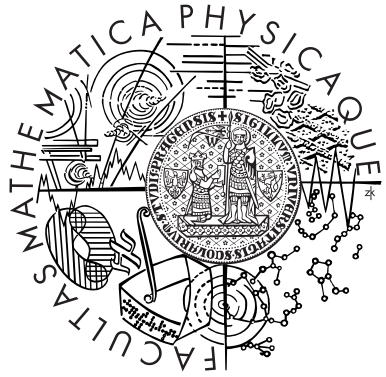
# Methods for Content-based Interactive Retrieval

Jakub Lokoč

Habilitation Thesis




Charles University
Faculty of Mathematics and Physics
Department of Software Engineering
Prague, 2019

Methods for Content-based Interactive Retrieval

Habilitation thesis

Jakub Lokoč
April, 2019

lokoc@ksi.mff.cuni.cz
http://www.ksi.mff.cuni.cz/∼lokoc
http://siret.ms.mff.cuni.cz/lokoc

Charles University
Faculty of Mathematics and Physics
Department of Software Engineering
Malostranské nám. 25
118 00, Prague 1
Czech Republic

# Acknowledgments

I would like to thank to all my colleagues and co-authors of my papers for their valuable time, contributions, influential discussions and intensive co-operation. My thanks go also to members and Ph.D. students of the *Siret Research Group* (SRG) which is led by Prof. Tomáš Skopal. In addition, I thank to all my excellent students whose Bachelor/Master theses and SW projects helped to investigate open research problems and examine promising ideas. I am very grateful for the opportunity to visit for a longer period Prof. Thomas Seidl at the RWTH Aachen University in Germany and Prof. Laszlo Böszörmenyi at the Alpen Adria University in Klagenfurt, Austria and to get inspired with their interesting work and research topics. For helpful comments, suggestions, and ideas that enabled to improve our results I thank to all the anonymous reviewers of our publications, audience at our presentations, and all the scientists I had the pleasure to meet and talk to. I am also grateful for the grants supporting my research, namely the Czech Science Foundation (GAČR) projects GAČR P202/11/0968, GAČR P202/12/P297, GAČR 201/09/0683, GAČR 15-08916S, GAČR 17-22224S, GAČR 19-22071Y and several projects from the Charles University Grant Agency. And, last but not least, I am very thankful to my family whose support enables me to do the job I like.

# Contents

**1 Thesis objectives** **1**

**2 Introduction** **3**
2.1 Motivation . . . . . 3
2.2 Elementary formal background . . . . . 5
2.2.1 Database ordering . . . . . 6
2.2.2 Vector spaces . . . . . 8
2.2.3 Distance and metric spaces . . . . . 10
2.3 Basic multimedia search approaches . . . . . 11
2.3.1 Metadata-based search . . . . . 12
2.3.2 Content-based search . . . . . 14
2.4 Indexing data structures for similarity search . . . . . 16

**I Models based on feature signatures** **19**

**3 Commentary for Part I** **20**
3.1 Motivation . . . . . 20
3.2 Feature signatures . . . . . 23
3.3 Efficient retrieval using adaptive distance measures . . . . . 26
3.3.1 Distance-specific approaches . . . . . 26
3.3.2 Metric indexing . . . . . 28
3.3.3 Ptolemaic indexing . . . . . 31
3.3.4 Parallel computing . . . . . 32

**4 Approximating the Signature Quadratic Form Distance Using Scalable Feature Signatures** **35**

**5 On Indexing Metric Spaces Using Cut-regions** **37**

**6 D-Cache: Universal Distance Cache for Metric Access Methods** **39**

**7 Ptolemaic Access Methods: Challenging the Reign of the Metric Space Model** **41**

## II Interactive video retrieval 43

**8 Commentary for Part II** **44**
8.1 Motivation . . . . . 44
8.2 Methods for interactive video retrieval . . . . . 47
8.2.1 Video analysis and preprocessing . . . . . 47
8.2.2 Search initialization with a query . . . . . 48
8.2.3 Visualization and browsing . . . . . 51
8.2.4 Relevance feedback . . . . . 53
8.3 Interactive video retrieval evaluation . . . . . 53
8.4 Video Browser Showdown participation . . . . . 55

