A Database System for Multimedia Analytics and Analysis

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Ralph Marc Philipp Gasser

Basel, 2022

Zusammenfassung

Abstract

Acknowledgements

Good luck.

This work was partly supported by the Swiss National Science Foundation, which is also thankfully acknowledged.

Contents

Zusammenfassung iii

Abstract v

Acknowledgements vii

List of Figures xi

List of Tables xiii

I Introduction 1

1 Introduction 3

1.1 Working with Multimedia Data 4

1.1.1 Multimedia Analysis & Retrieval 4

1.1.2 Multimedia Analytics 5

1.2 Research Gap and Objective 5

1.2.1 Motivation 7

1.2.2 Research Questions 7

1.3 Contribution 7

2 Applications and Use Cases 9

2.1 Use case 1: Multimedia Retrieval System 9

2.2 Use case 2: Analysis of Social Media Streams 9

2.3 Use case 3: Magnetic Resonance Fingerprinting (MRF) 9

3 Related Work 11

II Foundations 13

4 On Multimedia Analysis and Retrieval 15

4.1 Multimedia Data and Multimedia Collections 15

4.2 Multimedia Retrieval 15

4.2.1 Similarity and the Vector Space Model 15

4.2.2 Approximate Nearest Neighbor Search 15

4.2.3 Beyond Similarity Search 15
4.3 Online Multimedia Analysis 15
4.4 Multimedia Analytics 16
4.4.1 Beyond Similarity Search 16

5 On The Design of a Database Management System 17
5.1 The Relational Data Model 17
5.2 Storage, Indexes and Caching 17
5.3 Query Parsing, Planning and Execution 17
5.4 Architectual Considerations 17

III Dynamic Multimedia Data Management 19

6 Modelling a Database for Dynamic Multimedia Data 21
6.1 Adaptive Index Management 21
6.2 Generalization of Similarity Search 22
6.3 Cost Model for Retrieval Accuracy 22
6.4 Architecture Model 22

7 Cottontail DB 23

IV Discussion 25

8 Evaluation 27
8.1 Adaptive Index Management 27
8.2 Cost Model 27

9 Conclusion & Future Work 29

Appendix 31

Bibliography 33

Curriculum Vitae 35

Declaration on Scientific Integrity 37

List of Figures

6.1 Adaptive index structures overview. 21

List of Tables

1.1 List of research questions resulting from challenging assumptions one, two and three. 7

Part I

Introduction

1

Introduction

The term *multimedia* describes the combination of different forms of content – also called *modalities* – into a single, sensory experience that carries a higher level semantic. Those modalities include but are not limited to textual, visual (e.g., images or videos) or aural (e.g, music, sound effects, speech) input. In addition, signals produced by various sensors and devices can also be seen as media modalities, even though direct experience by a human consumer may require pre-processing through specialized hard- and software.

People interact with multimedia on a daily basis when watching videos on Netflix or YouTube, when listening to music on Spotify or when browsing a private image collection on their laptop. Multimedia content makes up a large part of today's Internet and its mere omnipresence makes it a major driving force behind its growth, as both volume and variety increases steadily. A large contributing factor are social media platforms, where users act both as consumers and producers of content. Current estimates [1] suggest, that there are roughly 4.66 billion active Internet users worldwide, of which 4.2 million can be considered active social media users. Facebook alone contributed to 144000 uploaded images per minute in 2020. And many more of these platforms, such as *Instagram* org *Twitter* serve billions of users with mixed, self-made content involving text, images, videos or a combination thereof. A similar study found [2], that by 2025 we will produce a yearly amount of 175 Zettabytes (i.e, 10^{21} bytes).

Given these numbers, the need for efficient and effective tools for *managing*, *manipulating*, *searching*, *exploring* and *analysing* multimedia data corpora comes at no surprise.

[1] Source: Statista.com, "Social media usage worldwide", January 2021
[2] Source: Statista.com, "Big Data", January 2021

1.1 Working with Multimedia Data

On a very high level, multimedia data collections consist of individual (raw) multimedia items, such as video, image or audio files. Each item, in turn, comprises of *content*, *annotations* and *metdata*. Unlike traditional data collections that contain only text and numbers, the content of the multimedia item itself is unstructured, which gives rise to a need for compact *feature representations* that can be handled by data processing systems [ZW14]. Traditionally, such feature representations are numerical vectors $f_i \in \mathbb{R}^d$ but could in theory be any mathematical object that can be processed by a computer, such as a tensor. Annotations, metdata and features may either be generated upon the item's creation (e.g., for technical metadata), as a result of data-processing and analysis or by manually adding the information at some stage of the item's lifecycle.

