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

List of Figures

List of Tables

Part I

Introduction

1

Introduction

The term *multimedia* describes the combination of different forms of digital media – also called *modalities* – into a single, sensory experience that carries a higher level semantic. Those modalities include but are not limited to images and videos (visual), music, sound effects and speech (aural) or textual information. As a contrast, more exotic media types such as 3D models or signals produced by sensors can also be seen as modalities, even though experience by a human consumer may depend on pre-processing by specialized hard- and software.

Nowadays, people encounter digital media and 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. (Multi-)media content makes up a large part of today's Internet and constitutes a major driving force behind its growth, as both volume and variety increases at an ever increasing pace. A large contributing factor are social media platforms, where users act both as consumers and producers of digital content. Current estimates[1] suggest, that there are roughly 4.66 billion active Internet users worldwide, of which 4.2 billion 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* or *Twitter*, serve millions of users with mixed, self-made content involving text, images, videos or a combination thereof. A similar study found, that by 2025 we will produce a yearly amount of 175 Zettabytes (i.e, 10^{21} bytes) worth of data[2].

Looking at these numbers, the need for efficient and effective tools for *managing*, *manipulating*, *searching*, *exploring* and *analysing* multimedia data corpora becomes very apparent, which has given rise to different areas of research.

[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 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 on a data level, which is why *feature representations* that reflect a media item's content in some way and that can be handled by data processing systems are required [ZW14]. Traditionally, such feature representations have often been numerical vectors $f_i \in \mathbb{R}^d$. However, in theory, any mathematical object that can be processed by a computer can act as a feature.

It is important to note, that a media item can comprises of all of the aforementioned components and that a media collection can contain items of different types. Furthermore, when looking at a media item's lifecycle, all of these aspects are not static; 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. Hence, any data management system aimed at multimedia retrieval must be able to cope with changes to that information. Both aspects are formalized in Section 4.1.

1.1.1 Multimedia Analysis

Multimedia analysis has its roots in *computer vision* and *pattern recognition*, which started in the early 1970s and deal with the automated, computer-aided analysis of visual information found in images and later videos. In the early days of computer vision, a lot of effort went into the engineering of feature representations that captured certain aspects of a media item's content, such as the colour distribution, texture or relevant keypoints [Low99; BTVG06] in an image. Once such features have been obtained, they can be used to perform different tasks such as classification, clustering or statistical analysis. 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)* and sometimes even be integrated with the downstream analysis.

Obviously, such analysis is not restricted to the visual domain and can be applied to other types of media such as speech, music, video or 3D models with specific applications, such as, speech recognition, music classification, movement detection in videos or classification of 3D models, all of which fall into the broader category of multimedia analysis.

1.1.2 Multimedia Analytics

Multimedia analytics aims at generating new knowledge and insights from multimedia data by combining techniques from multimedia analysis and visual analytics. While multimedia analysis deals with the different media types and how meaningful representations and models can be extracted from them, visual analytics deals with the user's interaction with the data and the models themselves [CTW+10; KKE+10]. Simply put, multimedia analytics can be seens as a back and forth between multimedia (data) analysis and visual analytics, wheras analysis is used to generate models as well as visualisations from data which are then examined and refined by the user and their input. This is an iterative process that leads to new knowledge being produced and may in and by itself lead to new information being attached to the multimedia items in the collection.

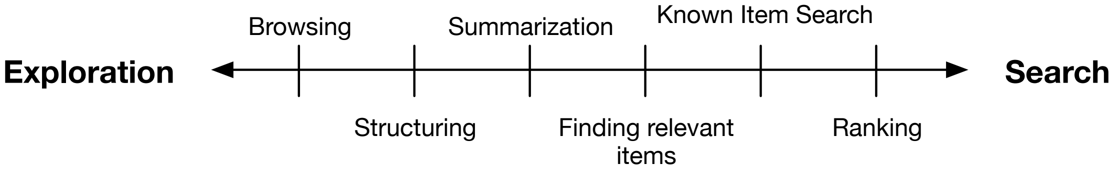

Figure 1.1 Exploration-search axis of multimedia analytics [ZW14].

For analytics on a multimedia collection, Zahalka et al. [ZW14] propose the formal model of an *exploration-search axis*, which is depicted in Figure 1.1. The model is used to characterize the different types of tasks carried out by the user. The axis specifies two ends of a spectrum, with *exploration* marking one end – in case the user knows nothing about the data collection – and *search* marking the other end – in case the user knows exactly which specific items of a collection they're interested in. During multimedia analytics, a user's activities oscillate between the two ends of the spectrum until the desired knowledge has been generated. Unsurprisingly, all of the depicted activites come with distinct requirements on data transformation and processing.