**9 On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015-2017** **58**

**10 Interactive Search or Sequential Browsing? A Detailed Analysis of the Video Browser Showdown 2018** **59**

**11 VIRET: A video retrieval tool for interactive known-item search** **61**

**12 Conclusions and discussion** **63**
12.1 Where are you heading models based on feature signatures? . 64
12.2 Interactive video retrieval challenges . . . . . 66

# Preface

For ages, the development of humanity is interconnected with an ability to share experience and knowledge. Already ancient nations found ways to represent, record and keep pieces of information in various forms (e.g., clay tablets, papyrus scrolls) and built libraries collecting them for future generations. For example, it is believed that the Great Library of Alexandria stored between 40.000 and 400.000 scrolls. Even though many of the scrolls were burnt during Caesar's Civil War, the accident did not stop efforts to write new scrolls, build new libraries and further improve recording approaches still limiting a faster knowledge expansion. The invention of mechanical printing in the mid-15th century credited to Johannes Gutenberg was a huge step forward for sharing knowledge with masses. It opened the doors for much more efficient reproduction and distribution of documents to many regions across the world. Libraries started to be populated with a large number of books which amplified challenges with their organization and manual indexing. Step by step, new inventions appeared to record also audio and other types of analog signals. Within few centuries, people mastered recording devices and piled up a huge amount of various documents and records. For manual management a new problem was on the horizon – how to search and access information in the pile? Librarians created and maintained huge catalogs, but it was expensive and unwieldy manual work.

The situation dramatically changed in the 20th century with the invention of modern computers and a shift to digital representations enabling algorithmic processing. For example, full-text search models [11] significantly improved accessibility of digitalized text documents. Furthermore, the greatest revolution in information amenability and availability was still to come – the invention of Internet and World Wide Web in the second half of the 20th century. Since the 21st century started, we have witnessed an incredible technological tsunami that significantly formed todays society. And the

tsunami is not weakening, it seems that it is getting even stronger. Today, the amount of information easily available to ordinary people is practically unlimited. The only condition is a pocket/wearable device with an Internet connection, which is no longer restricted just to a narrow community of people. Nowadays, almost everyone has an option to have a great library in his pocket, which boosts education, inventions, work productivity and entertainment as never before. So after just several thousands of years, people really mastered their ability to share experience and knowledge to an incredible extent. However, the new opportunities also opened a new huge pack of important complex challenges ranging from ethical, credibility and security issues to large-scale data management and information retrieval problems.

The proposed thesis aims at information retrieval [11], which is still a highly demanded research area with many open theoretical and practical challenges depending on types of data and tasks. The thesis focuses on similarity search in unstructured (e.g., multimedia or visual content [58]) data, which is attractive for many researchers and represents a very dynamic field. Many state-of-the-art trends from ten years ago [51] are getting obsolete due to new inventions. An exemplary shift is the extraordinary raise of deep machine learning approaches [64] that started to solve many problems that were considered very difficult for hand-crafted algorithmic processing. It seems that a lot of traditional approaches are now abandoned and enter a "winter" period. Nevertheless, also currently very popular models based on artificial neural networks remember several "winters". Only time will reveal, whether issues related to interpretability of accurate black-box deep models [39] will be sufficiently resolved for critical applications (e.g., diagnostic or control) or we could expect another shift towards traditional analytical and rule-based approaches.

My research has been carried out at the Faculty of Mathematics and Physics of the Charles University in Prague in years 2011–2019, mainly within the *Siret Research Group* (SRG)[1] led by Prof. RNDr. Tomáš Skopal, Ph.D. After an introductory chapter providing a motivation and short overview of popular content-based retrieval approaches, selected results are organized in two parts. Both parts start with commentaries summarizing the problematic and briefly surveying related works. The first part focuses on contributions in the area of efficient retrieval using models based on feature signatures. The second part summarizes contributions in the area of interactive video

[1] http://siret.ms.mff.cuni.cz

retrieval and evaluation. The thesis includes seven chapters formed by selected co-authored papers [101, 109, 165, 77, 102, 106, 108] from peer reviewed respected journals or conferences. The papers detail approaches related to efficient retrieval using models based on feature signatures and interactive video retrieval.