1.1.1 Multimedia Analysis & Retrieval

Multimedia analysis and multimedia retrieval are two sides of the same coin and are often used interchangeably in literature. Both areas of research have their roots in *computer vision* and *pattern recognition*, which started in the 1970s and deal with the automated, computer-aided analysis of visual information on images and videos. Classical tasks involve automatic classification of images, such as, labeling images as either depicting a dog or a cat. For the purpose of this thesis, and – one could argue – as a general definition, the distinction can be put as follows:

Multimedia analysis deals with the extraction and processing of meaningful feature representations from media items. In the early days of computer vision, a lot of effort went into the engineering of feature representations that appropriately captured certain aspects of a media item, such as the colour distribution, texture or keypoints in an image. With the advent of deep learning, the extraction of such features from, e.g., images could largely be automated through neural network architectures such as the *Convolutional Neural Network (CNN)*. Once a feature has been obtained, it can be used to perform different tasks such as classification, clustering or statistical analysis, all of which fall under the umbrella of multimedia analysis.

Multimedia retrieval can be seen a as a very special use cases of multimedia analysis. It's a dedidcated field of research that deals with the act of searching and finding an item of interest within a large multimedia collection. Even though this may sound like the main function of a database, it is a very different

task for multimedia than it is for structured data [BVB+07]. On the one hand, the structure of a relational database is a given and using languages like SQL, a user can specify exactly what elements from the database should be selected using predicates that either match or don't match the items in a collection. Retrieving multimedia data, on the other hand, comes with indirections due to the feature representations used for querying and the *semantic gap* associated with them. One model to work with the feature representations is by obtaining a (dis-)similarity score from the features, a process referred to as ranking. So in addition to extraction of appropriate features and effective ranking algorithsm, multimedia retrieval also concerns itself with aspects such as query formulation, results presentation, data storage etc.

1.1.2 Multimedia Analytics

In simple terms, *multimedia analytics* can be seens as a combination of methods from multimedia analysis with visual analytics. While multimedia analysis deals with the different media types and how meaningful representations can be extracted from them, visual analytics deals with the user's interaction with the data [CTW+10]. Multimedia analytics aim at generating knowledge from the multimedia data, for example, by summarizing data or by offering means to explore it.

1.2 Research Gap and Objective

The starting point for the research described in this thesis is the current state-of-the-art for data management in multimedia retrieval and analytics as briefly touched upon in the previous sections. Starting from the models and solutions proposed in [GS16; Gia18] and motivated by the "Ten Research Questions for Scalable Multimedia Analytics" [JWZ+16], this thesis challenges three basic assumptions currently employed and operated upon and explores the ramifications of doing so, with the higher level goal of furthering convergence between research conducted in *multimedia retrieval*, *multimedia analysis* and *multimedia analytics* on the one hand, and classical *database systems* on the other. The assumptions that are being challenged are namely:

Assumption 1: Staticity of data collections Most multimedia retrieval systems today make a clear distinction between an *offline* phase during which media items are analysed, features are generated and derived data is ingested into

a data management system, and an *online* phase, during which queries of the database take place. Usually, no changes to the data collection are being made during the online phase. This formal model is advertised by both [Gia18] and [Ros18] and to the best of our knowledge, most existing systems explicitly or implicitly function in this manner. This simplification allows for time consuming processes related to feature extraction and indexing to take place separated from any concurrent query activities and eases requirements on transaction isolation. While this is convenient from a perspective of system design, such a mode of operation is limiting when facing data that changes frequently, as is the case, for example, when doing real-time analysis or when having an application with CRUD support.

Assumption 2: Similarity search is nearest neighbor search Multimedia retrieval relies strongly on a notion of similarity search that is usually expressed as finding the k nearest neighboring feature vectors $\vec{v}_{i\in[1,k]} \in \mathrm{C}$ to a query vector $\vec{q} \in \mathbb{R}^d$ in a collection $\mathrm{C} \subset \mathbb{R}^d$ given a certain distance function. Very often, metrics such as the Euclidean or the Manhattan distance are employed for this comparison. This is also refered to as the *vector space model*. While this model is very concise and rather simple, it merely allows for ranking of potential results and finding the relevant or desired item(s). This model is, however, limited when looking at tasks such as summarization or structuring of data, as required for multimedia analysis and analytics. Furthermore, there are cases in which features are not real-valued vectors and similarity is not directly related to proximity.

Assumption 3: User defines execution Database management systems usually evaluate and select the execution plan for an incoming query during a step that is refered to as *query planning*. The underlying assumption here is that the database system has all the information required to determine the most effective execution plan in terms of cost parameters such as required I/O, CPU and memory usage. In similarity search, this is not the case since, for example, index selection relies on a lot of different aspects that, to some extent, can be parametrized by the client issuing a query or that may be subject to change. Therefore, the index used for executing a query is usually selected manually by the user issuing the query. This limits the amount of optimization that can be applied by the multimedia database system especially in the face of non-static data collections.