As data collections become large enough for the relevant units of information, such as feature representations, annotations and metadata to no longer fit into main memory, multimedia analytics and the associated data processing quickly becomes an issue of scalable data management [JWZ+16]. This data management aspect becomes very challening when considering the volume and variety of the multimedia data, the velocity at which new data is generated and the inherently unstructured nature of the media data itself.

1.1.3 Multimedia Retrieval

Traditionally, multimedia retrieval or content-based retrieval could be seen as a special niche within the multimedia analysis domain. It constitutes a dedidcated field of research that deals with searching and finding items of interest within a large (multi-)media 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, given the structure of a relational database and 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. For example, when considering a product database that contains price information for individual items, it is trivial to formulate a query that selects all items above a specific price threshold.

Retrieving multimedia data, on the other hand, comes with indirections due to the unstructured nature of the content, the feature representations used as a proxy for it and the *semantic gap* associated with these representations. A very popular model to work with the feature representations involves calculation of (dis-)similarity scores from the features and sorting and ranking of items based on this score. This is commonly referred to as the *vector space model* of multimedia retrieval and similiarty search. Over the years, many different combinations of features and ranking models have been proposed to facilitate content-based retrieval of different media types, such as, images, audio or video as have been different types of query formulation, such as *Query-by-Example*, *Query-by-Sketch* or *Query-by-Sculpting* [BGS+20].

Add sources: Survey for each modality and query type

When looking at concrete system implementations that facilitate interactive multimedia retrieval for an end-user today, the lines between multimedia retrieval and multimedia analytics quickly start to blur. This is because, in addition to the extraction of appropriate features and the conception of effective ranking algorithms, multimedia retrieval systems today also concern themselves with aspects such as query (re-)formulation and refinement, results presentation and efficient exploration [LKM+19]. In addition, multimedia retrieval systems do not simply operate on features representations anymore but combine *similarity search* on features and *Boolean retrieval* on annotations and metadata [RGL+20]. Therefore, one could argue that multimedia retrieval systems can be seen as a very specific type of multimedia analytics system. This makes all the arguments made about data processing and data management requirements for multimedia analytics applicable to multimedia retrieval as well.

1.2 Research Gap and Objective

It has been pointed out by Jonson et al. [JWZ+16] (p. 296) that "Multimedia analytics state of the art [...] has up to now [...] not explicitly considered the issue of data management, despite aiming for large-scale analytics.". Despite recent advances and the development of concrete architecture models for multimedia database management [GS16; Gia18] and implementations that refactor data management into distinct system components [Ros18], that statement, to some extent, still holds true today. While [GS16; Gia18] make important contributions towards a unified logical data-, query- and execution model required for effective search and exploration in multimedia collections, scalability aspects and the need for near real-time query performance, especially in the face of dynamic data, are not systematically considered. On the contrary, the proposed models – despite being seminal for data management in specific multimedia retrieval applications – postulate assumptions, that have considerable impact on the practical application of data management systems following them.

The starting point for the research described in this thesis is therefore the current state-of-the-art for data management in multimedia retrieval and analytics as briefly touched upon in the previous sections. Starting from and inspired by 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 in *multimedia data management* 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. These assumptions that are namely:

Assumption 1: Staticity of data collections Most multimedia retrieval systems today make a 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 data management system take place. Usually, no changes to the data collection are being made during the online phase. This model is proposed by both [Gia18] and [Ros18] and to the best of our knowledge, most existing multimedia retrieval and analytics systems implement this either explicitly or implicitly. 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.

Assumption 2: Similarity search is nearest neighbor search The vector space model used in multimedia retrieval relies on a notion of similarity search that is usually expressed as finding the k nearest neighboring feature vectors $\vec{v}_{i \in [1,k]} \in C$ to a query vector $\vec{q} \in \mathbb{R}^d$ in a collection $C \subset \mathbb{R}^d$ given a certain distance function. Very often, metrics such as the Euclidean or the Manhattan distance are employed for this comparison. While this model is very concise and rather simple, it merely allows for the ranking of potential results and, given that the underlying model and the query is precise enough, finding the relevant or desired item(s).

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 path in terms of cost parameters such as required I/O, CPU and memory usage. In multimedia retrievak, 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 often selected explicitly by the user issuing the query.

