of satisfying the multiple aspects underlying a query, i.e. the objective we want to maximize, such as novelty (produce a novel result list by removing the redundant documents) and coverage (produce a result list that covers multiple possible interpretations and latent aspects).

The very first diversify-oriented ranking algorithm, named MMR [7] (maximal-marginal-relevance) was motivated by the need of "relevant novelty" as a potentially superior criterion. The author tried to measures the relevance and novelty independently and provided a linear combination as the metric called "marginal relevance". As mentioned in the introduction, MMR is a seminal work for results diversification and it led to several follow-ups.

Zhai et al. [75] study both novelty and relevancy in the language modeling framework. They proposed an evaluation framework for subtopic retrieval, based on the metrics of subtopic recall and subtopic precision. The author also proposed a cost-based approach to combine relevance and novelty using the same rationale as MMR. This work was further formalized in [76], in this paper, queries, and documents are modeled using statistical language models, user preferences are modeled through loss functions. To this end, they derived a new retrieval model for subtopic retrieval, which is concerned with retrieving documents to cover many different subtopics of a general query topic.

In [9], Chen et. al. argue that a natural and practical approach when designing a retrieval algorithm is to optimize the expected value of the metric. They proposed an objective function that aims to find at least one relevant document for all users. In both cases, the topic of the query is considered only tacitly. Clarke et al [11] study diversification in the context of question answering. They focus on developing an evaluation framework that takes both novelty and diversity into consideration. In their work, questions and answers are treated as sets of "information nuggets", and relevance is a function of the nuggets contained in the questions and the answers. Based on the concept of information nuggets, they proposed to use α-NDCG to measure the quality of a ranking system[2].

2 We will further introduce α-NDCG in the experiment section

By analyzing the logs of a search engine, Radlinski et. al. [48] proposed to diversify search results using query-query reformulations. Given a query, the authors tries to generate a list of related queries in order to yield a more diverse set of documents. Similarly, Ziegler et. al. [78] proposed a novel method to balance and diversify personalized recommendation lists in order to reflect the user's complete spectrum of interests.

2.2.2 Visual-based Approaches

Due to the reliance on the textual information associated with an image, image search engines on the Web lack the discriminative ability to deliver visually diverse search results. It is also commonly known that an image has to be seen to fully understand its semantics, significance, beauty, or context, simply because it conveys information that words can not capture. Thus, a lot of work has been proposed to create a better visually diverse ranking of the image search results to satisfy user's information need.

Leuken et. al. [67] employed a unsupervised learning (clustering) scheme. A dynamic feature weighting function was used to remeasure the similarity between hand-crafted features, such as color histogram and edge histogram. The results were clustered using three different strategies, including folding, maxmin and reciprocal election.

Wang et. al. [71] categorized current image search result diversification approaches into two groups: clustering-based approaches and duplicate removal approaches. Based on their analysis, the authors claimed that it is non-trivial to determine the number of image clusters or the duplicate threshold, then proposed an algorithm called DRR (diverse relevance ranking) to optimize a performance metric that considers both relevance and diversity. First, this algorithm estimates the relevance scores of images with respect to the query term based on both the visual information of images and the semantic information of the associated textual information. Then, the authors estimates the semantic similarities of social images based on their textual information (tags). Based on the relevance scores and the similarities, the ranking list is generated by a greedy ordering algorithm which optimizes ADP (average diverse precision). Similar to our case, an experiment was conducted on a social image retrieval setting. Experimental results reveal that the DRR algorithm is able to achieve better performance compared to common social image ranking algorithms, such as time-based ranking, relevance-oriented ranking, and folding strategy introduce in [67].

To foster new technology for improving both relevance and diversification of image search results with explicit emphasis on the social media context, Ionescu et. al [29] introduced a new evaluation framework and dataset (named Div150) [26]) for benchmarking search result diversification techniques and discuss its contribution to the community by analyzing the results of the annual MediaEval Retrieving Diverse Social Images Task such as [27]. During the benchmark competition, various types of visual diversification approaches has been examined, such as clustering-based approaches [60; 45], multi-modality fusion-based approaches [50], relevance feedback based approaches [6] or our intent based approaches [69].

Apart from these algorithms developed for a generalized setting, some image search result diversification algorithms have been proposed to resolve

task-specific issues, such as landmark (geographical) re-ranking. For example, Qian et. al. [46] proposed a 4-step approach to summarize landmark viewpoints, Rudinac et. al. [56] employed a multi-modal graph to create a diverse, yet representative, image set to summarize geographic areas.

2.3 INTENT IN MULTIMEDIA RESEARCH

2.3.1 Intent in Image Search

Triggered by a task, end user starts to consult the search engine with a set of information needs using a set of query terms. Since [4], a huge amount of work has been done to analyze user's information need in the context of text retrieval. While in the domain of multimedia information retrieval, it is even harder to capture the diversity of information needs.

In [34], a survey of user intent in multimedia search, an overview is provided of papers that have made use of user intent to improve image search. Note that we are interested in papers that consider "user intent" to be the goal of the user, and not in work that users "intent" as a synonym for subconcept, sub-topic, or query aspect. An early effort [35] used a content-based intent classifier to adapt the display of search results. Recently, [63] used content-based image classification as one of a range of session-derived signals used to study user intent during an image search. This work is interesting because it focused on predicting intent by processing the sequence of images that users view in an image search session.

Equally important to why user form a query (i.e. the search intent), it is also interesting understanding why user annotate images and how people annotate images. According to Ames et al. [3], the tagging intent, again, was categorized by two dimensions: sociality and function. Some exemplary examples of this are, one motivation was named *self-organization: search and retrieval*. Which means for this particular type of users, they want to better manage their image by adding structured tags. Apart from organization purpose, a group of users was categorized into *self-communication: memory and context*. Participants enter tags to add context to a photograph, such as the names of the people that appear in it or the name of the place it was taken, in order to aid future recall of the situation. Similarly, another two groups are denoted as *social-organization* and *social-communication*, which aims to provide public reachable tags and add public contextual information respectively. Undoubtedly, users motivated by *organizational* purpose are likely to provide more structured, meaningful tags. While people motivated by *communication* purpose usually attach too specific, fine-grained tags such as name, a location which makes training a tag prediction model a lot more complicated by leveraging noisy data. Also, other work conducted by Nancy et al. [24] reveals that the motivation can be grouped into *Memory-Identity-Narrative*, *Maintaining Relationships*, *Self-Representation* and *Self-Expression*.

Recently, the work was done by Riegler et al. [51] introduced *intentional framing*, which goes beyond visual concepts, events, and scenes. Intentional framing was defined as *the sum of the choices made by photographers on exactly how to portray the subject matter that they have decided to photograph*. More

specifically, as introduced in Figure 1.2, though all images reflect the topic of "police", these four pictures still reflects four different photographers intents. Since on most photo sharing websites, it is the photographers themselves who upload and label the images [42], to understand intentional framing could be critical to refine or predict the tags associated with the photo. However, the authors didn't investigate the possible intents behind intentional framing.

Despite this, some work has been done to analyze the underlying pattern behind image production and sharing. By interviewing 10 people covering 40 different situations, Lux et al. [39; 41] studied the user intentions deriving photo production. Result shows the reason for photographers to capture an image can be classified into six reasons: *preserve a good feeling*, *preserve a bad feeling*, *share with family and friends*, *share with public*, *support a task* an *recall a specific situation*. Other work was carried out by Kindberg et al. [32]. In this research, the photographer's intent was categorized into six clusters dividing by two dimensions: social-individual dimension and affective-functional dimension.

2.3.2 Taxonomies of Image Intent

The earliest taxonomy of image search intent is most probably that of Fidel et al. [18], who define intent along a spectrum from a *Data Pole* (obtain source of information), the *Information Pole* (obtain an example of something or an object) and *In between* (retrieve images both as information sources and objects). To create their taxonomy they first categorized images into 12 attribute classes based on Jorgensen's research [31]. A user study was then conducted on 100 image search requests. They claimed that the Data Pole and Information Pole represents different retrieve tasks, and these two types of intent usually associate with different searching characteristics. For the queries categorized as Data Pole, the relevance can be determined ahead of the retrieving action. Usually, users already have a clear criterion on relevance level before they conduct the searching task, this type of query is easier to express using textual queries and other verbal clues. On the other hand, for the queries categorized as Information Pole, according to the author, the user recognize relevance criteria when they see them. Normally, these relevance criteria are latent and difficult to express with text queries and verbal clues. In some cases, two very different images might satisfy the same piece of information need.

Other early work includes Cunningham et al. [13], who investigated 98 video search queries and categorize the user search intent into 8 groups. Including *mental state*, *video*, *audio*, *learning*, *social*, *MSM*, *temporal* and *others*. For example, *mental state* explicit the reference to subject's emotional state, *visual* explicit references to visual aspects to target video, *MSM* explicit references to mainstream media. The author categorized these 8 query intents using an open-card sorting approach by asking a "why" question (i.e. why the video was viewed). As introduced by the author, a total number of 119 queries were investigated, 47 out of 119 queries can be categorized to the group of *mental state*, which is much more than the rest 7 categories.

Valuable image intent taxonomies have bee proposed on the basis of use studies. Lux et al. [38] conducted user research on 20 image search users on

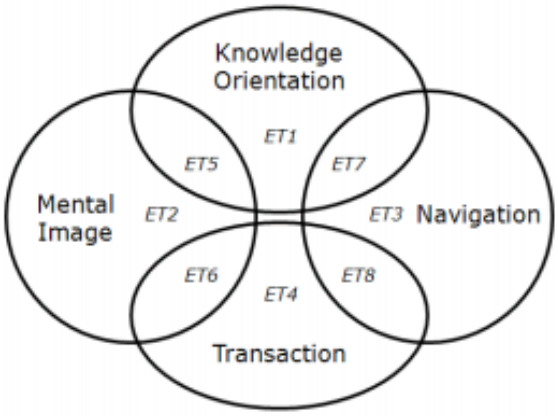

Figure 2.3: Venn diagram showing the classification scheme and the overlapping of classes [38]

Flickr, and build an intent taxonomy with four classes: knowledge orientation, transaction, mental image, and navigation. As illustrated in Figure 2.3, by taking into consideration the overlapping sections, the four classes intent taxonomy can be further divided into 8 sub-classes (Et1-ET8). The overlapping parts indicate one search activity could be placed in multiple intent classes. To be more concrete, pure knowledge-oriented search intent (ET1) indicates the user wishes to acquire knowledge by conducting a search action. In some cases, the users might already have a mental image in mind (ET2) or wish to perform a search action to see a concrete visual example (ET3). Transaction intent means the user wants to find an image to finish a certain task, such as give a presentation (ET4). The 8-classes intent taxonomy is shown in table 2.3.

Author	Intent	Description
Lux et al.	KO	Extract knowledge
	MI	Already have a mental image
	NG	Known existence of an image except it's visual content
	TS	Find or download for a task
	KO$\cap MI$	Extract knowledge and have a mental image
	MI$\cap TS$	Have mental image, download for a task
	KO$\cap NG$	Known existence of image for knowledge acquiring
	NG$\cap TS$	known existence of image for a task

Table 2.3: 8 Taxonomy of image search intent. KO: Knowledge Orientation, MI: Mental Image, TS: Transaction, NG: Navigational.

Apart from uncovering why users query images by interviewing 10 people covering 40 different situations, Lux et al. [39; 40] studied the user intent in photo production. According to the authors, photoproduction intents can be classified based on different dimensions. First of all, photo production intents can be categorized into *individual* and *social* classes. An photo production intent is *individual*, for example, when it reminds oneself of a beautiful, funny or hilarious moment. In addition to photo production in-

tent is *public* while sharing the pictures with friends and family in order or public to tell a story. Apart from *individual* and *social* dimension, the intents can also be categorized into *affective* and *functional* classes. To this end, by cross-cuts these two dimensions, the user proposed a 4-class photo production intent taxonomy, that is *individual-functional*, *individual-affective*, *social-functional* and *social-affective*. The second paper introduced a test dataset consist of 1309 photos annotated by photographers from Flickr. The result was a intent taxonomy containing six reasons why users take pictures: *preserve a good feeling*, *preserve a bad feeling*, *share with family and friends*, *share with public*, *support a task* and *recall a specific situation*. It should be noted that the taxonomy is not mutually exclusive, which means on photo could be placed into more than one intent class.

The work closest to our own is Hanjalic et al. [22], who employed a social-web mining approach, analyzing 4512 questions posted on Yahoo Answers that contained certain terms (i.e. "find" and "video"). The questions were sorted by crowd souring workers and then, using a card sorting process similar to our own manual content analysis process, five classes were defined: *Information*, *Experience:Learning*, *Experience: Exposure*, *Affect*, and *Object*. Instead of investigating video, we investigate images. The commonality is that our work also exploits a large collection of social media to identify a set of intent classes that may not be absolutely universal, but is "good enough" to support tasks such as retrieval.

3 A TAXONOMY OF INTENT CLASSES

As was introduced in Chapter 1, a number of studies have postulated a convergence between search result diversification [1]. There is an urgent need to address the unsatisfied search result problems caused by didnot tackling *underspecified query* into considering. We propose to tackle this problem by introducing an intent taxonomy. More concretely, the intent taxonomy should go beyond the conceptual content of the image and reflect the goals of the photographer. The taxonomy is created on the basis of sub-set of a large collection of social images, namely, the YFCC100M data set [66]. We create the classes using a manual content analysis [43] approach. The intent classes are then used to label the data set. We are interested in demonstrating the ability of intent to go beyond concept detection. For this reason, we need to ensure that the data set contains images depicts a wide range of concepts, but also that there are enough images per concept. This chapter will first discuss the specifics of the dataset creation, and then describe the content-analysis procedure which was used to create the intent taxonomy. In the end, we'll answer the research questions proposed as RQ.1, that is: *Do images with the same subject matter reveal different perspectives and goals of users (user intent)?* and *Do categories that express a certain perspective/goal (i.e., user intent categories) appear across different types of subject matter (concept categories)?*

3.1 TAXONOMY CREATION

3.1.1 Data Set and Set Up

Since we need to create an image corpus that depicting a wide range of concepts that could potentially form an intent taxonomy, the first step in creating the data set is to define a list of concepts $C = \{c_1, c_2, ..., c_k\}$ that we would like to focus on. We chose the NUS-Wide concepts for several reason: this concept set is comparable to other concept sets used in the literature, it corresponds to the most frequent tags in Flickr, it includes both general and specific concepts, and finally, the concepts are of different types, including scene, object, event, program, people and graphics [10].

The concepts are used as queries to retrieve images from the YFCC100M data set with a tag-based image retrieval system. The system was built upon nearly 100 million image documents (namely YFCC100M dataset). The title and tags were then indexed with BM-25 ranking function.

We choose YFCC100M for a variety of reasons: 1. To our knowledge it is the largest public image collection that has ever been released. 2. It offers

1 Based on ambiguous queries and visual redundancy

social images associated with user information. 3. It is publicly available, which supports reproducibility. Flickr has a batch tagging function that allows users to assign tags to a set of uploaded images [70]. Since batch tagging might affect the robustness of our intent data set, we only retain a maximum of three images with the same tags. For each concept (query), we collect top-200 relevant images. If less than 200 images are returned by the query, we use the total number of images returned. Our final data set contains 15618 images.

3.1.2 Content Analysis for Taxonomy Creation

Next we turn to describe the content analysis approach by which we created the intent taxonomy. The approach involves iterative labeling by an expert annotator [2]. The expert acquired his expertise by reading the papers in Section 2, and studying the taxonomies proposed there. During the labeling process he allowed his understanding of intent to evolve guided by what he observed in the data.

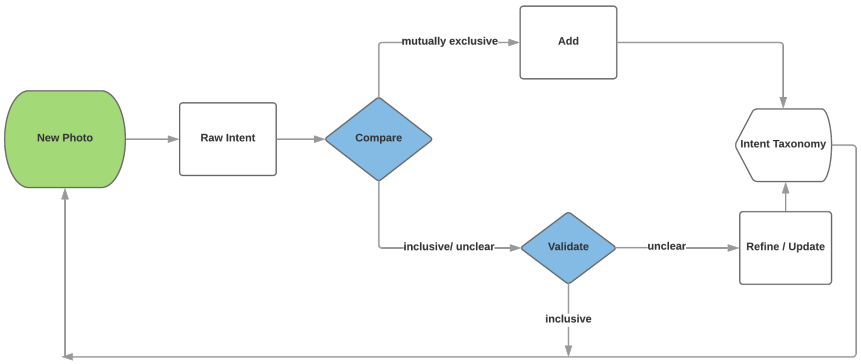

Figure 3.1: Diagram of iterative labeling process.