1.2.1 Motivation

Teh a

1.2.2 Research Questions

Challenging the aforementioned assumptions raises very specific questions that fundamentaly impact the design of a *multimedia database*. These questions are briefly summarized in Table 1.1.

Table 1.1 List of research questions resulting from challenging assumptions one, two and three.

RQ	Question	Related to
1	Which commonly used, secondary index structures for NNS (e.g., VA [WSB98], LSH [IM98], PQ [JDS11] based indexes) can cope with dynamic data collections and to what extent?	Assumption 1
2	Can we quantify and estimate how much the retrieval quality of index structures from RQ1 deterioration as changes are being made to the underlying data collections?	Assumption 1
2	How can we handle secondary index structures from RQ1 for which to expect deteriorated retrieval quality during query planning and execution?	Assumption 1
4	How can user knowledge about the the retrieval task at hand be factored into query planning without forcing the user the make explicit choices about how a query should be executed?	Assumption 1 & 3
5	88	

Obviously, the list of questions in Table 1.1 is not exhaustive and many more implications can be derived from the assumptions challenged so far. However, the questions are the one tackled by the research presented in this thesis.

1.3 Contribution

2

Applications and Use Cases

2.1 Use case 1: Multimedia Retrieval System

2.2 Use case 2: Analysis of Social Media Streams

2.3 Use case 3: Magnetic Resonance Fingerprinting (MRF)

3

Related Work

Part II

Foundations

4

On Multimedia Analysis and Retrieval

4.1 Multimedia Data and Multimedia Collections

Formalisation of what multimedia data is and how it can look like (video, audio, images, text + metadata etc.).

4.2 Multimedia Retrieval

4.2.1 Similarity and the Vector Space Model

4.2.2 Approximate Nearest Neighbor Search

Describe techniques for approximate nearest neighbor search (ANN). Focus on a more conceptual overview of the types of algorithms rather than just enumerating concrete examples; this can be used as a build-up for discussing properties of different index structures later.

4.2.3 Beyond Similarity Search

Retrieval and analytics techniques that go beyond simple similarity search (e.g. SOM, summarization, clustering)

4.3 Online Multimedia Analysis

Introducing an online analysis pipeline (e.g., Pythia / Delphi).

4.4 Multimedia Analytics

Describe how the combination of analysis

4.4.1 Beyond Similarity Search

5

On The Design of a Database Management System

Digression into design considerations of a database management system (storage, locking, query planning, execution model etc.)

5.1 The Relational Data Model

5.2 Storage, Indexes and Caching

5.3 Query Parsing, Planning and Execution

5.4 Architectual Considerations

Part III

Dynamic Multimedia Data Management

6

Modelling a Database for Dynamic Multimedia Data

Is there more?

6.1 Adaptive Index Management

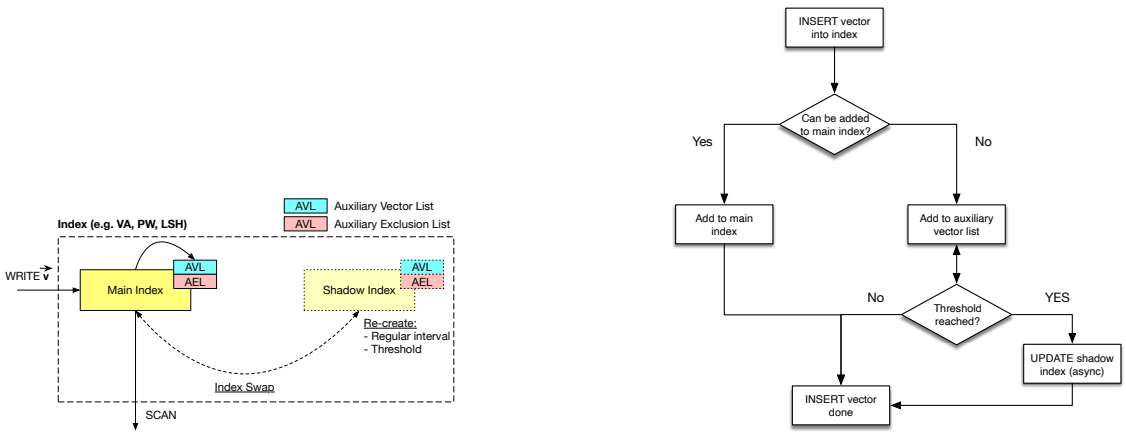

Figure 6.1 Adaptive index structures overview.

Describe model for index management in the face of changing data:

– Reason about properties of secondary indexes for NNS (e.q., PQ, VA, LSH) with regards to data change

– Derivation of error bounds possible (e.g., usable for planning)?! Use in query planning?