Prague, April 2019

Jakub Lokoč

# Chapter 1

# Thesis objectives

Content-based retrieval frameworks are developed to aid users with various types of information needs. In our work, we focus on large annotation-free datasets of complex unstructured objects (e.g., multimedia or 3D objects), where developed content-based retrieval models enable convenient fast access to particular objects based on an interaction with users. A typical considered interactive search scenario consists of iterative query formulation and various forms of result set visualization/inspection. This thesis presents a list of our contributions to the area of content-based interactive retrieval. In order to support responsiveness of interactive search systems, we have focused on efficient processing of similarity search queries for models enabling flexible object-specific representation of contents. In the video retrieval domain, we have designed interactive search systems for effective known-item and ad-hoc search tasks and participated in the organization of an interactive video retrieval competition fostering research in this area.

In years 2011 - 2016, the author of the thesis focused on traditional hand-crafted approaches that model contents of multimedia data using descriptors representing distributions of selected features and where the similarity between two multimedia objects is modeled as a distance function on the distributions. Given a designed or trained global feature space partitioning, the feature distributions of each object are usually aggregated into histograms with predefined fixed bins [161]. If the searched objects do not fit the partitioning (e.g., in dynamic databases), the retrieval effectiveness may drop. In such cases, the retrieval system should provide additional search options for users interactively operating the system. Models based on *feature signatures* [18] represent a flexible alternative to models based on histograms.

A feature signature is an object specific representation tailored for the modeled object. However, the flexibility of the representation comes at the cost of more complex and expensive similarity evaluation, compared to models based on histograms. Considering interactive search scenarios, slow query response times would represent an uncomfortable obstacle for retrieval systems employing feature signatures. Therefore, the main objective of the first part of this work is to investigate and design novel efficient approaches for models based on feature signatures. Most of the proposed indexing approaches were designed as more general methods that can be applied with other similarity models satisfying certain properties (e.g., metric axioms).

The second part summarizes contributions from years 2014-2019 focusing on interactive video retrieval. According to the results of the Video Browser Showdown [102] and TRECVID [9], there are classes of video retrieval tasks focusing on recall where interactive means of retrieval is necessary. In order to solve the tasks, the users have to combine frequent query reformulation, results visualization and browsing approaches. The objectives comprise both the development of methods and tools for successful interactive video retrieval and also efforts for fostering the research in the area of interactive video retrieval and evaluation (especially co-organization of the Video Browser Showdown). During the last six years, we have proposed several prototypes of interactive video retrieval tools. Whereas the first successful versions of the prototypes relied mostly on color-based search and representations based on position-color feature signatures, our recent objectives incorporated video analysis, feature extraction, and multi-modal temporal fusion strategies. The investigated approaches were confronted at international evaluation campaigns[1] and the results of the competitions were summarized and published in several comprehensive journal reports.

We conclude the thesis and outline directions of our future research in Chapter 12.

[1]Prototypes of our tool won the Video Browser Showdown competition (www.videobrowsershowdown.org) in years 2014, 2015 and 2018.

# Chapter 2

# Introduction

## 2.1 Motivation

The volume, complexity and diversity of digital data collected by devices measuring various physical effects (e.g., sound, light, temperature, scent, pressure) increase every year with new advancements of sensing and recording technologies. The devices can work practically non-stop creating huge repositories of data (Big data phenomenon) collected with a specified recording precision, providing a detailed record for observed events. For example, the lifelogging wearable devices [69] allow for digital recording of GPS locations, biometrics or multimedia snapshot sequences from our daily lives. The availability of such technologies in connection with available immense data storage resources still continue to drive the exponential data growth. Whereas the overall global volume of digital universe was estimated in 2012 to grow to 40 zettabytes[1] by 2020 [60], in 2018, Reinsel et al. [142] already reported about the so-called *Global Datasphere* estimated to grow to 175 zettabytes by 2025.