Specifically, our manual content analysis approach proceeds by examining images in turn and applying a preliminary intent label (see Figure 3.1). Each new image is judged as either belonging to an existing class or potentially requiring the introduction of a new class. Before introducing a new class, the annotator returns to the previous annotated images to ensure that it is not possible to accommodate the new image by updating the description of an existing class. If no existing class can be extended to accommodate the new image, a new intent class is introduced. Note that the process requires multiple iterations over the whole data set, and represents a full person month of image annotation work.

2 The author of this thesis

Figure 3.2: Example of two photos that appear similar at first glance, but ultimately land in two different intent categories, because of the difference in photographer goal.

The two images in Figure 3.2 demonstrate the process with a hypothetical example. The left image is previously labeled with "portrait" and the right image is new. The annotator examines all previous images and realizes that the right image does not fit into the "portrait" class because the picture was not taken with the same goal as a portrait. Another class, *situation_documentation* is then added to the taxonomy. The final taxonomy is shown in Table 3.2.

Here, we have used an intent taxonomy created by a single expert. In our future work, we are interested in the extent to which changes or refinements of this taxonomy impact the predictability of the intent classes, and the ability of the intent classes to improve the diversity of image retrieval results. We suspect that no single perfect taxonomy exists for all cases, but many taxonomies are probably "good enough". If this is the case, we are interested in optimizing the taxonomy to maximize both intuitiveness to users.

3.1.3 Taxonomy Validation

For the purpose of validating that intent class must be considered independently of topics, an analysis of our intent taxonomy was carried out from two perspectives: how intent classes i were shared among concept groups c and how concepts c are shared within different intent classes i. We discuss each here in turn.

In general, as can be seen in Figure 3.3, the 81 concept categories that form the basis of our dataset are distributed across various intent classes. For example, the intent classes *portrait*, *candid*, *setting*, *media_capture* and *landscape* are found associated with images from 67 out of the 81 concepts. On the other hand, some intent classes did not show such generality. The intent classes *product_presentation_by_person*, *domesticate* and *situation_documentation* are found to be associated with 8–16 concepts. On average, an intent class is associated with 46.21 different concepts. This means that each intent class is associated on average with just over half of all concepts (to be exact 56.79% of all concepts, with a standard deviation of 23.95). The results of this investigation show categories that express a certain perspective/goal (i.e., user intent categories) appear across different types of subject matter (concept categories).

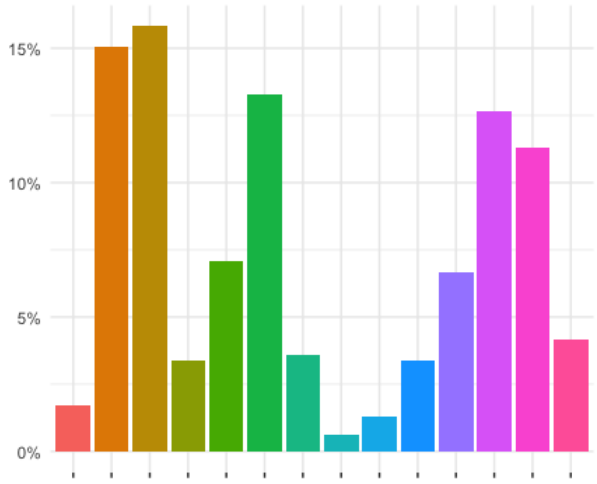

Figure 3.3: Distribution of intent classes, from left to right: art, candid, landscape, macro, media capture, portrait, product presentation, product presentation by person, situation documentation, social event private, social event public, setting, structure, wildlife.

We turn now to look at how intent classes are distributed across concept categories. On average a concept category is associated with just over 8 intent classes (to be exact: 8.42 and 2.01 standard deviation.) This also confirmed our hypothesis that images with the same subject matter (concept) usually reveal different perspectives and goals of users (user intent). These observations validate our intent taxonomy in the sense that they show that intent classes cross-cut image topic as represented by topic category, and that topics are associated with multiple intents. We note that the fact that our intent classes are effective for image search result diversification, provides further validation for our taxonomy. These experiments will be discussed in Chapter 4.

3.2 INTENT DISTRIBUTION

In order to explore the distribution of intent classes over concepts, we ranked the concepts with respect to the number of intent classes with which they are associated. The top-10 concepts are listed in Table 5.2 and the bottom-10 in Table 3.3.

Class	Description
product_presentation	photos made to demonstrate or sell an item or a product.
product_presentation_by_person	photos made to demonstrate or sell an item or a product worn by or held by a human model.
social_event_public	photos made to record social events held for the public.
social_event_private	photos made to record planned social events held for a certain group of people.
situation_documentation	photo made to document desired or undesired situation.
landscape	photos made to shows spaces within the world, sometimes vast and unending.
macro	photos made to show extreme close-up photos of very small subjects too small to be noticed.
structures	photos made to demonstrate landmarks, buildings and similar structures.
setting	photos made to depict a inanimate objects, could be either nature or man-made, reflect specific aspect of the object.
portrait	photo of a person or a group of people that captures the characteristics of the subject(s)(who are aware they are being photographed).
candid	photo of a person or a group of people that capture the characteristics of the subject(s)(who are not aware they are being photographed).
wildlife	photos capture varies forms of wildlife in their nature habitat.
media_capture	image captured to carry out other information source.
art	photos captured to demonstrate abstract, creative vision of a photographer.

Table 3.2: Taxonomy of intent classes resulting from our intent discovery process

Concept c	Num Intent Classes	Entropy
sun	12	3.16
frost	12	2.06
fish	12	3.09
tiger	11	2.80
animal	11	1.57
boat	11	2.61
earthquake	11	2.96
window	11	2.90
forest	11	2.86
fox	11	3.18

Table 3.1: Top-10 concepts ranked by the number of intent classes associated with the concept in the image collection.

Concept c	Num Intent Classes	Entropy
coral	2	0.72
running	5	2.07
dog	5	1.52
elk	6	2.01
dancing	6	2.10
statue	6	1.83
plane	6	1.81
buildings	6	2.28
whale	6	1.67
surfing	6	1.93

Table 3.3: Bottom-10 concepts ranked by the number of intent classes associated with the concept in the image collection.

We observe that concepts differ with respect to the number and distribution of associated intent classes. For some concepts, there are photos in the collection associated with a wide range of different intent classes. For other concepts, the number of intent classes is more limited. We capture these patterns with the entropy, which is reported in the tables.

More insight is provided by looking at four typical concepts in greater depth: *bird*, *building*, *fox* and *wedding* as illustration in Figure 3.4. We see that concepts are directly associated with one dominant intent class, such as *bird* (top left) and *buildings* (top right). Other concepts are associated with multiple dominant intents. For example, for *wedding*, **78%** of the photos were captured related to people (*portrait* and *candid*) or aims at recording a private social event. We noticed that for some concepts, such as *fox*, intent classes corresponded closely to what would be more conventionally considered different interpretations of an ambiguous query. In a social media context, "fox" it might be an animal, a logo, or even a band of musicians, as reflected in our data set.

To create the ground truth, we analyzed the distribution of intent classes over users. We aggregated the data set by user id and filtered out users with less than five photos, resulting in a set of 221 users. Figure 3.5 shows the entropy over the 14 intent classes with respect to each user.

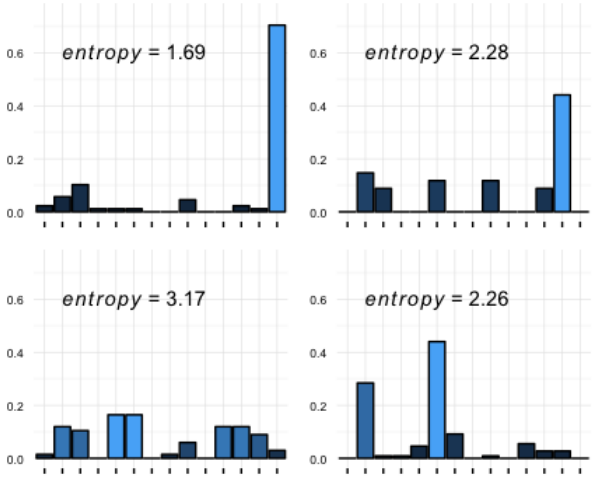

Figure 3.4: Distribution of intent classes with respect to four selected concepts plus entropy: bird (1.69), building (2.28), fox (3.17), wedding (2.26). Intent classes from left to right: art, candid, landscape, macro, media capture, portrait, product presentation, product presentation by person, setting, situation documentation, social event private, social event public, structure, wildlife.

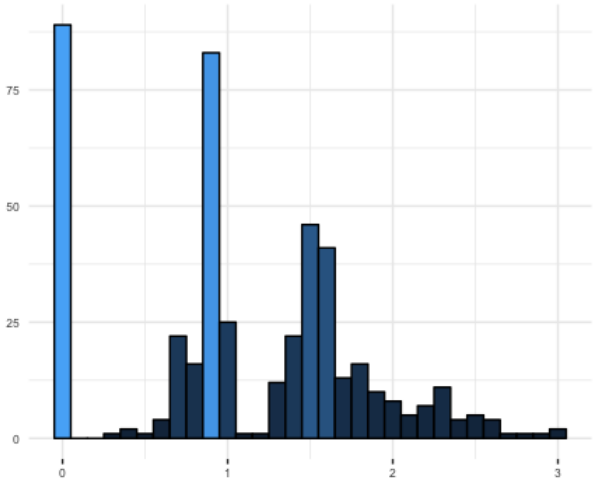

Figure 3.5: Intent entropy distributed over users, x-axis indicates entropy value, y-axis indicates number of users.

We see that the images of some users are associated with a stable set of intent classes and some have are unevenly associated with intent classes. A large number of users have one particular intent (i.e., entropy is 0). We note that this observation might be an artifact of the fact that for quite a few users the number of images per user is quite small. Overall the patterns in Figure 3.5 suggests that individual users have individual patterns of intent. We do not build further on this insight here, but only point out that in the future user-specific intent patterns could also be relevant for personalizing image search.

4 AUTOMATIC INFERENCE OF INTENT CLASSES

In the previous chapter, we have created a taxonomy of intent classes. As explained earlier, we introduced an iterative labeling process and a total number of 15618 images was labeled with 14 intent classes. Further analysis showed that images with the same subject matter (i.e. concept) reveal different goals of user intent and a user intent category appear across different types of concept categories. Before discussing how the intent taxonomy could possibly contribute to image search result diversification, it is necessary to *automate* the labeling process. In this chapter, we will introduce how state-of-the-art deep neural networks could help us automatically conduct Multimedia Content Analysis, more concretely, inference intent classes.

4.1 MODEL

First of all, we start to investigate whether a classifier can learn to predict the intent class of an image. In recent years, deep convolutional neural networks (CNNs) have shown a great progress in various visual recognition tasks [49]. Building on this progress, we adopt a transfer learning scheme. Transfer learning consists of using models that were trained for a certain task and leveraging the knowledge that they have acquired on a different, but related task [44]. In our case, we have a relatively small number of the new data set (15618 with intent labels) which is not sufficient to train a brand new network from scratch. Besides, since the intent dataset is a subset of YFCC100M data set and it's similar to the original ImageNet dataset. Thus we expect higher-level features in the ConvNet to be relevant to this intent dataset as well.

It is common knowledge that the more data an ML algorithm has access to, the more effective it can be, this is especially the case for deep neural networks since we need to learn a decently complicated function. Before we started the training process, all images were re-sized into (224,224) and we applied random horizontal flipping, choppingand re-scaling for data augmentation. The same pre-processing was performed on the test and validation set before training.

In our case, VGGNet (VGG16) [61] was adopted to extract visual features from image pixels (originally trained on ImageNet [15]). Instead of having a large number of tunable hyperparameters, VGGNet has achieved excellent performance even while used as a relatively simple pipeline.

An overview of the model architecture can be found in Figure 4.1. The architecture of VGGNet increased ConvNet depth in a fair setting (16/19 layers) and was agreed upon after using the standard configuration principle. The key aspects of "standard" can be listed as follows: First of all, all

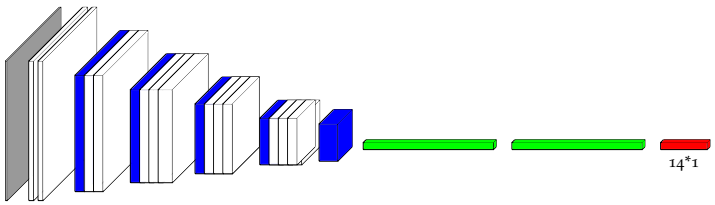

Figure 4.1: An overall architecture of VGG-16 model. Grey: input layer (image), white: conv layer, blue: pooling layer, green: fully connected layer, red: classification layer. (Note, each fully connected layer is a n by 1 dimensional)

CONV layers have a filter size of 3 by 3 with a stride of 1 and same padding. Secondly, all max-pooling layers use the same filter size of 2 by 2 with a stride of 2 pixels. Followed by a stack of CONV layers and max-pooling layers, three fully connected layers can be seen in Figure 4.1. The first two have 4096 channels each, and the last fully connected layer (going from 2048 neurons to 1000 class scores) was removed. As a result, the entire architecture of VGG16 excludes the final classification layer was used as a visual feature extractor.

We retrained a Softmax classifier using the cross-entropy Softmax loss on our image data set annotated with 14 intent classes using 70% of the intent data set. Meanwhile, 25% of the images were held out for the validation purpose and we left 5% for future analysis. This will be further discussed in the following section.

4.2 TRAINING

The first layers of the CNN network are pre-trained on ImageNet classification challenge dataset using tensorflow [1]. We learn the parameters for the rest of the layers by optimizing the categorical cross-entropy loss for the final classification layer of the network on our labeled intent data set. Each label j corresponding to an image i is represented as a 14-dimensional vector (reflects 14 intent classes defined in the intent taxonomy), where each element of y_j indicates the presence or absence of the corresponding intent class. The loss function can be written as:

$$L(\hat{y}, y) = -\sum_{j=1}^{14} y_j * log\hat{y}_j$$

and the overall cost function can then be written as:

$$J = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}_i, y_i)$$

We employ the Adam optimizer [33] for optimizing the loss objective for the network and rectified linear units (ReLU) as activation functions globally. We initialize the weights with the initializer originally proposed in [23] to

reduce vanishing gradient exploding problem. In our experiments, we use a dropout rate of 0.2 for all the projection layers to avoid over-fitting.

4.3 RESULTS AND EVALUATION

As shown in Figure 4.2, our model achieved 71% accuracy on the validation set. This performance level demonstrates that there is enough visual stability within intent classes to allow a classifier to generalize on them. We expect with a larger labeled intent dataset, model performance could be further improved.

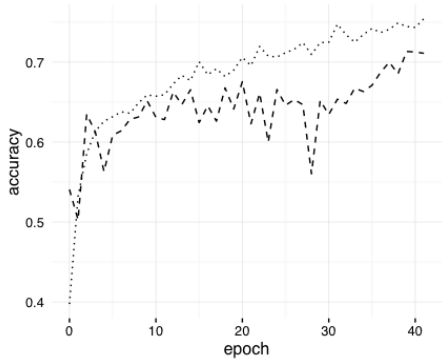

Figure 4.2: Intent class prediction: training set accuracy (dotted line) and validation set accuracy (dashed line).

Figure 4.3: Intent predicted as *landscape* from four different concepts, *road*, *zebra*, *city space*, *reflection*.

Nextwe analyze the ability of the classifier to distinguish intent classes in more depth. The data set we use here is the 5% of the labeled data (ca. 800 images) that was reserved (as mentioned above) for future analysis. Employing our trained model for intent classification, we predicted the outcomes for this small proportion of data.

As reflected in Figure 4.4, for most of the intent classes, over 70% of the images in the data set belonging to that category can be correctly predicted. However, for some particular intent classes, it is still non-trivial for our classifier to capture the patterns of the images belonging to this class. For example, for the classes *art* and *product_presentation*, our intent classifier can only 31.25% and 37.5% of the images can be identified correctly. We attribute this observation to the visual diversity of the images in these intent classes. However, we also point out, as Figure 4.4 indicates that for these two intent classes, there are fewer images in our dataset compared to other classes.

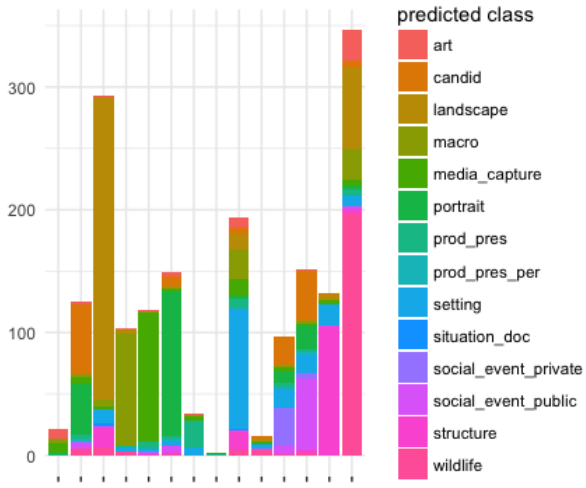

Figure 4.4: Discriminative ability of intent classes: bars are the intent classes (alphabetical order left to right). Height of the bars indicates the number of images in each class. Colors indicate class confusion.