- Systems perspective: How to cope with "dirty" indexes (e.g, auxilary data structure, offline optimization)

6.2 Generalization of Similarity Search

Describe generalized model for similarity search:

- NNS: Scan -> (Predicate) -> Distance Function -> Sort -> Limit -> (Predicate); no need for dedicated language feature aside from distance function
- Distance function is a binary function $D(q, v) \longrightarrow d$ (consequence: different types of distance functions, application of weights merely an operation before executing the function etc.)
- q and v can be elements of $\mathbb{R}^d, \mathbb{C}^d$ or even matrices
- Systems perspective: How enable planner to reason about function execution

6.3 Cost Model for Retrieval Accuracy

Describe cost model with following properties:

- Cost function: $f(a_{cpu}, a_{io}, a_{memory}, a_{accuracy}) \longrightarrow C$
- Means to estimate results accuracy from execution path (e.g., when using index) based on properties of the index
- Means to specify importance of accurate results (e.g., global, per-query, context-based i.e. when doing 1NN search) in comparison to other factors
- Cost usable in query planner
- What about execution time?

6.4 Architecture Model

Putting everything together into a unified systems model (base on previous work + aforementioned aspects).

7

Cottontail DB

Implementation chapter for Cottontail DB

Part IV

Discussion

8

Evaluation

8.1 Adaptive Index Management

Brute force vs. plain index vs. index with auxilary data structure

8.2 Cost Model

Benchmark effect of cost model in different settings (e.g. based on use cases from chapter 2)

9

Conclusion & Future Work

Appendix

Bibliography

[BVB+07] Henk M Blanken, Arjen P de Vries, Henk Ernst Blok, and Ling Feng. *Multimedia Retrieval*. Springer, 2007.

[CTW+10] Nancy A. Chinchor, James J. Thomas, Pak Chung Wong, Michael G. Christel, and William Ribarsky. Multimedia Analysis + Visual Analytics = Multimedia Analytics. *IEEE Computer Graphics and Applications*, 30(5):52–60, 2010. DOI: 10.1109/MCG.2010.92.

[Gia18] Ivan Giangreco. *Database Support For Large-Scale Multimedia Retrieval*. PhD thesis, University of Basel, Switzerland, August 2018.

[GS16] Ivan Giangreco and Heiko Schuldt. ADAMpro: Database Support for Big Multimedia Retrieval. *Datenbank-Spektrum*, 16(1):17–26, 2016. DOI: 10.1007/s13222-015-0209-y.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, pages 604–613, 1998.

[JWZ+16] Björn Þór Jónsson, Marcel Worring, Jan Zahálka, Stevan Rudinac, and Laurent Amsaleg. Ten Research Questions for Scalable Multimedia Analytics. In *International Conference on Multimedia Modeling*, pages 290–302. Springer, 2016.

[JDS11] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product Quantization for Nearest Neighbor Search. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(1):117–128, January 2011. DOI: 10.1109/TPAMI.2010.57.

[Ros18] Luca Rossetto. *Multi-Modal Video Retrieval*. PhD thesis, University of Basel, Switzerland, 2018.

[WSB98] Roger Weber, Hans-Jörg Schek, and Stephen Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *VLDB*, volume 98, pages 194–205, New York City, NY, USA. Morgan Kaufmann, 1998.

[ZW14] Jan Zahálka and Marcel Worring. Towards Interactive, Intelligent, and Integrated Multimedia Analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 3–12, 2014. DOI: 10.1109/VAST.2014.7042476.

Curriculum Vitae

Name	Ralph Marc Philipp Gasser
	Brunnenweg 10, 4632 Trimbach
Date of Birth	28.03.1987
Birthplace	Riehen BS, Switzerland
Citizenship	Switzerland

Education

since Jan. 2000	Ph. D. in Computer Science under the supervision of Prof. Dr. Heiko Schuldt, Databases and Information Systems research group, University of Basel, Switzerland
Sept. 1997 – Aug. 1999	M. Sc. in Computer Science, University of Basel, Switzerland
Sept. 1994 – Aug. 1997	B. Sc. in Computer Science, University of Basel, Switzerland

Employment

since Jan. 2000	Research and teaching assistant, Databases and Information Systems research group, University of Basel, Switzerland

Publications

1937

- **turing:1937**.

Hand-in in thesis and separately

Declaration on Scientific Integrity

includes Declaration on Plagiarism and Fraud

Author
Ralph Marc Philipp Gasser

Matriculation Number
2007-050-131

Title of Work
A Database System for Multimedia Analytics and Analysis

PhD Subject
Computer Sciences

Declaration
I hereby declare that this doctoral dissertation *"A Database System for Multimedia Analytics and Analysis"* has been completed only with the assistance mentioned herein and that it has not been submitted for award to any other university nor to any other faculty at the University of Basel.

Basel, DD.MM.YYYY

Signature

Hand-in separately