It is worth noting, that Assumption 1 and 3 both go against well-established design principles usually found it modern database systems. While it may be convenient from a perspective of system design, to assume a data collection to be static, such a mode of operation is utterly limiting when considering data that is subject to change, as is the case, for example, when doing real-time analysis or when having an application with CRUD support. The same argument can be made for manual index selection. Such an assumption may be simplifying the process of query planning but assumes, that a user is a technical expert. Furthermore, it limits the amount of optimization that can be applied by the data management system especially in the face of non-static data collections or changing query workloads.

As for Assumption 2, one can state that the described model is only able to accommodate the search-end of the *exploration-search axis*, assuming that features are, in fact, real valued vectors. It quickly becomes unusable for tasks such as browsing, structuring and summarization, delegating the required data processing to upper-tier system components. Refering to [JWZ$^+$16], it would however be desirable to offer such primitives at the level of data management.

1.2.1 Research Questions

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

Table 1.1 List of research questions resulting from challenging assumptions(AS) 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 changes to data and to what extent?	AS 1
2	Can we estimate and quantify deterioration of retrieval quality of index structures from RQ1 as changes are being made to the underlying data collections?	AS 1
3	How can we handle index structures from RQ1 for which to expect deterioration during query planning and execution?	AS 1
4	Can we device a model that (temporarily) compensates deterioration of retrieval quality of index structures?	S 1
5	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?	AS 1 & 3
6	How would a cost model that factors in desired retrieval accuracy look like and can it be applied during query planning?	AS 3
7	Assuming the cost model in RQ6 exists, at what levels of the system can it be applied (globally, per query, context)?	AS 3
8	Is there a measurable impact (e.g., on query execution time vs. accuracy) of having such a cost model?	AS 3
9	Can we generalize the model for similarity search (i.e., the vector space model) and what is the consequence of doing so?	AS 2
10	Do the existing applications and use-cases justify a generalization?	AS 2

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 what forms it can take (video, audio, images, text + metadata etc.). This formal model has the potential of being an original contribution, since we will make very explicit assumptions about what aspects of a multimedia item there are and which ones are mutable or immutable (e.g, content vs. annotations, metadata, features 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

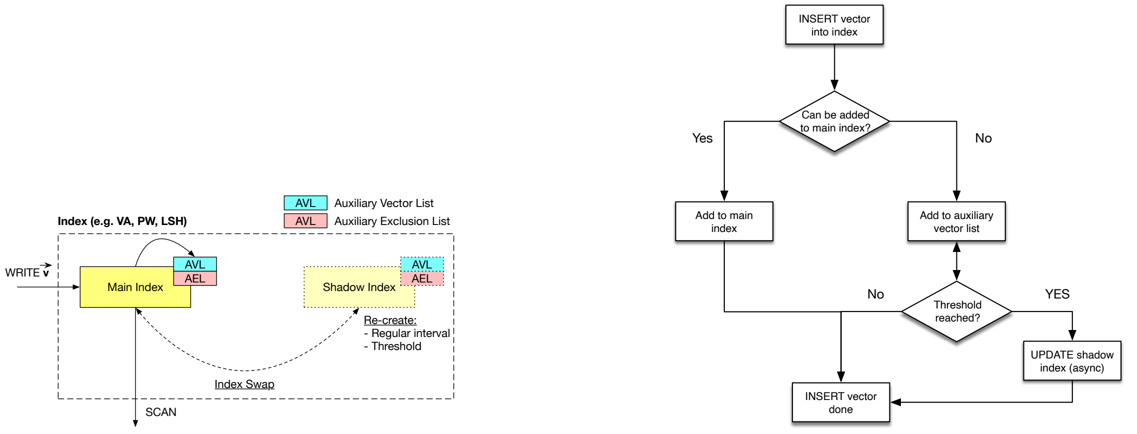

Figure 6.1 Adaptive index structures overview.

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

- 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

[BTVG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006.

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

[BGS$^+$20] Samuel Börlin, Ralph Gasser, Florian Spiess, and Heiko Schuldt. 3D Model Retrieval Using Constructive Solid Geometry in Virtual Reality. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 373–374. IEEE, 2020.

[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`.

[KKE+10] Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. Mastering the Information Age: Solving Problems with Visual Analytics, 2010.

[LKM+19] Jakub Lokoč, Gregor Kovalčík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. Interactive Search or Sequential Browsing? A Detailed Analysis of the Video Browser Showdown 2018. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–18, 2019.

[Low99] David G Lowe. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. Ieee, 1999.

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

[RGL+20] Luca Rossetto, Ralph Gasser, Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, Tomas Soucek, Phuong Anh Nguyen, Paolo Bolettieri, Andreas Leibetseder, et al. Interactive Video Retrieval in the Age of Deep Learning - Detailed Evaluation of VBS 2019. *IEEE Transactions on Multimedia*, 2020.

[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