It is also interesting to observe that certain intent classes are easily confused. This effect occurs case of *candid* and *portrait*. By definition, we differentiate *candid* and *portrait* by the awareness of the person or people appearing in the photo. Awareness is reflected in a subtle difference in direction of the gaze of the person in the image, and people judging images also might confuse these classes. Moreover, there could be multiple persons within an image (some aware while others not).

In this chapter, we have proved that we are able to build a visual classifier to interfere with the rational behind photographic act with reasonable predict accuracy. Besides, by conducting an analysis on the discriminative ability of intent classes internally, we point out for the majority of intent classes, the visual classifier has a relatively good performance against high visual variability and high visual similarity. In the following chapter, we will carry out a set of experiments on multimodal social image retrieval that show that our intent classifier is able to contribute productively to the diversification of image search results.

5 SEARCH RESULT DIVERSIFICATION

In this chapter, we will first introduce the setup of the datasets we used and discuss the evaluation framework used to interpret the results. This is followed by a set of experiments, including employing our pre-trained intent classifier in the previous Chapter 4 for image search result diversification. Since there are two basic approaches currently being adopted in research into search result diversification. One is the text-based approach. Another one is visual-textual joint search result diversification. We will investigate the performance of applying text-based search result diversification algorithm on image-based search result diversification. Apart from that, we'll also compare the performance derived from different approaches.

5.1 EXPERIMENTAL SETUP

5.1.1 Dataset and Setup

We evaluate the ability of our intent-based approach to improving the diversity of image search results using the task definition, development data set, and ground truth that was released by the MediaEval 2017 Retrieving Diverse Social Images Task [26]. The task aims at image search result diversification in the context of social media. Given a ranked list of query-related photos retrieved from Flickr using text queries, participating systems are expected to refine the results by providing a set of images that are relevant to the query and, at the same time, offer a visually diversified summary.

According to the task organizer, *relevance* and *diverse* were defined as:

Relevance: an image is considered to be relevant for the query if it is a common visual representation of the query topics (all at once). Bad quality photos (e.g., severely blurred, out of focus) are not considered relevant in this scenario.

Diverse: a set of images is considered to be diverse if it depicts different visual characteristics of the query topics and subtopics with a certain degree of complementarity, i.e. most of the perceived visual information is different from one image to another.

The development data set consists of 110 Creative Commons licensed general-purpose, multi-topic queries defined by the task organizers. Meanwhile, the test set consist of 84 queries. In Table 5.1, we provide a set of example queries.

Example Queries
accordion player
animals at zoo
birthday candle
bee on a flower
gold chain necklace
hamburger and French fries
stormy sky
sunflower field
trees reflected in water
worker in factory

Table 5.1: Example complex, general purpose and multi-concept search queries in our dataset.

Each query is represented with up to 300 Flickr photos and their associated social metadata: title, description, geo-tagging information, number of views, and number of posted comments. For each query, a list of top 50 photos that are considered to be both relevant and divers should be returned.

Data	Number of Queries	Number of Images
Dev	110	32487
Test	84	24986

Table 5.2: Statistics of the MediaEval 2016 Retrieving Diverse Image Dataset.

To encourage participation of groups from different communities such as information retrieval, multimedia and computer vision, the task organizers also provided varies types of descriptors. To be more concrete, the entire data set consist of:

1. **Visual descriptors**:
   a) Convolutional neural network based descriptor (CNN), model trained on ImageNet data set.
   b) Auto color correlogram (acc).
   c) Color and Edge Directivity Descriptor (cedd).
   d) Color Layout (cl).
   e) Edge histogram (eh).
   f) Fuzzy Color and Texture Histogram (fcth).
   g) Gabor descriptor.
   h) Joint Composite Descriptor (jcd).
   i) Pyramid of histograms of orientation gradients (phog).
   j) Scalable color.
   k) Speeded up robust features (surf).
2. **Textual Descriptors**:
   a) Word embeddings (50d) trained on Wikipedia data set.

b) Term frequency (tf).

c) Document frequency (df).

d) Term frequency - Inverse document frequency (tf-idf).

3. **Others**:

a) Photo taken date.

b) Photo taken geo-location.

c) Number of comments.

d) Number of views.

e) Original rank given by Flickr relevance algorithm.

f) Title, description and tags.

g) Credibility descriptors to estimate the quality of tag-image content relationships.

In terms of ground truth collection, all images are labeled by 17 expert annotators in terms of both relevance and diversity. According to the organizer, the data were properly distributed among the annotators such that each query was labeled by at least three different annotators. The ground truth data consists of binary relevance ground truth (rGT) and diversity ground truth (dGT). For each query and its corresponding images, the relevance is judged by aggregating results from yes/no annotations using a majority voting scheme. To create the ground truth for image search result diversity, images are grouped into clusters based on visual characteristics of the target concepts such as sub-locations, temporal information with a certain degree of complementarity. As an example, images associated with query term *accordion_player* are grouped into 18 clusters including *accordion player on boat*, *accordion player standing street art*, *accordion player on stage*, *accordion statue* etc. Note that only relevant images are annotated for diversity.

5.2 EVALUATION FRAMEWORK

A great number of studies has been to conduct on evaluation metrics for search result diversification. These metrics usually focus on rank-based evaluation of retrieved results using ground truth information such as determined by manual judgments or implicit feedback from user behavior data. Some metrics are computed at a specific position at k since top-ranked results need more attention. In the following section, we will introduce the evaluation metrics we employed during the experiment.

5.2.1 Precision, Cluster Recall and F1

With the relevance and diversity judgments, several evaluation metrics have been used in the experiment. Compared to conventional IR metrics, search result diversification employs a different set of metrics for performance judgment. Below we'll list the evaluation measurements for the search result diversification problem.

Precision@K: Precision is defined as the fraction of retrieved documents that are relevant:

$$Precision = \frac{NumRelevantItemsRetrieved}{NumItemsRetrieved}$$

In our case, we evaluate Precision on top K images. It should be noted that not all images are relevant and only relevant images have a diversity label. In other words, in our experiment, higher precision usually has a higher chance to get better diversity results.

CR@K: Cluster Recall(CR) is a measure that assesses how many different clusters from the ground truth are represented among the top K results, thus, it reflects the diversification quality of a given image result set. Cluster Recall is defined as:

$$CR = \frac{NumClustersRevrieved}{NumClusters}$$

F1@K: In a search result diversification scenario, F1 is defined as harmonic mean of Precision@K and CR@K:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{ClusterRecall}}$$

5.2.2 α-nDCG

In [11], Clake et al. proposed a framework for evaluation that systematically rewards novelty and diversity. It was developed based on the concept of *information nugget*. As described in [14], an information nugget is defined as an atomic piece of information about the target that is interesting and is not part of an earlier question in the series or an answer to an earlier question in the series. Then user's information need can be represented as a set of information nuggets $\{n_1, n_2, ..., n_m\}$. Given a user u and document d, the relatedness of user query can be predicted with:

$$P(R = 1|u, d) = 1 - \prod_{i=1}^{m}(1 - P(n_i \in u) * P(n_i \in d))$$

Given the fact that the absence of user behavior and feedback, the author assume information nuggets are independent and equally likely to be relevant, thus $P(n_i \in u)$ is able to be represented as a constant γ.

Meanwhile, the authors assume that a human assessor reads d and reaches a binary decision (0 denotes not relevant and 1 denotes relevant) regarding each nugget n using $J(d, i)$. They use α to express how confident we believe the human assessor has given the right annotation ($0 \leq \alpha \leq 1$). Then $P(n_i \in u) = \alpha$ if n_i is relevant to d, otherwise $P(n_i \in u) = 0$. Then the relevance prediction function can be rewritten as:

$$P(R = 1|u, d) = 1 - \prod_{i=1}^{m}(1 - \gamma\alpha J(d, i))$$

Considering a specific piece of information nugget appears in first $k-1$ documents, then the document d_k will provide no additional information gain to the user. Thus, the probability of the previous $k-1$ document do not contains information nugget n_i is the joint probability which can be represented as $\prod_{j=1}^{k-1}(p_{n_i} \notin d_j)$, thus the probability of user still being interested in document k is equal to the probability of document k containing information nugget n_i multiplied the probability of all previous $k-1$ documents not containing the information nugget n_i.

Finally, the relevance of kth document is relevant given user u and previous documents is equal to:

$$P(R_k = 1|u, d_1, d_2, ..., d_k) = 1 - \prod_{i=1}^{m}(1 - \gamma\alpha J(d_k, i)(1-\alpha)^{r_i,k-1})$$

where $r_i, k-1$ is the number of documents ranked up to position $k-1$ that have been judged to contain nugget n_i (equals to $\sum_{j=1}^{k-1} J(d_j, i)$), and $(1-\alpha)^{r_i,k-1}$ is the probability of the previous $k-1$ document do not contains information nugget n_i. This can be further simplified to:

$$P(R_k = 1|u, d_1, d_2, ..., d_k) = \gamma\alpha \sum_{i=1}^{m} J(d_k, i)^{r_i,k-1}$$

Finally, the information gain vector given the kth document can be defined as:

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_i,k-1}$$

As claimed by the authors, α is a hyper-parameter set to 0.5, and given the definition of information gain G, we compute nDCG using the conventional manner.

Since this version of nDCG favors novel documents that rewards novelty through the gain value, it was named as α-nDCG.

5.2.3 NDCG-IA and MAP-IA

As was explained in [2], the classical information retrieval metrics such as MAP and NDCG compute the degree of relevance given an ordering of documents. However, the ideal is bound to be different depending on the aspects the results are evaluated against when a query may be of multiple intents.

It is worth noting that, as was discussed in Chapter 2, the intent-based approach is different from our approach from different perspectives: First of all, the intent-aware framework was originally developed for text retrieval scenario, while our approach is developed for image retrieval task. Secondly, compared to the intent-based framework for text retrieval, our approach is built based on the hypothesis of *intentional framing*, which assumes that what an image is about is also reflected in how that image was taken. In our case, the document (i.e., social image) is created by the user himself. Thirdly, in the text retrieval case, such as ERR-IA, the intent is predicted based on

user query. While for intent-based image retrieval, the intent is predicted on the document to be retrieved.

Given a distribution of categories (or topics, information nuggets) of a query $P(c|q)$, for each intent of the query, we treat any document does not match that intent as irrelevant document and compute intent-dependent MAP and NDCG. Then we average intent-dependent MAP and NDCG by each intent and aggregate the results as MAP-IA and NDCG-IA. Formally, they're defined as:

$$MAP - IA = \sum_c P(c|q)MAP(q,k|c)$$
$$NDCG - IA = \sum_c P(c|q)NDCG(q,k|c)$$

5.2.4 ERR-IA

Proposed by [8], intent-aware expected reciprocal rank (ERR-IA) approach the diversification problem in the framework of intent-based ranking.

Given the fact that ERR-IA is developed base on intents modeled on user query, the author firstly introduced the *cascade model*. The cascade model assumes that the user views search results from top to bottom and at each position, the user has a certain probability of being satisfied. Once the user discovered the documents which satisfy his/her information need, the user terminates searching behavior and neglect the rest of the documents. If we assume each document has a probability of R_I which satisfy user, given a list of documents and their corresponded probability, the likelihood of the user is satisfied and stops at position r can be defined as:

$$\prod_{i=1}^{r-1}(1-R_i)R_r$$

If we define a cascade=based utility function $\phi = \frac{1}{r}$ as Expected Reciprocal Rank (again, r is the rank where the user finds the document he/she was looking for). In this case, as the top-ranked documents satisfy user's information need, i.e. $r = 1$, $\phi = 1$. On the other hand, as r goes to $+\infty$, the utility goes to 0.

Given n documents in the ranking, then ERR can be computed as:

$$ERR := \sum_{r=1}^{n} \frac{1}{r} P(user \quad stops \quad at \quad position \quad r)$$
$$= \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1}(1-R_i)R_r$$

If we use $P(t|q)$ denotes for the distribution of intents (information nuggets) given a query q, the user that interest in a topic t at position r can be re-written as:

$$\prod_{i=1}^{r-1}(1-R_i^t)R_r^t$$

Then the probability of the user stopping at position r over all topics is:

$$\sum_t P(t|q) \prod_{i=1}^{r-1} (1 - R_i^t) R_r^t$$

And ERR and be extended intent-aware form, that is ERR-IA:

$$ERR - IA = \sum_{r=1}^{n} \frac{1}{r} \sum_t P(t|q) \prod_{i=1}^{r-1} (1 - R_i^t) R_r^t$$

5.3 DIVERSIFY SEARCH RESULTS

Having discussed the data set and metrics of search result diversification evaluation, the final section of this chapter illustrate different methods for diversifying search results, including the intent-based approach, text-based approach (MRR-based & clustering based), visual-based approach and multimodal (text and visual) based approach.

As was mentioned in Section 5.1, only relevant images are annotated for diversity. Naturally, a higher relevance score is likely to leverage better ranking result in terms of both relevance and diversity. For this reason, we re-ranked the initial Flickr results using Okapi-BM25 before each of the different approaches.

5.3.1 Intent-Based Search Result Diversification

Our intent-based approach is developed to tackle to problem of "underspecified queries" for social image retrieval. We build on the assumption that user intent, i.e., the goals that users are pursuing when they take photos or search for images, has visual reflexes that can be captured by automatic visual classifiers.

As was discussed in Chapter 4, we introduced an intent-classifier using transfer learning scheme, that is able to classify the photographer's intent into 14 classes based on the visual signals. We also confirmed that the visual classifier is able to bring reasonable performance in the face of high visual variability and high visual similarity.

Our intent-based re-ranking approach is illustrated in Figure 5.2. It adopts the same general strategy of clustering and alternating images from clusters used by UPMC@MediaEval16. UPMC@MediaEval16 adopts a four-step approach. The first step is to create a refined initial ranked list by re-ranking the Flickr baseline using textual features (vector space model with tf-idf weights) with the aim of increasing precision. After that, the top N images are grouped using Hierarchical Clustering and the CNN features released with the data set. For each of the resulting sub-clusters, the images are resorted based on their position in the refined initial ranked list. Finally, the results are re-ranked by alternating images from different clusters.

There are two reasons why we believe intentional framing could be potentially helpful: 1. Intentional framing naturally reflects aspects of photos that are expressed by photographers. 2. Intentional framing could be processed in a direct, light-weight fashion without extensive computational resources.

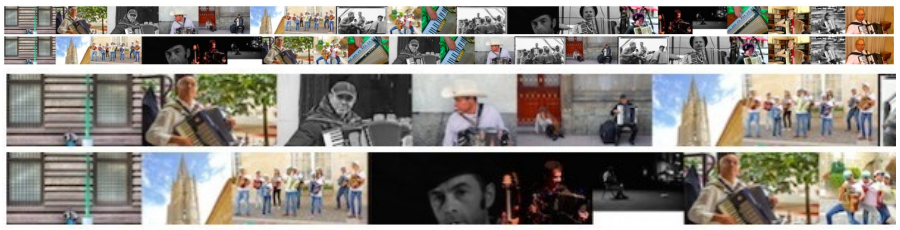

Figure 5.1: Results for the query "accordion player": the original ranking (top, first row) and the results-list that has been re-ranked based on intent classes (top, second row), and partial enlarged version (bottom).

In view of this motivation, we conducted an experiment on intent-based search result diversification.

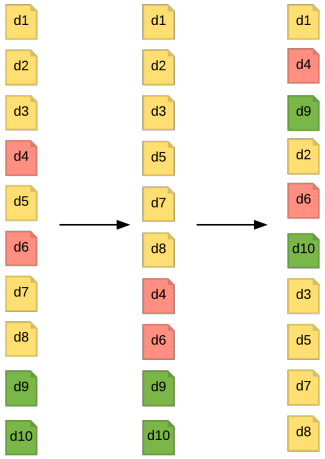

Figure 5.2: Re-ranking search result based on clusters of intents, colors of documents reflects different predicted intents. From left to right: initial ranking result, intent clusters, re-ranked result.

Our intent-based approach adopts the same strategy as "UMPC@MediaEval16" approach. However, the difference is that instead of creating clusters by applying a clustering method to the initial ranked list, we create clusters by classifying the images in intent classes and grouping together images that belong to the same class. For each query and its ranked results list returned by Flickr, we apply a three-step approach to diversify the search results. The first step in this process is to predict the intent classes of retrieved image list (i.e., Flickr baseline). Once the predictions are made, we cluster the first N results based on intent class. In our case, N equals to 50. Following this step, we pick the top-one photo without replacement for each cluster, under the assumption that new clusters reflect diversity as captured by intent. Finally, these new clusters are sorted internally based on the original rank position of the images and concatenated to create the final re-ranked list.