A significant portion of the datasphere take multimedia data that can be simply recorded and easily stored/uploaded online by billions of active devices owned by ordinary users or installed in cars, streets, IoT, et cetera. Furthermore, various industrial, agricultural, physical, biological, medical or smart city projects start to implement multimedia data into their standard processes, analytics and workflows. Therefore, the thesis mostly discusses multimedia data (images, videos), even though many presented models can

[1] One zettabyte equals $2^{70}$ bytes.

be applied also for different types of recorded data established or emerging in various domains (e.g., network security [89]).

Multimedia data are usually recorded and stored as multidimensional arrays of discretized measured/captured intensities, optionally accompanied with other attributes (e.g., creation time, GPS location). In addition, various metadata (e.g., author, license or keywords) could be available too. All these data can be stored in a single container file with a standardized format. Once a collection of multimedia files reaches a critical limit for manual sequential search, retrieval systems are necessary to fulfill user needs. A relatively easy task is to provide data access for needs concerning available attributes like creation time/location or other structured metadata (so called *attribute-based or structured retrieval*). For example, "find photos from Prague in June 2018" or "find all Czech movies". Since each of such available attributes has a known specification comprising name, semantic and data type/domain, a data schema can be prepared for a structured query language enabling exact query formulation. However, once the user searches for a content somewhere inside the "raw" multidimensional array of numbers (so called *content-based or unstructured retrieval*), the retrieval task becomes more difficult.

The classical formats of the collected multimedia data were designed for the mainstream usage – easy recording/playing, considering a limited storage capacity (e.g., JPEG compressed matrix of RGB pixels). However, the "raw" data comprising potentially a lot of useful but hidden information (in a large haystack) represents a challenge for multimedia retrieval systems. The users are not searching for a given (sub)matrix of numbers, they mostly ask for semantic information easily recognizable by the human perception system. However, for machines this information is hard to identify from a grid of numbers, which is often denoted as the *semantic gap problem*. It took several decades of research to identify promising directions to bridge the gap. After all, the results of biological evolution of the mammal brain were important inspiration for the current successful trend for bridging the gap – machine learning employing specially designed models of artificial neural networks [64]. In addition to searching by provided semantic concepts, several other types of representations are usually extracted from the raw data to aid users with various information needs.

Before we proceed with the description of models and popular approaches in the following sections, we close the motivation section with three important objectives of multimedia retrieval systems:

- *Effectiveness* of a retrieval system is a capability to meet and fulfill user information needs, which is probably the most crucial property of the system. However, assessing the effectiveness of information retrieval systems is a difficult problem on its own as information needs are often subjective and thus unwieldy evaluations with users are necessary. In order to foster research and optimize comparative evaluation processes for new models, repeatable automatized laboratory-based methodologies were established (e.g., Cranfield paradigm).

- *Efficiency*, representing the speed of query evaluation, is gaining in importance with the increasing size of the collections where sequential query processing can be too time consuming for a given model. In cases when it is impossible to design a sufficiently efficient solution, the systems often switch to approximate retrieval that can trade the effectiveness for efficiency. However, such trade-off is possible only if the effectiveness is not the most crucial part of the system.

- Beside the classical trade-off between effectiveness and efficiency of retrieval models, systems can provide also intuitive and responsive *interactive interfaces* in order to let (ordinary) users to conveniently query and browse a multimedia database. For many types of retrieval tasks, integration of users into the search process and design of suitable interactive interfaces improves search performance (e.g., as demonstrated for video retrieval [102]).