Figure 5.1 presents a comparison of the original retrieved result and diversified retrieved result. Given the query "accordion player", our intent-based classifier re-ranked the initial ranking results based on predicted photographer's intent: "candid", "structure", "social_event_public", "portrait" etc.

Note that the main difference between our diversification approach and UPMC@MediaEval16 is that we use intent classification. However, another difference is that we do not refine the initial ranking using text-based features. The omission of such an initial re-ranking step in our approach is not critical in the context of this paper, which focuses on diversification rather than on precision.

We point out three other aspects of the intent-based diversification approach, which make it possibly more useful compared with the rest approaches in practice. These aspects arise because, in contrast to unsupervised learning (clustering), our approach directly tries to predict pre-defined intent classes. We mention these aspects since they represent advantages of our approach over an unsupervised learning approach, even though the two achieve the same performance with respect to the diversity metric. First, intent-based diversification has the advantage of better understandability since the classification result is able to directly provide a user-interpretable indication of the reason behind the ranking. The retrieval system can provide the user with an explanation for its prioritization of search results. Second, once the model has been trained, we do not necessarily need to fine-tune the hyperparameters, i.e., the position to cut the dendrogram. Third, image labels are generated off-line at indexing time, and a clustering step at query time, which increases the system response time, is not necessary.

5.3.2 Maximal Marginal Relevance (text-based)

Originally proposed in [7], Maximal-Marginal-Relevance presented a method for combining query-relevance with information-novelty in the context of text retrieval. More concretely, in an ordered list of retrieved documents, if the latter document has a high similarity compared with top-ranked documents, we consider the latter document as a duplicate. MRR is defined as follows:

$$MRR = argmax_{d_i \in RS}[\lambda Sim_1(d_i, Q) - (1 - \lambda) max_{d_j \in S} Sim_2(d_i, d_j)]$$

where R is relevant documents in collection, S represents for selected documents. λ is the threshold employed to control the degree of relevance and diversity.

To implement the MRR function, one first needs to build a similarity matrix between every two documents. We collected all titles, descriptions, and tags associated with all images in both training set and test set and formed a corpus. We extract tf-idf features with respect to each document (i.e. title, description and tags associated with one image), then compute the pair-wise similarity between documents using cosine similarity (i.e. Sim_2). In the second step, we add the top-ranked document into our selected document list S. After then, we iterate through all the top-ranked documents given by the base retrieval system (i.e. Sim_1), and use their original relevance score minus the maximal similarity score by comparing current document d_i with each document d_j in the selected document set S.

As declared before, λ is used for control the degree between relevance and diversity. It can be clearly observed that if λ was set to 1, MMR computes the

standard relevance-ranked list. In contrast, if λ was set to 0, MMR computes a maximal diversity ranking among the documents.

In our experiment, λ was set to 0.5, the similarity function of base retrieval system (Sim_1) we used is Okapi BM25, and, as mentioned before the similarity function of pair-wise similarity (Sim_2) is cosine similarity between two documents computed based on their tf-idf features. We're interested in whether such a search result diversification algorithm originally proposed for text retrieval is suitable for the social image retrieval case.

5.3.3 K-means Clustering (text-based)

Our text-based approach involves K-means Clustering with image ranking based on BM-25 re-ranking on text descriptors, such as image title, tagsand description. The intuition behind approaches for supervised object classification is that, as it works with varies classification problems, the deep features must project data in a feature space where they are somehow separable. Thus, simple clustering algorithms such as K-means might be working well. And this hypothesis is validated in [20].

K-means is an efficient clustering algorithm provided that the number of clusters is known. The usual K-means clustering algorithm is known to require a number of small tricks to avoid common problems, such as to initialize the number of k and centroid. To get rid of these common drawbacks of K-means clustering, we initialize the number of k by dynamically conduct average the silhouette analysis. In general, silhouette coefficient is a measurement that determines how well each object lies within its cluster and how compact the clusters are. For each data point x_i, silhouette coefficient is computed with:

$$s_i = \frac{\mu_{out}^{min} x_i - \mu_{in} x_i}{max(\mu_{out}^{min} x_i, \mu_{in} x_i)}$$

where μ_{in} is the mean distance from x_i to data points in its own cluster, and μ_{out} is the mean distance from x_i to data points in its closet cluster. Then for all the data points, we evaluate clustering quality by taking the average value across all point-wise silhouette scores:

$$sc = \frac{1}{n} \sum_{i=1}^{n} s_i$$

If silhouette score close to 1, it implies all the data points are close to its own clusters while away from its closet clusters. In our experiment, we initialize a list of k (number of clusters) which range from 2 to 20 because according to our observation, the number of clusters (diversity ground truth) in our training data is usually less than 20. For each k, we compute the silhouette score, and estimate the number of k using:

$$k = argmax(sc_k)$$

During the experiment, we first re-rank the Flickr baseline with tf-idf weights. Since in this case, we are interested in how textual descriptors can be effective in image search result diversification, the most critical issue is to learn a good representation for each "short document" consisting of ti-

tle, descriptionand tags. To achieve this, we adopted the idea of *weighted word embedding aggregation* proposed by Cedric et al. [5]. More concretely, for each term associated with an image, we use its 50-dimensional word embedding vector (Word embeddings were supplied by the task organizers.). Each image is thus represented as a set of vectors. For an image with m terms, we have set of m 50-dimensional vectors. To model an image, we take the coordinate-wise maximum and minimum of the set of m vectors. We concatenate the two resulting vectors (min and max) to arrive at a 100-dimensional vector, which is our final text-based image representation. For each query, we have a set of 300 image vectors, to which we apply k-means clustering with average silhouette analysis.

5.3.4 K-means Clustering (visual-based)

Our experiment of K-means Clustering is basically the same as "UPMC@MediaEval16" approach (and our intent-based search result diversification).

Again, the first step is to create a refined initial ranked list by re-ranking the Flickr baseline using textual features (vector space model with tf-idf weights). This is followed by the feature extraction step. We employed the Convolutional neural networks (CNN) features provided by the task organizers. As was described in 5.1, the pre-trained model was trained on ImageNet dataset. We then directly applied K-means clustering on the extracted CNN features. Same as what we have introduced in the previous subsection, We employed a heuristic approach to initialize k. Specifically, we treat k as a variable and initialize $k(1, 20]$ and apply k-means clustering for n times. For each k, we evaluate clustering performance with silhouette analysis and select the best k with respect to the achieved silhouette score.

Compared to "UPMC@MediaEval16" approach, our approach replaced Hierarchical Clustering with K-means Clustering. Compared to intent-based search result diversification, our visual-based approach tuned the classification problem into an unsupervised learning problem, then images were alternated into different clusters.

5.4 RESULTS

Table 5.3 reports the results in terms of the official MediaEval 2017 evaluation metrics Precision, Cluster Recall, F1, ERR-IA and α-NDCG. In general, higher precision is usually associated with relatively higher cluster recall and F1 scores because non-relevant images have no associated diversity cluster label. This phenomenon can be clearly observed comparing k-means (visual) run and the rest of the experiments (re-ranked using tf-idf weights).

To investigate how initial ranking precision is likely to impact the cluster recall results, we selected two representative runs including k-means (re-rank+text) and intent-based approach (intent with re-rank) and plot their precision with corresponded cluster recall on Figure 5.3

Data	Approach	P@10	P@20	P@50	CR@10	CR@20	CR@50	F1@10	F1@20	F1@50
Test	k-means (visual)	0.6381	0.6601	0.6264	0.2682	0.5698	0.7538	0.4734	0.5830	0.6549
Test	k-means (rerank+text)	0.7250	0.7036	**0.6814**	**0.4479**	**0.6142**	**0.7802**	**0.5336**	0.6343	**0.7001**
Test	maximal marginal relevance (rerank+mrr)	0.7186	0.7002	0.6791	0.4341	0.6073	0.7798	0.5168	0.6221	0.6931
Test	intent (rerank+intent)	**0.7464**	**0.7262**	0.6693	0.4345	0.6125	0.7746	0.5274	**0.6462**	0.6913

Data	Approach	ERR-IA@10	ERR-IA@20	ERR-IA@50	α-NDCG@10	α-NDCG@20	α-NDCG@50
Test	k-means (visual)	0.6087	0.6015	0.6120	0.5710	0.5685	0.6137
Test	k-means (rerank+text)	0.6880	0.6729	0.6772	**0.6454**	0.6282	0.6573
Test	maximal marginal relevance (rerank+mrr)	0.6531	0.6614	0.6513	0.5328	0.6007	0.6254
Test	intent (rerank+intent)	**0.6934**	**0.6806**	**0.6835**	0.6428	**0.6325**	**0.6582**

Table 5.3: Results for search result diversification on test data evaluated by varies metrics.

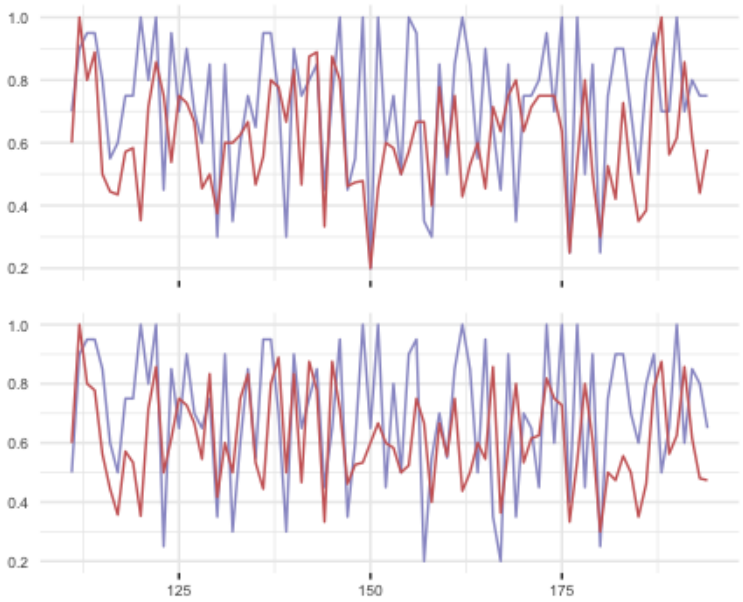

Figure 5.3: Comparison between *text-rerank+intent* (run4) (above) and *text-rerank+text run* (run2) (below) over all query id (x-axis), purple: P@20, red: CR@20.

Figure 5.3 shows that both metrics fluctuate widely with respect to different queries. This figure indicates that for most cases, a higher precision value (purple line) is associated with a higher cluster recall value (red line). We measured the Pearson coefficient between P@20 and CR@20 for *text-rerank+intent* (0.41) and *text-rerank+text* (0.35), which reveals that the intent-based approach is more sensitive to initial ranking precision. The standard deviations are comparable: $\sigma = 0.17$ for *text-rerank+text* and $\sigma = 0.18$ for *text-rerank+intent*.

Back to Table 5.3, the comparison between different measurements shows that our intent-based the approach appears to give a boost to relevance as measured by P@10, P@20, F1@20, ERR-IA@K, *alpha*-NDCG@20 and *alpha*-NDCG@50. This answered our research question **RQ3**, the intent-based classifier is able to contribute productively to social image diversification. Figure 5.3 also shows that the performance of our intent-based search result diversification varies across different queries. For some particular queries, the intent-based approach is able to achieve 100% in terms of cluster recall, while for some specific queries, such as query 150, the cluster recall could decrease to 20%. We will investigate this issue by conducting failure analysis by diving into the query-level analysis in the next chapter.

What is surprising is that the text-based representation achieves a better clustering result on the test set compared with the visual CNN representation. The text-based approach (text re-rank+text) and our intent-based strategy (text re-rank+intent) perform comparably on the test set. As can be seen in table 5.3, our k-means clustering on weighted word embedding aggregation textual features achieved the highest score in terms of CR@10, CR@20 and F1@10. Besides, our experiment using Maximal Marginal Relevance (with re-rank) also achieved a reasonable score on different metrics. This confirmed our hypothesis of **RQ4**, that is we can directly employ text-

based search result diversification algorithms, such as maximal marginal relevance for social image search result diversification.

We point out three other aspects of the intent-based diversification approach, which make it possibly more useful compared with other approaches in practice. These aspects arise because, in contrast to unsupervised learning (clustering), our approach directly tries to predict pre-defined intent classes. We mention these aspects since they represent advantages of our approach over an unsupervised learning approach, even though the two achieve a similar performance with respect to the diversity metric. First, intent-based diversification has the advantage of better understandability since the classification result is able to directly provide a user-interpretable indication of the reason behind the ranking. The retrieval system can provide the user with an explanation for its prioritization of search results. Second, once the model has been trained, we do not necessarily need to fine-tune the hyper parameters, i.e., the position to cut the dendrogram. Third, image labels are generated off-line at indexing time, and a clustering step at query time, which increases the system response time, is not necessary

In this chapter, we have introduced our experiment setup, data set, evaluation metrics and experiment results for search result diversification under a social image retrieval context. The experiment was conducted on "MediaEval Retrieving Diverse Social Images Dataset". Results reveals that our proposed intent-based search result diversification algorithm is able to contribute productively on social image search result diversification. Also, we showed that search result diversification algorithms original developed for text-based search result diversification such as Maximal-Marginal-Relevance is able to bring reasonable results as well. However, we also figured out that for some specific queries, the intent-based approach didn't perform as well as text-based approaches. An investigation of the failure analysis, together with feature modality analysis is provided in the next chapter.

6 FAILURE ANALYSIS

In the previous section, we conducted experiments on social image search result diversification with four different conditions, including intent-based approach, text-based approach (k-means and maximal marginal relevance) and visual-based approach. The experiment demonstrated that our intent-based approach is able to contribute productively to social image diversification. Meanwhile, diversifing images with textual descriptors using k-means clustering on word embeddings or conventional algorithm developed for text search result diversification can also, achieve reasonable results. In this chapter, we will conduct a query-level error analysis to investigate the rationale behind our proposed algorithms, and the relationship between the diversification results over different feature modalities.

6.1 SELECTION OF QUERIES

We have shown that higher precision is usually associated with relatively higher cluster recall by computing Pearson coefficient between P@20 and CR@20 (0.41 and 0.35 for intent-based and text-based approach respectively). Also, Figure 5.3 reveals that intent-based approach, in general, has better performance compared with other ranking algorithms, while fails on several specific cases.

To gain deeper insights, we will conduct query-level error analysis and select the most representative queries in the "MediaEval Retrieving Diverse Social Images" dataset. The key strategies of query selection can be listed as follows:

First of all, the query and corresponded images should belong to the test data set. This is because the features we used were trained on the training set, it is interesting to observe how different approaches perform on unknown examples. Apart from this, the organizer of "MediaEval Retrieving Diverse Social Image Task" released the final query-level performance with respect to various evaluation metrics on the test set.

Secondly, we will investigate the queries that have high precision associated with relative low cluster recall. As was claimed before, it is useless to look into queries with a high precision and cluster recall, simply because irrelevant images does not have a diversity ground truth. Compared to this, the queries got a relatively high precision and low recall are the most representative ones because most of the images are relevant, while these images are from limited number of labeled clusters.

Thirdly, we will only consider top 20 retrieved results. As we observed from the result, the cluster recall naturally goes up when k increases. We believe it is, for most cases, not because our diversification algorithms per-

Concept	Approach	P@5	P@10	P@20	CR@5	CR@10	CR@20
camera lens	intent	1	1	1	0.1176	0.2941	0.3529
	textual	1	1	1	0.1176	0.2941	0.3529
	visual	1	1	1	0.2500	0.3333	0.3333

Table 6.1: Query-level (camera lens) performance comparison with respect to three diversification algorithms, intent-based (re-rank+intent), text-based (re-rank+kmeans on Embeddings), visual-based (re-rank+kmeans on CNN)

form better and better, but because of the increasing number of images have a higher probability to cover images from different clusters.

It is worth mentioning that though this chapter is called *failure analysis*, we are also interested in why a specific method works on specific queries. So we will not limit ourselves to the *failure* cases, but also the *success* cases.

6.2 FAILURE ANALYSIS

Having discussed how to select the representative queries, guided by the strategies, we selected 6 queries from the "MediaEval Retrieving Diverse Social Image" task test set. In this section, we will conduct the query-level error analysis based on the evaluation result in terms of multiple metrics with respect to different search result diversification experiments, involving different feature modalities.

6.2.1 Clear Queries

The first query we selected is *camera lens*. As can be seen in Table 6.1, the accuracy of concept *camera lens* is 1, which indicates that all of the retrieved images are relevant (have the diversity ground truth labeled by annotators). However, the performance against cluster recall at different levels of k across different diversification algorithms is low. The performance of *intent-based* approach has the exact same level of diversification performance as *text-based approach*. While *visual-based* approach performs slightly better in terms of CR@5 compared with the rest approaches.

In general, as listed in [1], annotators assigned 12 general clusters with respect to *camera lens*. This means among top 20 retrieved images, different diversification algorithms is able to cover about 4 clusters. It should be mentioned that through visual-based CR20 is a little bit lower than that of intent-based and textual based, the performance of visual-based CR@5 is 2 times better. For this query, visual-based diversification approach is able to cover 3 different diversity clusters ($0.25 * 12$) among top 5 images.

We plotted a screen shot of retrieved images associated with query *camera lens* in Figure 6.1. This figure demonstrated that the retrieved social images were extremely "compact". To be more concrete, the query *camera lens* itself is a clear one which has a low level of extrinsic issue (from both ambiguity

1 *lens*(8), *old cameras*, *person taking a picture*(1), *zoom*(10),*camera and accessories*,*multiple cameras*,*camera in a case*,*camera non professional*,*professional cameras*(1),*Semi professional cameras*,*Polaroid camera*,*inside the camera*

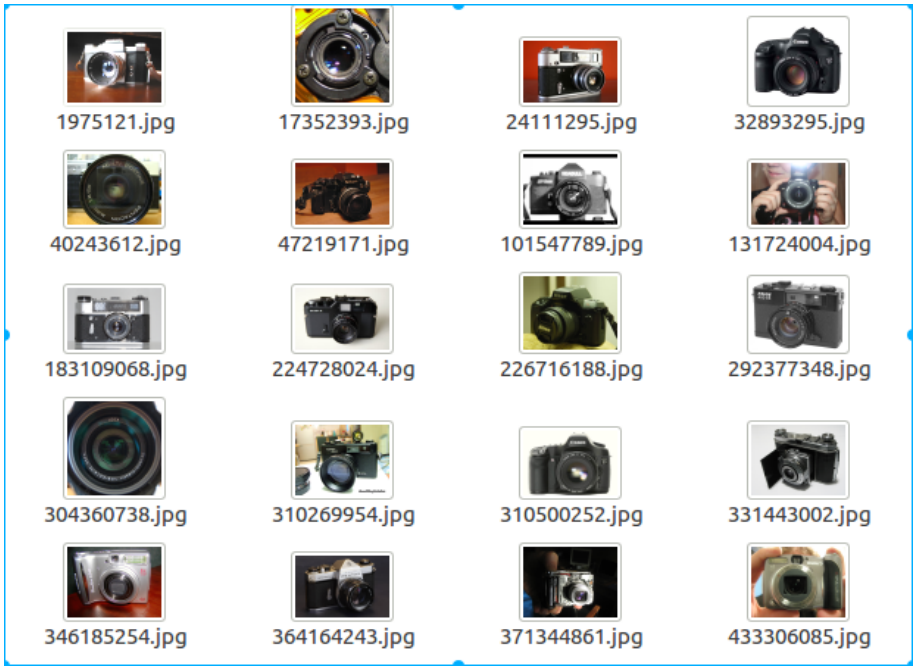

Figure 6.1: A general overview of images retrieved by concept *camera lens*

level and specification level), and it is reflected by the images the system returned back.

If we look into the intent predicted by our classifier introduced in Chapter 4, only 4 different intent classes were produced by the classifier: include *setting*, *product_presentation*, *macro* and *structure* (misclassified). If we look into the clusters given by users, it's clear to see that the diversity clusters (such as the professional level of camera, number of cameras in image etc...) simply can not be interpreted by our intent classifier. We have to admit that for this concept, we have a diversity cluster mis-match issue.

To summarize, given the fact that the query is relatively clear, and diversity clusters given by user mismatch our intent taxonomy, the intent-based approach has a low diversification performance on this query.

A similar issue can be observed on the query *students in the classroom*. As can be found in Table 6.2, compared to the previous concept, the cluster recall is a little bit higher. In this case, a total number of 19 clusters were given by annotators[2]. This indicates that in general, around 9-11 diversity clusters can be covered by three different approaches among top 20 images.

Again, we added a screen shot to demonstrate the retrieved images associated with query *student in the classroom* (Figure 6.2). If we look into these images, the majority of them involves a group of people studying in a room, some of the images only shows the classroom (might be irrelevant) itself.

In this case, our intent-based classifier does not works well, since 65% (13 out of 20) of the images were predicted as belonging to the same intent

2 student side view, artistic presentation, reading, kindergarten, primary school, teaching, conference, class photography, presentation, writing, computer class, university, secondary school, black and white, tess india, army, interacting, attentive, praying

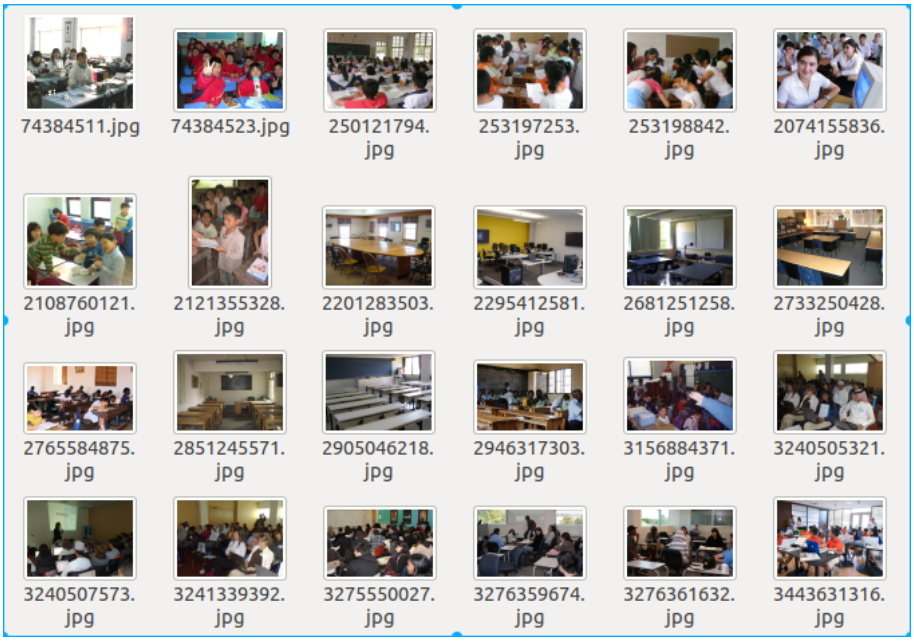

Figure 6.2: An overview of images retrieved by concept *student in the classroom*

class: *social_event_private*. In our taxonomy design, *social_event_private* were defined as *photos made to record planned social events held for a certain group of people*, which is correct in terms of classification, but not meaningful for diversification (because the predicted intent distribution is skewed). Apart from *social_event_private*, 3 retrieved images were predicted as *candid*, 2 were predicted as *portrait*, 2 were predicted as *setting*.

Concept	Approach	P@5	P@10	P@20	CR@5	CR@10	CR@20
students in the classroom	intent	1	0.7	0.75	0.2632	0.3684	0.5263
	textual	1	0.7	0.75	0.2857	0.3571	0.5000
	visual	1	0.9	0.85	0.2857	0.4211	0.6316

Table 6.2: Query-level (students in the classroom) performance comparison with respect to three diversification algorithms, intent-based (re-rank+intent), text-based (re-rank+kmeans on Embeddings), visual-based (re-rank+kmeans on CNN)

Compared with the intent-based approach, our text-based approach achieve the same level performance across different k on different metrics. The performance of visual-based approach on CR@20 is slightly better than both intent-based approach and text-based approach, 2 more diversity clusters were returned.

Similar to *camera lens*, *student in the classroom* is a clear query with less extrinsic issues. In this case, our predicted intent classes were too general to distinguish the diversity clusters (such as activities of students, age of students). Thus our intent classifier failed to diversify images with respect to this query.

6.2.2 Underspecified Queries

Another query we examined is *bow and arrow*. It is clearly that *bow and arrow* is a query with low level of ambiguity, however, it might be associated with various latent aspects. Before we dive into its different latent aspects by manually look at retrieved images, let's first look at Table 6.3. Given 16 diversity clusters annotated by users[3], all of our diversification approaches can bring reasonable performance regards to CR@5. In this case, 3 different diversity clusters can be discovered by all three methods among top 5 images. What's more, our intent-based approach is able to leverage better performance with respect to CR@10 and CR@20. To be more concrete, at the level of CR@20, intent-based search result diversification is able to retrieve 10 diversity clusters out of 16 diversity clusters among top 20 retrieved images, which is better than textual-based approach (9) and visual-based approach (8).

Concept	Approach	P@5	P@10	P@20	CR@5	CR@10	CR@20
bow and arrow	intent	0.8	0.8	0.8	0.1875	0.3750	0.6250
	textual	0.8	0.8	0.85	0.1875	0.3125	0.5625
	visual	0.8	0.8	0.8	0.1875	0.2500	0.5000

Table 6.3: Query-level (bow and arrow) performance comparison with respect to three diversification algorithms, intent-based (re-rank+intent), text-based (re-rank+kmeans on Embeddings), visual-based (re-rank+kmeans on CNN)

An example of predicted intent classes can be seen in Figure 6.3, as we can percept from the images, all of these 6 samples are exact *bow and arrow* image, however, in a social image context, the concept is associated with multiple latent aspects (which did not specified within the query). For example, by taking the first image, the photographer wants to demonstrate *bow and arrow* as landmarks, buildings and similar structures.

The nature of the query (low level of ambiguity together with multiple latent aspect) makes our intent-based approach easier to create diversify clusters. If we compare our intent-based prediction with labels given by human annotators, we'll figure out that: 1. Our intent-based prediction can clearly distinguish several clusters assigned by the annotators. For example, *bow and arrow monument* and *bow and arrow statue* usually can be predicted as "structure", the images related to man or women using *bow and arrow* usually can be predicted as "candid" or "portrait". Images related to kid can be predicted as "social_event_public" since there involves a group of person (usually a coach teaching kids use bow and arrow). Meanwhile, cluster *toy bow and arrow* and *bow and arrow alone* were predicted as "setting" since these images were made to depict an inanimate object, could be either nature or man-made, reflect specific aspect of the object. 2. Through our intent taxonomy cannot perfectly match the diversity clusters labeled by users (it can never be), it is robust enough to create clusters to diversify social images for this query.

3 bow and arrow monument, Statue using bow and arrow, toys using bow and arrow, bow and arrow as decoration, ancient bow and arrow hunter, bow and arrow guard, costume with bow and arrow, nowaday bow and arrow hunter, fictional computer character using bow and arrow, painting containing a bow and arrow, bow and arrow alone, kid using bow and arrow, man

Figure 6.3: Concept *bow and arrow* predicted as different intent classes, (from left to right, up to bottom) structure, candid, portrait, social_event_public, setting, landscape.

Concept	Approach	P@5	P@10	P@20	CR@5	CR@10	CR@20
girl reading	intent	0.8	0.8	0.85	0.4444	0.7778	0.8889
	textual	0.8	0.8	0.85	0.4444	0.7778	0.7778
	visual	0.8	0.8	0.8	0.3333	0.4444	0.7778

Table 6.4: Query-level (girl reading) performance comparison with respect to three diversification algorithms, intent-based (re-rank+intent), text-based (re-rank+kmeans on Embeddings), visual-based (re-rank+kmeans on CNN)

We also examined a similar query named *girl reading*. In this case, only 9 clusters were annotated by users[4]. In Table 6.6, we listed the performance comparison on *girl reading* with respect to different metrics. Since for this query, we have less diversity clusters, our CR@K naturelly starts at a high value. 2 out of 5 diversity clusters can be discovered by visual based approach, while 3 for textual based approach and intent-based approach at CR@5. Eventually, among top 20 retrieved images, our intent-based approach is able to retrieve 8 out of 9 diversity clusters, 7 for textual based approach and visual based approach.

We picked out several sample top-ranked images which can be found in Figure 6.4. As we can see from the images (from up to bottom, then left to right), these images illustrates girl reading book in the forest, cartoon of girl reading, girl reading (didnot aware she was captured by the photographer), a close-up face of girl reading, a girl reading a book (statue) and painting of a girl reading. It's clear that *girl reading* have relatively low degree of ambiguity, however, it could also associated with various reading context or

targeting with bow and arrow, women targeting with bow and arrow, targeting with bow and arrow close-up, bows and arrows clothes store.

4 Fake, Black and white, Paiting and statues, ,Kids close up, Kids far up, Adults close up, Adults far up, Image filters, Nature

Figure 6.4: Concept *girl reading* predicted as different intent classes, including landscape, media capture, candid, portrait, setting, art.

other latent aspects. Our intent taxonomy, also produced a relatively high level of overlap with human annotations.

6.2.3 Ambiguous Query

The query *sea horse* is a good illustration of ambiguous query. As one of Chapter 1, extrinsic diversity was defined as diversity as uncertainty about the information need, and it can be further grouped into two sub-groups, that is "ambiguous queries" and "underspecified queries". In the first case, query might refer to different interpretations. By analyzing a number of retrieved results given the query "sea horse", we figured out that in the most of the cases, it is hard to interpret the user's intent. As was illustrated in Figure 6.5, the majority of images retrieved given the query "sea horse" can be categorized as three clusters: seahorse (small marine fishes), horses running along the sea shore or a type of military helicopter named "seahorse".

Given the retrieved images, the first cluster of the retrieved images (seahorse as small marine fishes) were predicted as macro (photos made to show extreme close-up photos of very small subjects too small to be noticed) or wildlife (photos capture various forms of wildlife in their nature habitat) by our intent-based classifier. In most cases, the pictures indicates horses run along with the sea shore were predicted as landscape. Meanwhile, the images related to helicopter were predicted as setting, that is photos made to depict a inanimate objects, could be either nature or man-made, reflect specific aspect of the object.

The performance comparison results can be seen in Table 6.5. The result reveal that the intent-based approach and the visual-based approach performs better than that of text-based approaches. For the top 5 results, 4

Figure 6.5: Concept *Sea horse* predicted as different intent classes, (from left to right,) macro, landscape, setting.

clusters can be retrieved among 8 clusters annotated by the user. However, when increasing the k to 10, both intent-based approach and visual based approach failed to discover new clusters. Finally, if we look at the cluster recall among top 20 images, both of intent-based approach and visual-based approach is able to return 6 out of 8 image clusters. The two missing clusters of our intent-based approach were "Gray scale" and "multiple". This is reasonable since our intent taxonomy was not designed to detect such attributes, such as color of the image or number of objects in the image. In this case, it is similar as the query called "camera lens", because there is a group annotated by the users called "multiple lens".

Concept	Approach	P@5	P@10	P@20	CR@5	CR@10	CR@20
	intent	1.0	0.7	0.8	0.5000	0.5000	0.7500
sea horse	textual	1.0	0.8	0.8	0.0.375	0.5000	0.6250
	visual	1.0	0.8	0.8	0.5000	0.5000	0.7500

Table 6.5: Query-level (sea horse) performance comparison with respect to three diversification algorithms, intent-based (re-rank+intent), text-based (re-rank+kmeans on Embeddings), visual-based (re-rank+kmeans on CNN)

6.2.4 Incorrect Prediction Failure Case

Due to the fact that our intent-based classifier was trained on a relatively small amount of data set, for certain query it fails to make prediction because these images has never been appeared in our training data. An example is the retrieved images given query *beer bottles*. For this particular case, our intent-based classifier fails to make predictions. As an result, the images demonstrate beer bottle on a product shelf sometimes were predicted as "portrait" or "social_event_public" or even "structure" while "product_presentation" or at least "setting" is expected. We believe this issue can be partially resolved by adding more training data together with correct intent labels.

Concept	Approach	P@5	P@10	P@20	CR@5	CR@10	CR@20
beer bottles	intent	0.8	0.9	0.95	0.2 (iv)	0.4 (iv)	0.6 (iv)
	textual	0.8	0.9	0.95	0.2	0.6	0.8
	visual	1	1	1	0.4	0.5	0.7

Table 6.6: Query-level (beer bottles) performance comparison with respect to three diversification algorithms, intent-based (re-rank+intent), text-based (re-rank+kmeans on Embeddings), visual-based (re-rank+kmeans on CNN), *iv* means invalid since most of the labels were incorrect.

Apart from the issue of amount of training data. We have used an intent taxonomy created by a single expert. In our future work, we are interested in the extent to which changes or refinements of this taxonomy impact the predictability of the intent classes, and the ability of the intent classes to improve the diversity of image retrieval results. We suspect that no single perfect taxonomy exists for all cases, but many taxonomies are probably "good enough". If this is the case, we are interested in optimizing the taxonomy to maximize both intuitiveness to users and the performance of the intent classifier.

6.3 SUMMARY

In the previous section, we conducted error analysis by selecting 6 representative queries from three target groups: clear queries, ambiguous queries and underspecified queries.