## 2.2 Elementary formal background

Throughout the presented work, classical mathematical models provide an essential abstraction for the description of data representations, relevance scoring and efficient retrieval processes. Even though a multimedia file is already a digitized approximation of a real world signal for computers, the file can be still treated just as a desired instance of a real world object satisfying user needs. Given a maximal size $m$ (in bytes) of a multimedia file in a given format, the universe $\mathcal{U}$ of all considered distinct multimedia files is finite ($|\mathcal{U}| \leq 2^{8m}$). In practice, only a subset $\mathcal{S} \subset \mathcal{U}$ forms a database which is the subject of data management and retrieval. It is important to note that multimedia retrieval becomes a challenge from a certain size of $\mathcal{S}$, while searching a collection of a few photos or searching a shot in one

known film can be satisfactorily solved by sequential browsing. However, for large databases it is necessary to prioritize some elements of $\mathcal{S}$ before others, based on a specification of search intents in a provided query interface. In our work, we have focused on similarity/dissimilarity based relevance score models inducing a ranking on a database $\mathcal{S}$ with respect to a query. The following text summarizes just basic elementary concepts from the perspective of investigated methods. For a comprehensive overview of various information retrieval approaches, we refer the reader to related books (e.g., [193, 151, 11]).

### 2.2.1 Database ordering

In large multimedia databases, querying represents an essential approach to investigate database subsets which contain searched objects with a high probability. To obtain such subsets, retrieval engines usually rely on relevance score models that can be used to rank database objects with respect to a user query. As a simple example, searching for previously watched videos in a web browser history could be accomplished with a scoring model based on a one-dimensional time domain attribute, provided that the time attribute is logged for visited pages and users can partially estimate the time of watching the videos. However, more advanced relevance score models are necessary for information needs targeting the contents of the searched item (e.g., semantic or visual features).

Usually, relevance score models do not use directly the original multimedia files, but operate on representations designed for a particular task. The representations are created by a descriptor extraction function $f_e : \mathcal{U} \rightarrow \mathbb{U}$ that captures a distribution of suitable features present in a database object, and maps the object to a representation universe $\mathbb{U}$. The particular dataset is then represented as a set $\mathbb{S} \subset \mathbb{U}$, where each element from $\mathbb{S}$ is called *object descriptor*. In order to search the dataset, users provide queries from a universe $\mathbb{U}_\mathbb{Q}$ considered for a given retrieval scenario over $\mathbb{U}$. A general relevance score model can be defined for a data representation universe $\mathbb{U}$ and query object representation universe $\mathbb{U}_\mathbb{Q}$ as a function $\sigma : \mathbb{U}_\mathbb{Q} \times \mathbb{U} \rightarrow \mathbb{R}$.

There are two different concepts to be considered when designing a relevance score model – similarity and dissimilarity based scoring. For the similarity based concept, the higher score models more similar/relevant items with respect to a query. The dissimilarity based concept is motivated by a geometric intuition, where closer objects are more similar. Hence, the lower

score models more relevant items.

When designing an effective relevance score model, an essential goal is to find a function that corresponds to user judgments. However, this is a very difficult task as the relevance assessment is subjective, comprises cognitive processes, depends on the type of retrieval needs, domain knowledge, and context. For example, in the medical domain there might be topics with a strong disagreement among judges [126]. Relevance score models frequently incorporate the notion of similarity, which is used by individuals to categorize and classify stimuli. In order to deal with similarity perception/assessment and design proper mathematical models, psychological studies are necessary [180, 152]. On the other hand, relevance score functions can be successfully approximated in practice by simplified models for specific multimedia retrieval processes involving user interaction. In other words, the absence of perfect relevance score models can be bridged by the ability of users to interactively inspect a candidate set and identify relevant items. In addition, retrieval systems supporting more relevance score models provide users an additional interaction option to switch between models based on their observed performance.

Despite subjectivity of relevance score assessment, the effectiveness of a relevance score model is often evaluated automatically in a laboratory setting using so-called ground-truth datasets (e.g., [129, 140, 52]) representing specific retrieval tasks. Given a ground-truth dataset, the effectiveness of a relevance score model can be evaluated by various performance indicators assessing ranked result sets. Two popular concepts for the indicators are *precision* defined as $|Relevant \cap Result|/|Result|$ and *recall* defined as $|Relevant \cap Result|/|Relevant|$, where $Result \subset \mathbb{S}$ denotes the result of a query (e.g., top $k$ ranked items) and $Relevant \subset \mathbb{S}$ represents the set of relevant database objects with respect to the query. For example, precision-recall curve is used to present the dependence of precision on recall, considering a retrieval model with a threshold controlling the number of relevant objects in the result set. Both indicators can be combined into a single score (e.g., F1-score or average precision). For immense databases with a lot of relevant results and users satisfied with a fraction of them, precision at k represents another popular measure. The indicators of effectiveness are often used to compare different models on a particular ground-truth dataset. The evaluation of interactive search systems is discussed in Part II.