For clear queries, we find out that the visual-based approach tends to achieve better performance than the intent-based approach and the text-based approach. In this case, intent-based approach might fail due to several reasons: First of all, there exist a mismatch between our intent taxonomy and anotated intent classes. For example, the annotated clusters *old camera*, *professional camera*, *semi-professional camera* or *multiple lens* given query *camera lens* can not be inferenced by our intent classifier. This is because our intent-based approach is designed to infere the intent of a photographic act, instead of predicting object attributes. Secondly, we observed that for several clear queries, the predicted intent classes are likely to be skewed and lies into one specific intent class. This might post a negative impact on our final re-ranking list since we clustering and alternating images from clusters.

For underspecified queries, intent-based approach is able to achieve better performance compared to that visual-based appraoch and text-based approach. Though the query has low level of ambiguity, the retrieved results might have various context information. For example, *bow and arrow* might be referred to as an object, or activities associated with human-beings. Our intent taxonomy tends to have a greater overlap with annotated clusters, and it is reflected on our query-level performance results: intent-based appraoch is able to cover more clusters than that of text and visual based approaches.

For ambiguous queries, our intent-based approach seems to bring comparable performance with visual-based approach, slightly better than that of text-based approach. It should be noted that the "MediaEval Retrieving Diverse Social Image" dataset is not a fair dataset to look into. Since the

provided queries are, mostly, contains multiple words and specific to certain interpretation.

Apart from these different query types, we observed on several queries, our intent-based approach fails to create a diverse ranking result. This is due to the issue of insufficient training samples. As was described in Chapter 4, our intent-based classifier employs transfer learning which uses a total number of 19618 images for training and validation. While during error analsis, we do observe the images associated with several queries, such as *beer bottles*, has never been labeled in our intent dataset. As the result, our intent-based classifier fails to make predictions on these images. Thus the re-ranked list is not valid anymore.

7 CONCLUSION AND OUTLOOK

7.1 CONCLUSION

In the thesis, we proposed a novel method, namely an intent-based approach, for social image search result diversification. The underlying assumption is that the visual appearance of social images is impacted by the underlying photographic act, i.e., why the images were taken. We believe the goals of the photographer provide a simple, easily understandable explanation for the differences observed between photos. Better understanding of the rationale behind the photographic act could potentially benefit social image search result diversification.

To investigate this idea, we employ a manual content analysis approach to create a taxonomy of intent classes. The intent taxonomy and labeled data set were produced by an expert annotator, who examined each image in turn. Finally, we produced a 14 classes intent taxonomy. We proved that there are noticeable patterns in social images that go beyond the conceptual content of the image and reflect the goals of the photographer. We discovered that images with the same subject matter reveal different perspectives and goals of users, also, categories that express a certain user intent appear across different types of subject matters.

We adopt a conventional transfer learning scheme to predict the intent class of an image. Our experiments show that a CNN-based neural network classifier is able to capture the visual difference between the classes in the intent taxonomy. We cluster images of the Flickr baseline based on predicted intent class and generate a re-ranked list by alternating images from different clusters. Our results reveal that, compared to conventional diversification strategies, intent-based search result diversification is able to bring a considerable improvement in terms of cluster recall with several extra benefits.

Since the text-based approach employs associated text or tags as information cues to convert a multimedia document retrieval problem into a conventional text retrieval scheme, and, image retrieval can benefit from years of research experience from text retrieval, we also examined text-based search result diversification algorithms on image search result diversification task. Our experimental result reveals that Maximal Marginal Relevance (MMR) original proposed for diversify textual documents is able to bring reasonable performance on image search result diversification task.

Our failure analysis demonstrated that our intent-based approach, in general, is more effective that visual-based approach and text-based approach, this is especially the case for *underspecified queries* (query associated with multiple latent aspects). However, due to several reasons, the intent-based search result diversification approach might fail in the case of *clear queries*.

First of all, such user queries with low level of ambiguity usually tend to have skewed intent categories based on predicted intent class. Secondly, when having retrieved results given clear queries, annotators is likely to group the image clusters based on image attributes, such as color, number of objects in the image, which is not reflected by our intent taxonomy. Apart from *clear queries*, our intent-based approach might fail on queries with insufficient training data.

7.2 OUTLOOK

7.2.1 Selection or Combination of Diversification Strategies

By conducting error analysis, we observed that visual-based search result diversification algorithm works well on *clear queries*, while intent-based approach achieved better performance in terms of *underspecified queries*. Previously, a lot of research effort has been put on quantifying query ambiguity [12]. It is interesting to investigate the relationship between query ambiguity degree and search result diversification strategy. In the thesis, we conduct experiments for image search result diversification over multiple single algorithms, it is promising to discover the strategy to stacking multiple diversification approaches to produce a robust re-ranked list.

7.2.2 Optimize Intent Taxonomy

Our study has contributed to enhancing our understanding of intentional framing. The potential of "why" dimension has been established, and future research can be undertaken to explore how intentional framing can further enhance image retrieval.

Here, we have used an intent taxonomy created by a single expert. In our future work, we are interested in the extent to which changes or refinements of this taxonomy impact the predictability of the intent classes, and the ability of the intent classes to improve the diversity of image retrieval results. We suspect that no single perfect taxonomy exists for all cases, but many taxonomies are probably "good enough". If this is the case, we are interested in optimizing the taxonomy to maximize both intuitiveness to users and the performance of the intent classifier.

7.2.3 Explore User-Specific Intent Pattens

We noticed that the images of some users are associated with a stable set of intent classes and some have are unevenly associated with intent classes. A large number of users have only one particular intent (i.e., entropy is 0). We note that this observation might be an artifact of the fact that for quite a few users the number of images per user is quite small. In the future, better understanding user-specific intent patterns could also be relevant for personalizing image search.

7.2.4 Multimedia Understanding for More Tasks

Overall, this study strengthens the idea of indexing images with respect to dimensions of human perception could be important for image retrieval. However, the recognition of tangible properties of data, such as objects and scenes, have overwhelmingly covered the spectra of applications in multimedia. We believe, better understanding of social, affective and subjective attributes, could be beneficial for varies tasks, such as visualization and social media analysis.

BIBLIOGRAPHY

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*, pages 265–283, 2016.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 5–14, 2009.

[3] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 971–980, 2007.

[4] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, 1992.

[5] C. D. Boom, S. V. Canneyt, T. Demeester, and B. Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, 2016.

[6] B. Boteanu, M. G. Constantin, and B. Ionescu. LAPI @ 2017 retrieving diverse social images task: A pseudo-relevance feedback diversification perspective. In *Working Notes Proceedings of the MediaEval 2017 Workshop co-located with the Conference and Labs of the Evaluation Forum (CLEF 2017), Dublin, Ireland, September 13-15, 2017.*

[7] J. Carbinell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum*, 51(2):209–210, 2017.

[8] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S. Wu. Intent-based diversification of web search results: metrics and algorithms. *Inf. Retr.*, 14(6):572–592, 2011.

[9] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 429–436, 2006.

[10] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from national university of singapore.

In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*, 2009.

[11] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 659–666, 2008.

[12] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 104–109, *San Francisco, CA, USA*, 2002. Morgan Kaufmann Publishers Inc.

[13] S. J. Cunningham and D. M. Nichols. How people find videos. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2008, Pittsburgh, PA, USA, June 16-20, 2008*, pages 201–210, 2008.

[14] H. T. Dang, J. J. Lin, and D. Kelly. Overview of the TREC 2006 question answering track 99. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14-17, 2006*, 2006.

[15] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[16] M. J. Eppler and J. Mengis. The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The Information Society*, 20(5):325–344, 2004.

[17] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 49–56, 2004.

[18] R. Fidel. The image retrieval task: implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia*, 3:181–199, 1997.

[19] É. Gaussier. Statistical language models for information retrieval - chengxiang zhai, morgan & claypool, 2008; xiii+125 pp, ISBN 978-1-59829-590-0. *Computational Linguistics*, 36(2):279–281, 2010.

[20] J. Guérin, O. Gibaru, S. Thiery, and E. Nyiri. CNN features are also great at unsupervised classification. *CoRR*, abs/1707.01700, 2017.

[21] A. Hanjalic. Multimedia search: From relevance to usefulness. *IEEE MultiMedia*, 22(1):2–3, 2015.

[22] A. Hanjalic, C. Kofler, and M. Larson. Intent and its discontents: the user at the wheel of the online video search engine. In *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*, pages 1239–1248, 2012.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034, 2015.

[24] N. A. V. House. Flickr and public image-sharing: distant closeness and photo exhibition. In *Extended Abstracts Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 2717–2722, 2007.

[25] W. H. Hsu, L. S. Kennedy, and S. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, October 23-27, 2006*, pages 35–44, 2006.

[26] B. Ionescu, A. Gînsca, B. Boteanu, M. Lupu, A. Popescu, and H. Müller. Div150multi: a social image retrieval result diversification dataset with multi-topic queries. In *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016, Klagenfurt, Austria, May 10-13, 2016*, pages 46:1–46:6, 2016.

[27] B. Ionescu, A. Gînsca, M. Zaharieva, B. Boteanu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2016: Challenge, dataset and evaluation. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016.*, 2016.

[28] B. Ionescu, A. Popescu, A. Radu, and H. Müller. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools Appl.*, 75(2):1301–1331, 2016.

[29] B. Ionescu, A. Popescu, A.-L. Radu, and H. Müller. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools and Applications*, 75(2):1301–1331, 2016.

[30] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, 2000.

[31] C. Jorgensen. *Image Attributes: An Investigation*. PhD thesis, Syracuse, NY, USA, 1995.

[32] T. Kindberg, M. Spasojevic, R. Fleck, and A. Sellen. The ubiquitous camera: an in-depth study of camera phone use. *IEEE Pervasive Computing*, 4(2):42–50, 2005.

[33] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[34] C. Kofler, M. Larson, and A. Hanjalic. User intent in multimedia search: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 49(2):36:1–36:37, 2016.

[35] C. Kofler and M. Lux. Dynamic presentation adaptation based on user intent classification. In *Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, October 19-24, 2009*, pages 1117–1118, 2009.

[36] T. Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[37] Y. Liu, D. Xu, I. W. Tsang, and J. Luo. Textual query of personal photos facilitated by large-scale web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):1022–1036, 2011.

[38] M. Lux, C. Kofler, and O. Marques. A classification scheme for user intentions in image search. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume, Atlanta, Georgia, USA, April 10-15, 2010*, pages 3913–3918, 2010.

[39] M. Lux, M. Kogler, and M. del Fabro. Why did you take this photo: a study on user intentions in digital photo productions. In *Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access*, pages 41–44. ACM, 2010.

[40] M. Lux, M. Taschwer, and O. Marques. A closer look at photographers' intentions: a test dataset. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia, CrowdMM@ACM Multimedia 2012, Nara, Japan, October 29, 2012*, pages 17–18, 2012.

[41] M. Lux, M. Taschwer, and O. Marques. A closer look at photographers' intentions: a test dataset. In *Proceedings of the ACM Multimedia 2012 workshop on Crowdsourcing for Multimedia*, pages 17–18. ACM, 2012.

[42] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, May 2006.

[43] K. A. Neuendorf. *The Content Analysis Guidebook*. Sage, 2016.

[44] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[45] L. Peng, Y. Bin, X. Fu, J. Zhou, Y. Yang, and H. T. Shen. Cfm@mediaeval 2017 retrieving diverse social images task via re-ranking and hierarchical clustering. In *Working Notes Proceedings of the MediaEval 2017 Workshop co-located with the Conference and Labs of the Evaluation Forum (CLEF 2017), Dublin, Ireland, September 13-15, 2017.*, 2017.

[46] X. Qian, Y. Xue, X. Yang, Y. Y. Tang, X. Hou, and T. Mei. Landmark summarization with diverse viewpoints. *IEEE Trans. Circuits Syst. Video Techn.*, 25(11):1857–1869, 2015.

[47] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43(2):46–52, 2009.

[48] F. Radlinski and S. T. Dumais. Improving personalized web search using result diversification. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 691–692, 2006.

[49] W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017.

[50] J. Renders and G. Csurka. NLE@MediaEval'17: Combining cross-media similarity and embeddings for retrieving diverse social images. In *Working Notes Proceedings of the MediaEval 2017 Workshop), Dublin, Ireland, September 13-15, 2017.*, 2017.

[51] M. Riegler, M. Larson, M. Lux, and C. Kofler. How 'how' reflects what's what: Content-based exploitation of how users frame social images. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 397–406, 2014.

[52] S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[53] S. E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304, 1977.

[54] S. E. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[55] A. Rorissa. A comparative study of Flickr tags and index terms in a general image collection. *Journal of the American Society for Information Science and Technology*, 61(11):2230–2242, 2010.

[56] S. Rudinac, A. Hanjalic, and M. Larson. Finding representative and diverse community contributed images to create visual summaries of geographic areas. In *Proceedings of the 19th International Conference on Multimedia 2011, Scottsdale, AZ, USA, November 28 - December 1, 2011*, pages 1109–1112, 2011.

[57] M. Sanderson. Christopher d. manning, prabhakar raghavan, hinrich schütze, *Introduction to Information Retrieval*, Cambridge University Press 2008. *Natural Language Engineering*, 16(1):100–103, 2010.

[58] R. L. T. Santos, C. MacDonald, and I. Ounis. Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, 2015.

[59] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):754–766, 2011.

[60] O. Seddati, N. Ben-Lhachemi, S. Dupont, and S. Mahmoudi. UMONS @ MediaEval 2017: Diverse social images retrieval. In *Working Notes Proceedings of the MediaEval 2017 Workshop co-located with the Conference*

and Labs of the Evaluation Forum (CLEF 2017), Dublin, Ireland, September 13-15, 2017., 2017.

[61] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[62] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[63] M. Soleymani, M. Riegler, and P. Halvorsen. Multimodal analysis of image search intent: Intent recognition in image search from user behavior and visual content. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, pages 251–259, 2017.

[64] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, October 23-27, 2006*, pages 707–710, 2006.

[65] A. Sun, S. S. Bhowmick, N. Nguyen, K. Tran, and G. Bai. Tag-based social image retrieval: An empirical evaluation. *Journal of the Association for Information Science and Technology*, 62(12):2364–2381, 2011.

[66] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016.

[67] R. H. van Leuken, L. G. Pueyo, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 341–350, 2009.

[68] R. Veltkamp, H. Burkhardt, and H.-P. Kriegel. *State-of-the-art in content-based image and video retrieval*, volume 22. Springer Science & Business Media, 2013.

[69] B. Wang and M. Larson. Exploiting visual-based intent classification for diverse social image retrieval. In *Working Notes Proceedings of the MediaEval 2017 Workshop co-located with the Conference and Labs of the Evaluation Forum (CLEF 2017), Dublin, Ireland, September 13-15, 2017.*, 2017.

[70] M. Wang, B. Ni, X. Hua, and T. Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*, 44(4):25:1–25:24, 2012.

[71] M. Wang, K. Yang, X. Hua, and H. Zhang. Towards a relevant and diverse search of social images. *IEEE Trans. Multimedia*, 12(8):829–842, 2010.

[72] X. Wang, D. Chakrabarti, and K. Punera. Mining broad latent query aspects from search sessions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 867–876, 2009.

[73] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Image and Video Retrieval, Second International Conference, CIVR 2003, Urbana-Champaign, IL, USA, July 24-25, 2003, Proceedings*, pages 238–247, 2003.

[74] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 183–192, 2010.

[75] C. Zhai, W. W. Cohen, and J. D. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *SIGIR Forum*, 49(1):2–9, 2015.

[76] C. Zhai and J. D. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.

[77] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *SIGIR Forum*, 51(2):268–276, 2017.

[78] C. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 22–32, 2005.

Beyond Concept Detection: The Potential of User Intent for Image Retrieval

Bo Wang[1], Martha Larson[1,2]
[1]Delft University of Technology, Delft, Netherlands
[2]Radboud University, Nijmegen, Netherlands
b.wang-6@student.tudelft.nl,m.a.larson@tudelft.nl

ABSTRACT

Behind each photographic act is a rationale that impacts the visual appearance of the resulting photo. Better understanding of this rationale has great potential to support image retrieval systems in serving user needs. However, at present, surprisingly little is known about the connection between *what* a picture shows (the literally depicted conceptual content) and *why* that picture was taken (the photographer intent). In this paper, we investigate photographer intent in a large Flickr data set. First, an expert annotator carries out a large number of iterative intent judgments to create a taxonomy of intent classes. Next, analysis of the distribution of concepts and intent classes reveals patterns of independence both at a global and user level. Finally, we report the results of experiments showing that a deep neural network classifier is capable of learning to differentiate between these intent classes, and that these classes support the diversification of image search results.