This thesis considers relevance score models that rely on a rigorous definition of query and descriptor universes, and employ a similarity/dissimilarity

based relevance score model. Please note that a special case $\mathbb{U}_{\mathbb{Q}} = \mathbb{U}$ is often assumed as well as additional restrictions on selected relevance score functions. Based on considered restrictions on the particular representation universe $\mathbb{U}$, two formal frameworks are frequently utilized for multimedia retrieval – vector and distance spaces.

### 2.2.2 Vector spaces

A multi-dimensional vector of $n$ numerical attributes $v \in \mathbb{R}^n$ is probably the most popular descriptor type in the area of information retrieval. For example, documents are usually transformed to vectors using word embeddings, while images/videos are preprocessed with extraction models to feature vectors. In addition, other mathematical entities can be considered for the description of content-based features. A matrix-based descriptor $x \in \mathbb{R}^{n \times m}$ can be used to represent a feature map of an image, or an object can be modeled by a feature representation function $F : \mathbb{R}^n \to \mathbb{R}$ [18]. For such representations, it is often desired to handle various data processing, organization and retrieval operations with standard linear algebra methods. The concept of vector spaces over a field represents a popular formal framework for mathematical entities with provided addition and scalar multiplication operations satisfying a set of axioms. The following definitions of a vector space and norm, considering the field of real numbers $\mathbb{R}$, are taken from a linear algebra textbook for informatics [78]:

**Definition 1 (Vector space)** *A vector space over the field of real numbers $\mathbb{R}$ is a set $\mathbb{U}$ consisting of elements called vectors, with a vector addition operation $+ : \mathbb{U} \times \mathbb{U} \to \mathbb{U}$ and multiplication of a vector by scalar $* : \mathbb{R} \times \mathbb{U} \to \mathbb{U}$. The operations satisfy $\forall a, b \in \mathbb{R}, u, v \in \mathbb{U}$: $(\mathbb{U}, +)$ is an Abelian group with the neutral element $o$, $a * (b * u) = (ab) * u$ (associativity), $1 * u = u$, $(a + b) * u = a * u + b * u$ and $a * (u + v) = a * u + a * v$ (distributivity).*

Considering vector spaces for multimedia modeling and retrieval is motivated by sound formal rules for various desired operations, for example:

- Unlike general distance spaces remembered in the next section, vector spaces enable construction of new vectors "out of" the dataset $\mathbb{S} \subset \mathbb{U}$. For example, it is possible to define a centroid vector $c_i \in \mathbb{U}$ for all $n$ dataset vectors $o_i \in \mathbb{S}$ as $c_i = \frac{1}{n} * (o_1 + \cdots + o_n)$, modify the database

by subtracting the centroid vector from all database objects, or update a query vector based on relevance feedback [143].

- Given a finite basis, it is possible to introduce the vector space dimensionality $n$ and isomorphism $\mathbb{U} \rightarrow \mathbb{R}^n$. Hence, data processing and management approaches can operate directly in the coordinate space $\mathbb{R}^n$ and employ its standard concepts. For example, hyperplanes $\sum_{i=1}^{n} a_i x_i = b$ that can be used to organize database objects to buckets.
- Considering two vector spaces over a field $\mathbb{R}$ with a finite basis, it is possible to define linear mappings between them. The mappings can be described by matrix operations over the corresponding coordinate spaces. For example, a mapping to a suitable subspace can be used to conveniently reduce the number of dimensions of a designed representation (e.g., for the reduction of the representation size or data visualization).

For many retrieval applications, it is necessary to define a non-negative quantity for each vector, so-called *norm*. The quantity corresponds to the length of the vector, where the zero vector should correspond to zero length, while the length of all the other vectors should be strictly positive. The norm is defined as:

**Definition 2 (Norm)** *Let $\mathbb{U}$ be a vector space over the field of real numbers $\mathbb{R}$ with operations $+$, $*$. A norm is a mapping $\|.\| : \mathbb{U} \rightarrow \mathbb{R}$ satisfying $\forall c \in \mathbb{R}, \forall u, v \in \mathbb{U}$ the triangle inequality $\|u + v\| \leq \|u\| + \|v\|$, $\|c * v\| = |c| * \|v\|$ and $\|v\| \geq 0$, where $\|v\| = 0$ only for zero vector.*

Considering that each image is modeled as a feature vector $v \in \mathbb{R}^n$, the Euclidean norm $\|.\|_2$ can be used to compute a dissimilarity based relevance score with respect to a query $q$ as $\|q - v\|_2 = (\sum_{i=1}^{3}(q_i - v_i)^2)^{1/2}$. The relevance score computed for all database objects then induces a ranking on $\mathcal{S}$ that could be used to focus on database images with similar features.

Another popular approach to model a similarity based relevance score between two vectors from $\mathbb{R}^n$ is the cosine similarity [197, 11], employing also the dot product. The similarity is defined for $q, v \in \mathbb{R}^n$ as $s_{cosine}(q, v) = \sum_{i=1}^{n}(q_i \cdot v_i)/(\|q\|_2 \cdot \|v\|_2)$. The cosine similarity of two vectors ranges from -1 to 1, assigning the highest score 1 to vectors pointing in the same direction and the lowest score -1 to vectors pointing in the opposite direction. Please

note that the similarity of two vectors can be easily turned to a distance by $1 - s_{cosine}(q, v)$ [162].

### 2.2.3 Distance and metric spaces

The concept of distance spaces represents a universal abstraction suitable for modeling multimedia representations as arbitrary digital records. There are generally no restrictions on the representation universe $\mathbb{U}$, no addition and scalar multiplication of the elements, nor their axiomatization. As a consequence, notions like the origin or coordinate system cannot be generally assumed. The concept is centered around the similarity of two objects modeled as a distance function $\delta : \mathbb{U} \times \mathbb{U} \to \mathbb{R}_0^+$, where similar elements have a small distance and vice versa. A comprehensive survey summarizing a dictionary of various distances was presented by Deza and Deza [53]. The book is also the source for the two following definitions.

**Definition 3 (Distance space)** *A distance space $(\mathbb{U}, \delta)$ is a set $\mathbb{U}$ equipped with a distance $\delta : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ satisfying $\forall x, y \in \mathbb{U}$: $\delta(x, y) \geq 0$ (non-negativity), $\delta(x, y) = \delta(y, x)$ (symmetry) and $\delta(x, x) = 0$.*

For example, a video retrieval task could be defined to search for shots with a similar frame sequence as in a provided query shot. Assuming variable long sequences where each element of the sequences is represented as a multi-dimensional point in $\mathbb{R}^n$, similarity of two sequences could be modeled by the Dynamic Time Warping distance [31]. Generally, many distance measures have been proposed for various types of tasks and spaces $\mathbb{U}$, each distance satisfying various properties. For multimedia retrieval, a popular choice are metric distances that enable also efficient retrieval using metric access methods [193].

**Definition 4 (Metric space)** *A metric space $(\mathbb{U}, \delta)$ is a set $\mathbb{U}$ equipped with a distance $\delta : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ satisfying $\forall x, y, z \in \mathbb{U}$: $\delta(x, y) \geq 0$ (non-negativity), $\delta(x, y) = \delta(y, x)$ (symmetry), $\delta(x, y) = 0$ if and only if $x = y$ (separation) and $\delta(x, y) \leq \delta(x, z) + \delta(z, y)$ (triangle inequality).*

Modeling data as objects of a metric space provides freedom for the choice of representation universe $\mathbb{U}$ and at the same time provides clues for database experts designing indexing methods [193]. However, there has been also criticism towards metric axioms. For example, Tversky [180] analyzed features

of similarity in a psychological review and demonstrated that similarity judgments do not have to be symmetric or transitive. In order to benefit from both unrestricted modeling of distance spaces and metric indexing structures, Skopal [162] presented a unified framework for metric and non-metric distance spaces.