CCS CONCEPTS

• **Information systems → Multimedia information systems**; **Information retrieval**; **Information retrieval diversity**; **Image search**; *Information systems applications*; *Query intent*; *Personalization*; *Rank aggregation*;

KEYWORDS

Multimedia retrieval; user intent; multimedia indexing; diversification;

1 INTRODUCTION

Today's challenges in the area of multimedia content analysis and retrieval have attracted enormous attention from both academia and industry. By employing techniques for processing and analyzing multimedia content, we have successfully brought automatic recognition of the literally depicted content of images (e.g., object recognition, scene labeling) to the next level [5].

However, because researchers have focused nearly exclusively on the literally depicted content (the "what" dimension), the research field of multimedia content analysis have overlooked opportunities to better understand the reasons for which people take photos in the first place (the "why" dimensions).

Figure 1: Example search results for the query "sailing". The intent (goal) of the photographer is different for each picture. Plausible characterizations of these goals are: provide an overall impression of the space in the wider world where sailing takes place (top left), depict a sailboat as an object (top right), capture a portrait of someone sailing (bottom left), capture information on sailing from another media source (bottom right).

1.1 The Potential of Intent

The area of multimedia information retrieval has seen exceptions to this tendency towards "what" at the cost of "why", which include recent work on user search intent [11, 15, 18]. The purpose of this paper is to build on previous work and to provide concrete, large-scale substantiation of the importance of moving multimedia analysis beyond concept detection, and the potential of techniques that automatically detect visual classes related to user intent to improve image retrieval. Contrasts in user intent are illustrated in Fig. 1, which shows four images uploaded to Flickr relevant to "sailing". Despite the fact that all four are relevant to the same topic, clear differences can be observed.

We argue that user intent should be exploited in image retrieval since the goals of the photographer provides a simple, easily understandable explanation for the differences observed between photos. If we were to restrict the descriptions of photos used for indexing to containing only visual concepts, then both pictures in the top row could be described as "Photo depicting a sailboat in water". We

MUSA2'17, October 27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-5509-4/17/10…$15.00
https://doi.org/10.1145/3132515.3132521

could then capture the fact that users see a clear difference between the two by describing the first as “sailboat in a lot of water" and “sailboat near a dock”. However, such descriptions are not entirely satisfying since the exact amount of water or the presence of the dock is not the key characteristic differentiating these photos. Instead, the difference can be simply and naturally explained by user intent, i.e., the goal of the photographer in capturing the image.

We build on the assumption that user intent, i.e., the goals that users are pursuing when they take photos or search for images, has visual reflexes that can be captured by automatic visual classifiers. Our motivation for assuming a connection between the reason why a user takes a photo, and the visual content of the photo, is the phenomenon of *intentional framing*. Riegler et al. [18] define intentional framing as, ‘the sum of the choices made by photographers on exactly how to portray the subject matter that they have decided to photograph.’ This definition implies that intentional framing is observable in the photo, and also captured the fact that intent cannot be reduced to the concepts that are literally depicted in the image. Our notion of intent is related to the idea of *broad latent aspects* from conventional web search [24]. This type of query aspect is described by [24] as having two properties: broad latent aspects apply to a broad set of queries, and, users queries frequently leave these aspects unspecified.

Note that the user intent behind a query does not reduce to a sub-topic of a query. To illustrate this point, we return to Fig. 1 to discuss the two pictures in the bottom row. If we considered these pictures sub-topics of sailing, then we would loose connection with the intent of capturing a portrait and the intent of capturing media information that apply to a large range of queries beyond “sailing”. In other words, intent cross-cuts the topic (conceptual content) of a photo and provides us with an additional, easily understandable dimension with respect to which images can be indexed for image retrieval.

1.2 Challenges and Contributions

In this work, we investigate the ability of automatic intent inference to improve image retrieval, with a focus on retrieval algorithms that diversify image results lists. Given that intentional framing is noticeable to users as visible in photos, but cannot be directly reduced to topical or conceptual categories, it is surprising that user intent has yet to be fully exploited in image retrieval systems. Because the intent of the user is actually the goal of the user, we anticipate that user intent potentially has an important contribution to make to image retrieval. We believe that there are three major roadblocks that explain the relative lack of research on intentional framing in image search, which we now discuss in turn. The three major contributions of this paper address each of these three roadblocks.

The first roadblock is the lack of intent taxonomies (definitions of intent classes) and data sets annotated with intent labels. With the exception of early work by Lux et al. [15], we do not know of work that has asked photographers about their intent. However, asking photographers might not always be necessary. As mentioned above, people also perceive intent classes when they look at images. However, there is no standard set of intent classes that has universal applicability, and inner-annotator agreement on intent judgments is difficult to achieve. We address this roadblock by moving away from the idea that we should assume that a single authoritative set of intent classes is necessary before progress can be made on intent-based image retrieval. Instead, we pursue the idea that any well-informed intent taxonomy that has been applied consistently in order to label an image data set will be able to drive forward the state of the art in applying user-intent to improve image retrieval. To this end, we build an intent taxonomy on the basis of the analysis of a large collection of social images by a single expert annotator. The first contribution of this paper is this taxonomy and a large set of images labeled with the intent classes that it contains. This resource will allow reproduction of our research results and serves as a point of reference for future extension or modification of the intent taxonomy.

The second roadblock standing in the way of applying intent-based approaches to image retrieval is general assumptions about the nature of intent. We believe that the currently dominant assumption is that the visual variability among photos associated with a single intent class is too high for a classifier to be able to generalize. Likewise, the visual similarity among photos associated with different intent classes is also assumed to be too high. The second contribution of this paper is to show that these assumptions should not be made: with enough labeled data it is possible to train a classifier capable of automatically inferring the intent classes of images.

The third roadblock is the assumption that the usefulness of intent in image retrieval is likely to be limited. The third contribution of this paper is a set of experiments on multimodal social image retrieval that show that our intent classifier is able to contribute productively to the diversification of image search results.

The remainder of the paper is organized as follows. In Section 2, we cover related work, focusing on multimedia intent taxonomies that have been proposed in the literature. Then, in Section 3 we describe the iterative process by which we create our intent taxonomy and label our image data set. Next, we train a classifier using the labeled data, and report the results of experiments demonstrating its effectiveness. Finally, in Section 6, we apply our intent classes to information retrieval. Section 7 of the papers provides conclusions and an outlook.

2 RELATED WORK

This section covers research that has trained image classifiers to predict intent, as well as papers that have proposed multimedia intent taxonomies. We use this work to inform the creation of our own taxonomy on the basis of a large collection of social photos, which is discussed in Section 3.

2.1 Intent in Image Search

In [11], a survey of user intent in multimedia search, an overview is provided of papers that have made use of user intent to improve image search. Note that we are interested in papers that consider “user intent” to be the goal of the user, and not in work that users “intent” as a synonym for sub-concept, sub-topic, or query aspect. An early effort [12] used a a content-based intent classifier to adapt the display of search results. Recently, [20] used content-based image classification as one of a range of session-derived signals used to study user intent during image search. This work is interesting

because it focused on predicting intent by processing the sequence of images that users view in an image search session.

2.2 Taxonomies of Image Intent

The earliest taxonomy of image search intent is most probably that of Fidel et al. [4], who define intent along a spectrum from a *Data Pole* (obtain source of information) and the *Information Pole* (obtain an example of something or and object). To create their taxonomy they first categorized images into 12 attribute classes based on Jorgensen's research [10]. A user study was then conducted on 100 image search requests.

Other early work includes Cunningham et al. [2], who investigate 98 video search queries and categorize the user search intent into 8 groups. For example, *mental state* explicit the reference to subject's emotional state, *visual* explicit references to visual aspects to target video, *MSM* explicit references to mainstream media.

Very valuable image intent taxonomies have bee proposed on the basis of use studies. Lux et al. [13] conducted user research on 20 image search users on Flickr, and build an intent taxonomy with four classes: knowledge orientation, transaction, mental image and navigation.

By interviewing 10 people covering 40 different situations, Lux et al. [14, 15] studied the user intentions on photo production. The result was a intent taxonomy containing six reasons why users take pictures: *preserve a good feeling*, *preserve a bad feeling*, *share with family and friends*, *share with public*, *support a task* and *recall a specific situation.*

The work closest to our own is Hanjalic et al. [6], who employed a social-web mining approach, analyzing 4512 questions posted on Yahoo Answers that contained certain terms (i.e. "find" and "video"). The questions were sorted by crowdsourincg workers and then, using a card sorting process similar to our own manual content analysis process, and five classes were defined: Information, Experience:Learning, Experience: Exposure, Affect, and Object. Instead of investigating video, we investigate images. The commonality is that our work also exploits a large collection of social media to identify a set of intent classes that may not be absolutely universal, but is "good enough" to support tasks such as retrieval.

3 INTENT DISCOVERY

In this section, we describe the intent taxonomy (set of intent classes). The taxonomy is created on the basis of sub-set of a large-collection of social images, namely, the YFCC100M data set [21]. We create the classes using a manual content analysis [16] analysis approach. The intent classes are then used to label the data set. We are interested in demonstrating the ability of intent to go beyond concept detection. For this reason, we need to ensure that the data set contains images depicting a wide range of concepts, but also that there are enough images per concept. This section first discusses the specifics of the data set creation, and then describes the content-analysis procedure which was used to create the intent taxonomy.

3.1 Data set and set up

The first step in creating the data set is to define a list of concepts $C = \{c_1, c_2, ..., c_k\}$ that we would like to focus on. We chose the NUS-Wide concepts for several reason: this concept set is comparable to other concept sets used in the literature, it corresponds to the most frequent tags in Flickr, it includes both general and specific concepts, and finally, the concepts are of different types, including scene, object, event, program, people and graphics [1].

The concepts are used as queries to retrieve images from the YFCC100M data set with a tag-based retrieval system. We choose YFCC100M for a variety of reasons: 1. To our knowledge it is the largest public image collection that has ever been released. 2. It offers social images associated with user information. 3. It is publicly available, which supports reproducibility. Flickr has a batch tagging function that allows users to assign tags to a set of uploaded images [23]. Since batch tagging might affect the robustness of our intent data set, we only retain a maximum of three images with the same tags. For each concept (query), we collect top-200 relevant images. If less than 200 images are returned by the query, we use the total number of images returned. Our final data set contains 15618 images.

3.2 Taxonomy Creation

Next we turn to describing the content analysis approach by which we created the intent taxonomy. The approach involves iterative labeling by an expert annotator. The expert acquired his expertise by reading the papers in Section 2, and studying the taxonomies proposed there. During the labeling process he allowed his understanding of intent to evolve guided by what he observed in the data.

Figure 2: Example of two photos that appear similar at first glance, but ultimately land in two different intent categories, because of the difference in photographer goal.

Specifically, our manual content analysis approach proceeds by examining images in turn, and applying a preliminary intent label. Each new image is judged as either belonging to an existing class, or potentially requiring the introduction of a new class. Before introducing a new class, the annotator returns to the previous annotated images to ensure that it is not possible to accommodate the new image by updating the description of an existing class. If no existing class can be extended to accommodate the new image, a new intent class is introduced. Note that the process requires multiple iteration over the whole data set, and represents a full person month of image annotation work.

The two images in Figure 2 demonstrate the process with a hypothetical example. The left image is previously labeled with

"portrait" and the right image is new. The annotator examines all previous images and realizes that the right image does not fit into the "portrait" class because the picture was not take with the same goal as a portrait. Another class, *situation_documentation* is then added to the taxonomy. The final taxonomy is shown in Table 1.

3.3 Validation of the Intent Taxonomy

For the purpose of validating that intent class must be considered independently of topics, an analysis of our intent taxonomy was carried out from two perspectives: how intent classes i were shared among concept groups c and how concepts c are shared within different intent classes i. We discuss each here in turn.

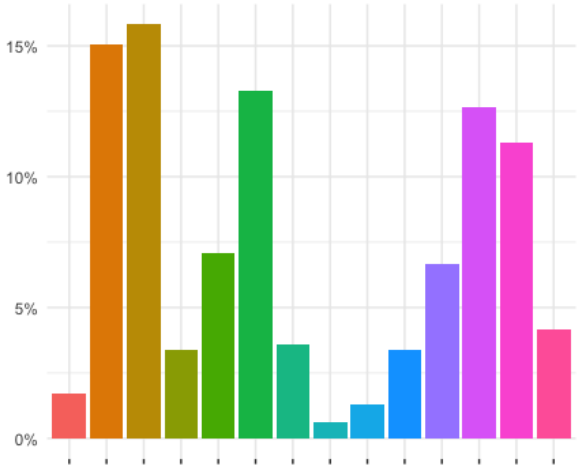

Figure 3: Distribution of intent classes, from left to right: art, candid, landscape, macro, media capture, portrait, product presentation, product presentation by person, situation documentation, social event private, social event public, setting, structure, wildlife.

In general, as can be seen in Figure 3, the 81 concept categories that form the basis of our data set are distributed across various intent classes. For example, the intent classes *portrait*, *candid*, *setting*, *media_capture* and *landscape* are found associated with images from 67 out of the 81 concepts. On the other hand, some intent classes did not show such generality. The intent classes *product_presentation_by_person*, *domesticate* and *situation_documentation* are found to be associated with 8–16 concepts. In average, an intent class is associated with **46.21** different concepts. This means that each intent class is associated on average with just over half of all concepts (to be exact **56.79%** of all concepts, with a standard deviation of **23.95**).

We turn now to look at how intent classes are distributed across concept categories. On average a concept category is associated with just over 8 intent classes (to be exact: **8.42** and **2.01** standard deviation.) These observations validate our intent taxonomy in the sense that they show that intent classes cross-cut image topic as represented by topic category, and that topics are associated with multiple intents. We note that the fact that our intent classes are effective for image search result diversification, provides further validation for our taxonomy. These experiments will be discussed in Section 6.

4 DISTRIBUTION OF INTENT CLASSES

In order to explore the distribution of intent classes over concepts, we ranked the concepts with respect to the number of intent classes with which they are associated. The top-10 concepts are listed in Table 2 and the bottom-10 in Table 3.

Concept c	Num Intents	Entropy
sun	12	3.16
frost	12	2.06
fish	12	3.09
tiger	11	2.80
animal	11	1.57
boat	11	2.61
earthquake	11	2.96
window	11	2.90
forest	11	2.86
fox	11	3.18

Table 2: Top-10 concepts ranked by the number of intent classes associated with the concept in the image collection.

Concept c	Num Intents	Entropy
coral	2	0.72
running	5	2.07
dog	5	1.52
elk	6	2.01
dancing	6	2.10
statue	6	1.83
plane	6	1.81
buildings	6	2.28
whale	6	1.67
surfing	6	1.93

Table 3: Bottom-10 concepts ranked by the number of intent classes associated with the concept in the image collection.

We observe that concepts differ with respect to the number and distribution of associated intent classes. For some concepts, there are photos in the collection associated with a wide range of different intent classes. For other concepts, the the number of intent classes is more limited. We capture these patterns with the entropy, which is reported in the tables.

More insight is provided by looking at four typical concepts in greater depth: *build*, *building*, *fox* and *wedding* as illustration in Figure 4. We see that concepts are directly associated with one dominant intent class, such as *bird* (top left) and *buildings* (top right). Other concepts are associated with multiple dominant intents. For example, for *wedding*, **78%** of the photos were captured related to people (*portrait* and *candid*) or aims at record a private social event. We noticed that for some concepts, such as *fox*, intent classes corresponded closely to what would be more conventionally

Class	Description
product_presentation	photos made to demonstrate or sell an item or a product.
product_presentation_by_person	photos made to demonstrate or sell an item or a product worn by or held by a human model.
social_event_public	photos made to record social events held for the public.
social_event_private	photos made to record planned social events held for a certain group of people.
situation_documentation	photo made to document desired or undesired situation.
landscape	photos made to shows spaces within the world, sometimes vast and unending.
macro	photos made to show extreme close-up photos of very small subjects too small to be noticed.
structures	photos made to demonstrate landmarks, buildings and similar structures.
setting	photos made to depict a inanimate objects, could be either nature or man-made, reflect specific aspect of the object.
portrait	photo of a person or a group of people that captures the characteristics of the subject(s)(who are aware they are being photographed).
candid	photo of a person or a group of people that capture the characteristics of the subject(s)(who are not aware they are being photographed).
wildlife	photos capture varies forms of wildlife in their nature habitat.
media_capture	image captured to carry out other information source.
art	photos captured to demonstrate abstract, creative vision of a photographer.

Table 1: Taxonomy of intent classes resulting from our intent discovery process

considered different interpretations of an ambiguous query. In a social media context, "fox" it might be an animal, a logo, or even a band of musicians, as reflected in our data set.

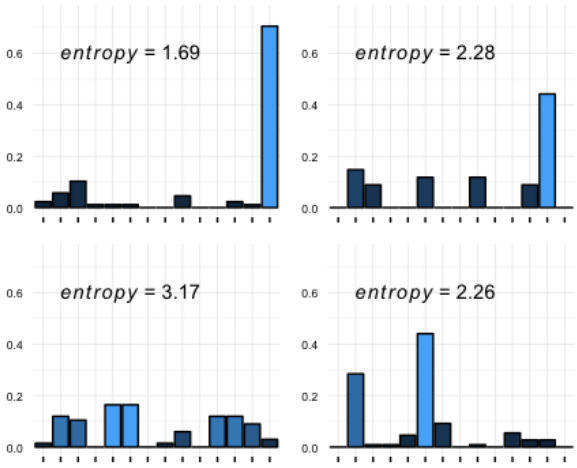

Figure 4: Distribution of intent classes with respect to four selected concepts plus entropy: bird (1.69), building (2.28), fox (3.17), wedding (2.26). Intent classes from left to right: art, candid, landscape, macro, media capture, portrait, product presentation, product presentation by person, setting, situation documentation, social event private, social event public, structure, wildlife.

To create the ground truth, we analyzed the distribution of intent classes over users. We aggregated the data set by user id, and filtered out users who had captured less than five photos, resulting in a set of 221 users. Figure 5 shows the entropy over the 14 intent classes with respect to each user.

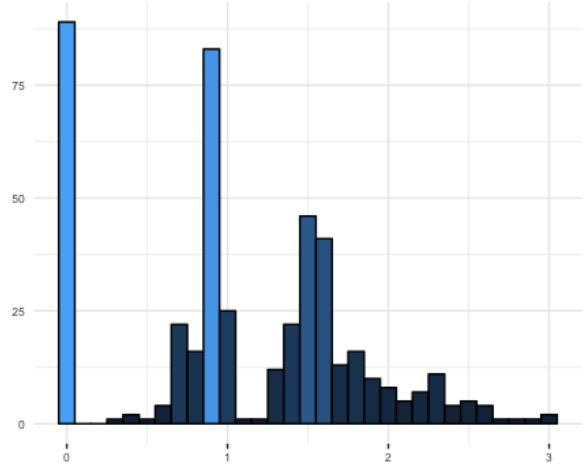

Figure 5: Intent entropy distributed over users, x-axis indicates entropy value, y-axis indicates number of users.

We see that the images of some users are associated with a stable set of intent classes and some have are unevenly associated with intent classes. A large number of users have one particular intent (i.e., entropy is 0). We note that this observation might be an artifact of the fact that for quite a few users the number of images per user is quite small. Overall the patterns in Figure 5 suggests that individual users have individual patterns of intent. We do not build further on this insight here, but only point out that in the future user-specific intent patterns could also be relevant for personalizing image search.

5 INTENT CLASSIFICATION

Next, we turn to investigate whether a classifier can learn to predict the intent class of an image. To this end, we adopt a conventional transfer learning scheme. Transfer learning consists of using models that were trained for a certain task and leveraging the knowledge that they acquired on a different, but related task [17]. In our case, VGGNet [19] was adopted to extract visual features from image pixels (originally trained on ImageNet [3]). The last fully connected layer (going from 2048 neurons to 1000 class scores) was removed and the rest of the networks serves as a feature extractor. We re-trained a Softmax classifier using a cross-entropy Softmax loss on our image data set annotated with 14 intent classes using 70% of the intent data set. Meanwhile, 25% of the images were held out for the validation purpose and we left 5% for future analysis. Before training, all images were re-sized into (224,224) and we applied random horizontal flipping, chopping and re-scaling for data augmentation. We evaluate with respect to accuracy.

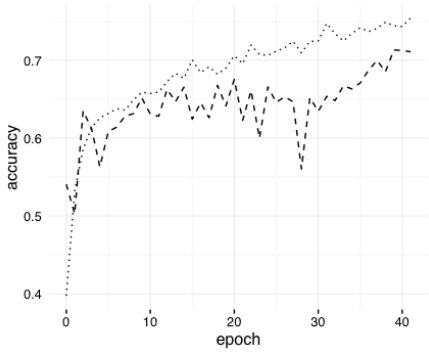

Figure 6: Intent class prediction: training set accuracy (dotted line) and validation set accuracy (dashed line).

As shown in Figure 6, our model achieved 71% accuracy on the validation set. This performance level demonstrates that there is enough visual stability within intent classes to allow a classifier to generalize them. We expect with a larger labeled intent data set, model performance could be further improved.

5.1 Discriminative Ability of Intent Classes

Next we analyze the ability of the classifier to distinguish intent classes in more depth. The data set we use here is the 5% of the labeled data (ca. 800 images) that was reserved (as mentioned above) for future analysis. Employing our trained model for intent classification, we predicted the outcomes for this small proportion of data.

As reflected in Figure 8, for most of the intent classes, over 70% of the images in the data set belonging to that category can be correctly predicted. However, for some particular intent classes, it is still non-trivial for our classifier to capture the patterns of the images belonging to this class. For example, for the classes *art* and *product_presentation*, our intent classifier can only identify 31.25% and 37.5% images correctly. We attribute this observation to the visual diversity of the images in these intent classes. However, we

Figure 7: Intent predicted as *landscape* from four different concepts, *road, zebra, city space, reflection.*

also point out, as Figure 8 indicates that for these two intent classes, there are fewer images in our data set compared to other classes.

It is also interesting to observe that certain intent classes are easily confused. This effect occurs case of *candid* and *portrait*. By definition, we differentiate *candid* and *portrait* by the awareness of person or people appearing in the photo. Awareness is reflected in a subtle difference in direction of the gaze of the person in the image, and people judging images also might confuse these classes. Moreover, there could be multiple persons within an image (some aware while others not).

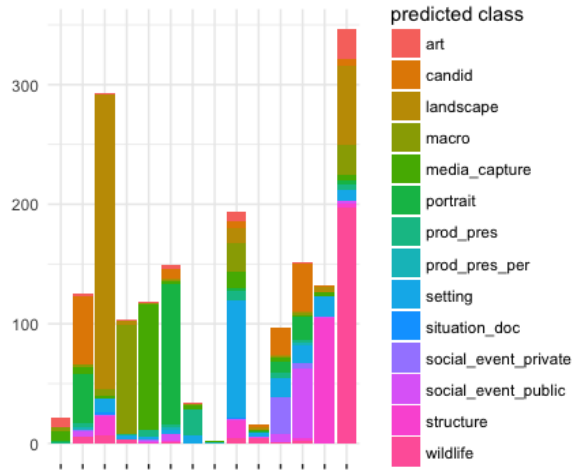

Figure 8: Discriminative ability of intent classes: bars are the intent classes (alphabetical order left to right). Height of the bars indicates the number of images in each class. Colors indicate class confusion.

6 SEARCH RESULT DIVERSIFICATION

An effective image retrieval system should be able to present results that are both relevant and that are covering different (i.e., diverse)

Figure 9: Results for the query "accordion player": the original ranking (top) and the results-list that has been re-ranked based on intent classes (bottom).

interpretations of the query. By diversifying the pool of possible results, one can increase the likelihood of providing the user with the information needed [8]. There are two reasons why we believe intentional framing could be potentially helpful: 1. Intentional framing naturally reflects aspects of photos that are expressed by photographers. 2. Intentional framing could be processed in a direct, light-weight fashion without extensive computational resources. In view of this motivation, we conducted an experiment on intent-based search result diversification.

6.1 Data set and set up

We evaluate the ability of our intent-based approach to improve the diversity of image search results using the task definition, development data set, and ground truth that were released by the MediaEval 2016 Retrieving Diverse Social Images Task [7]. The development data set consists of 110 general-purpose, multi-topic queries defined by the task organizers and 33,000 Creative Commons licensed images collected from Flickr.

Additionally, all images are labeled by expert annotators in terms of both relevance and diversity. For each query and its corresponding images, the relevance is judged by aggregating results from yes/no annotations. To create the ground truth for image search result diversity, images are grouped into clusters based on visual similarity. For example, images associated with query term *accordion_player* are grouped into 18 clusters including *accordion player on boat, accordion player standing street art.* Note that only relevant images are annotated for diversity.

Diversification performance is evaluated in terms of Cluster Recall at 5, 10 and 20 (CR@X), a measure that assesses how many different clusters from the ground truth are represented among the top X results, thus, it reflects the diversification quality of a given image result set.

We compared our intent-based diversification approach with the Flickr baseline [9]. We also compare our approach with a state-of-the-art diversification approach, designated "UPMC@MediaEval16", which is chosen because it was the top-performing approach on the MediaEval Retrieving Diverse Social Images Task in 2016 [22]. UPMC@MediaEval16 adopts a four-step approach. The first step is to create a refined initial ranked list by re-ranking the Flickr baseline using textual features (vector space model with tf-idf weights) with the aim of increasing precision. After that, the top N images are grouped using Hierarchical Clustering and the CNN features released with the data set. For each of the resulting sub-clusters, the images are re-sorted based on their position in the refined initial ranked list. Finally, the results are re-ranked by alternating images from different clusters.

6.2 Intent-based Reranking

Our intent-based re-ranking approach is illustrated in Figure 10. It adopts the same general strategy of clustering and alternating images from clusters used by UPMC@MediaEval16. However, the difference is that instead of creating clusters by applying a clustering method to the initial ranked list, we create clusters by classifying the images in intent classes and grouping together images that belong to the same class. For each query and its ranked results list returned by Flickr, we apply a three step approach to diversify the search results. The first step in this process is to predict the intent classes of retrieved image list (i.e., Flickr baseline). Once the predictions are made, we cluster the first N results based on intent class. In our case, N equals to 20. Following this step, we pick the top-one photo without replacement for each cluster, under the assumption that new clusters reflect diversity as captured by intent. Finally, these new clusters are sorted internally based on the original rank position of the images and concatenated to create the final re-ranking list. Figure 9 presents a comparison of original retrieved result and diversified retrieved result.

Note that the main difference between our diversification approach and UPMC@MediaEval16 is that we use intent classification. However, another difference is that we do not refine the initial ranking using text-based features. The omission of such an initial re-ranking step in our approach is not critical in the context of this paper, which focuses on diversification rather than on precision.

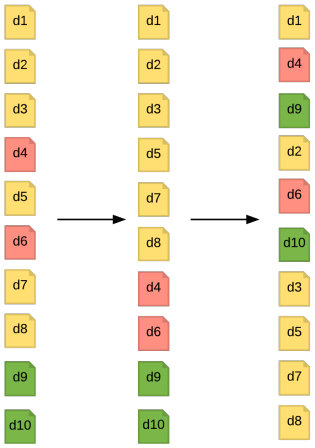

Figure 10: Re-ranking search result based on clusters of intents, colors of documents reflects different predicted intents. From left to right: initial ranking result, intent clusters, re-ranked result.

6.3 Result

Table 4 reports the results in terms of the official MediaEval 2016 evaluation metric CR@20.

Approach	CR@20	P@20
Flickr Baseline	36.1%	56.2%
UPMC@MediaEval16	49.4%	69.6%
Intent-based diversification	49.4%	56.2%

Table 4: Performance comparison against MediaEval 2016 retrieve diverse social image task over 110 queries. CR@5 and CR@10 of our approach are 27.9%, 37.5% respectively.

A significant increase can be found in terms diversity between the original Flickr baseline and UPMC@MediaEval. The most interesting aspect of this table the fact that the intent-based approach is able to achieve the same level of diversity as the state-of-the art UPMC@MediaEval16 approach. We take this result as a confirmation of the ability of classifiers trained to recognize intent to improve image retrieval algorithms. In Table 4, we also report the precision (P@20) of the three approaches. The precision gives us insight into the degree to which our approach could possibly improve if we refined the initial list with text-based features as UPMC@MediaEval16 does. Improvements in precision feed improvements in diversity, since the ground truth with respect to which cluster recall is calculated includes relevant images only.

We point out three other aspects of the intent-based diversification approach, which make it possibly more useful compared with UPMC@MediaEval16 in practice. These aspects arise because, in contrast to unsupervised learning (clustering), our approach directly tries to predict pre-defined intent classes. We mention these aspects since they represent advantages of our approach over an unsupervised learning approach, even though the two achieve the same performance with respect to the diversity metric. First, intent-based diversification has the advantage of better understandability since the classification result is able to directly provide a user-interpretable indication of the reason behind the ranking. The retrieval system can provide the user with an explanation for its prioritization of search results. Second, once the model has been trained, we do not necessarily need to fine-tune the hyper parameters, i.e., the position to cut the dendrogram. Third, image labels are generated off-line at indexing time, and a clustering step at query time, which increases the system response time, is not necessary.

7 CONCLUSION AND OUTLOOK

In this paper, we have introduced a taxonomy of intent for social photos. Additionally, we have confirmed that the visual content of images are connected to human perceptions of photographer's intent by employing a transfer learning scheme for intent classification. Moreover, we have shown that intent-based search result diversification can achieve on performance on par with that of state-of-the-art image search result diversification algorithms on a social image retrieval benchmark. Overall, this study strengthens the idea of indexing images with respect to dimensions of human perception could be important for various tasks. The study has contributed to enhancing our understanding of intentional framing. The potential of "why" dimension has been established, and future research can be undertaken to explore how intentional framing can further enhance image retrieval.

Here, we have used an intent taxonomy created by a single expert. In our future work, we are interested in the extent to which changes or refinements of this taxonomy impact the predictability of the intent classes, and the ability of the intent classes to improve the diversity of image retrieval results. We suspect that no single perfect taxonomy exists for all cases, but many taxonomies are probably "good enough". If this is the case, we are interested in optimizing the taxonomy to maximize both intuitiveness to users and the performance of the intent classifier.

REFERENCES

[1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009.*

[2] Sally Jo Cunningham and David M. Nichols. 2008. How people find videos. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2008, Pittsburgh, PA, USA, June 16-20, 2008.* 201–210.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA.* 248–255.

[4] Raya Fidel. 1997. The image retrieval task: implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia* 3 (1997), 181–199.

[5] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang. 2015. Recent Advances in Convolutional Neural Networks. *CoRR* abs/1512.07108 (2015).

[6] Alan Hanjalic, Christoph Kofler, and Martha Larson. 2012. Intent and its discontents: the user at the wheel of the online video search engine. In *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012.* 1239–1248.

[7] Bogdan Ionescu, Alexandru-Lucian Gînsca, Bogdan Boteanu, Mihai Lupu, Adrian Popescu, and Henning Müller. 2016. Div150Multi: a social image retrieval result diversification dataset with multi-topic queries. In *Proceedings of the 7th International Conference on Multimedia Systems, MMSys 2016, Klagenfurt, Austria, May 10-13, 2016.* 46:1–46:6.

[8] Bogdan Ionescu, Alexandru-Lucian Gînsca, Maia Zaharieva, Bogdan Boteanu, Mihai Lupu, and Henning Müller. 2016. Retrieving Diverse Social Images at MediaEval 2016: Challenge, Dataset and Evaluation. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016.*

[9] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. 2016. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools Appl.* 75, 2 (2016), 1301–1331.

[10] Corinne Jorgensen. 1995. *Image Attributes: An Investigation.* Ph.D. Dissertation. Syracuse, NY, USA. AAI9625856.

[11] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User Intent in Multimedia Search: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.* 49, 2 (2016), 36:1–36:37.

[12] Christoph Kofler and Mathias Lux. 2009. Dynamic presentation adaptation based on user intent classification. In *Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, October 19-24, 2009*. 1117–1118.

[13] Mathias Lux, Christoph Kofler, and Oge Marques. 2010. A classification scheme for user intentions in image search. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume, Atlanta, Georgia, USA, April 10-15, 2010*. 3913–3918.

[14] Mathias Lux, Marian Kogler, and Manfred del Fabro. 2010. Why did you take this photo: a study on user intentions in digital photo productions. In *Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access*. ACM, 41–44.

[15] Mathias Lux, Mario Taschwer, and Oge Marques. 2012. A closer look at photographers' intentions: a test dataset. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia, CrowdMM@ACM Multimedia 2012, Nara, Japan, October 29, 2012*. 17–18.

[16] Kimberly A Neuendorf. 2016. *The content analysis guidebook*. Sage.

[17] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359.

[18] Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'How' Reflects What's What: Content-based Exploitation of How Users Frame Social Images. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*. 397–406.

[19] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[20] Mohammad Soleymani, Michael Riegler, and Pål Halvorsen. 2017. Multimodal Analysis of Image Search Intent: Intent Recognition in Image Search from User Behavior and Visual Content. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*. 251–259.

[21] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[22] Sabrina Tollari. 2016. UPMC at MediaEval 2016 Retrieving Diverse Social Images Task. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016*.

[23] Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.* 44, 4 (2012), 25:1–25:24.

[24] Xuanhui Wang, Deepayan Chakrabarti, and Kunal Punera. 2009. Mining broad latent query aspects from search sessions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*. 867–876.

COLOPHON

This document was typeset using LaTeX. The document layout was generated using the `arsclassica` package by Lorenzo Pantieri, which is an adaption of the original `classicthesis` package from André Miede.