We end the subsection with definitions [193] of two frequently used types of queries for distance based similarity search. In addition to the query object, both types of queries consider also a constraint limiting the number of returned objects from the database.

**Definition 5 (Range and kNN queries)** *Given dataset* $\mathbb{S} \subset \mathbb{U}$*, a distance function* $\delta$*, a query object* $q \in \mathbb{U}$ *and* $r_q \in \mathbb{R}_0^+, k \in \mathbb{N}$*, the range query* $R(q, r_q) = \{o \in \mathbb{S}; \delta(q, o) \leq r_q\}$ *and the* $k$ *nearest neighbor query* $kNN(q) = \{\mathbb{X} \subset \mathbb{S}; |\mathbb{X}| = k \wedge \forall x \in \mathbb{X}, y \in \mathbb{S} - \mathbb{X} : \delta(q, x) \leq \delta(q, y)\}$.

Whereas the range query requires a domain knowledge to set up the query radius appropriately, the k-NN query enables a convenient selection of the $k$ most relevant objects for the price of more complex query processing algorithms. Examples of both types of queries are presented in Figure 2.1.

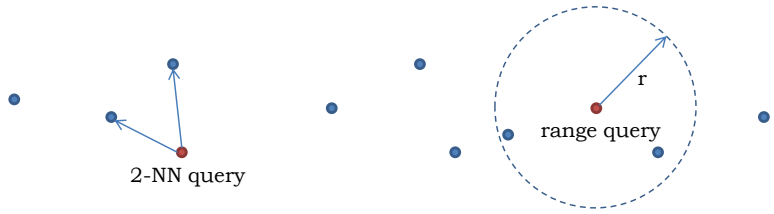



Figure 2.1: 2-NN and range queries using the Euclidean distance in $\mathbb{R}^2$.

## 2.3 Basic multimedia search approaches

In this section, we remember two popular types of multimedia search approaches.

### 2.3.1 Metadata-based search

A popular intuitive option to search a large collection is by making use of keywords describing semantic concepts in multimedia objects or matching related metadata[2]. Given sufficiently annotated objects, multimedia retrieval can be handled by classical text-based retrieval models (e.g., boolean or vector space models [11]) supported by highly efficient inverted indexes [197]. However, the multimedia data are usually insufficiently annotated or not annotated at all (except basic metadata). Hence, an effective assignment of keywords is an important challenge to provide keyword search. After a discussion of traditional methods to collect annotations, several deep learning based techniques enabling automatic content-based annotation are highlighted.

The most straightforward way to assign keywords is to employ users (or domain experts in cases a deep domain knowledge is required) and let the users to annotate the data. However, such approach is not always feasible. Especially for immense collections and non-trivial annotation tasks a large crowd of experienced annotators is necessary. On the other hand, for simple annotation tasks (e.g., marking objects on photographs) there are already well-established web portals considering crowd-sourcing as a powerful way to annotate data (e.g., the Amazon Mechanical Turk [144]). It has been also shown, that the power of the crowd can help experts to be more effective and efficient [55]. Furthermore, there are emerging specialized social networks that can connect a large number of experts, allowing them to annotate and discuss data related to their domain/expertise [173]. As a side effect, a lot of metadata can be extracted from such social networks, enabling various forms of retrieval in the underlying multimedia data.

The World Wide Web is also effectively mined for automatic annotations of multimedia data appearing on web pages. Google[3], Bing[4], Yahoo[5] and other titans of the Internet searching have demonstrated that associating specific keywords from the surrounding text with the multimedia data can be used for effective and efficient retrieval (without analyzing the content of the data). As the most popular search portals have millions of users, the portals

[2]A special type of metadata are also structured attributes (e.g., date, author, GPS) associated to multimedia data, which can be used to search data by traditional database models (e.g., relational model).

[3]www.google.com

[4]www.bing.com

[5]www.yahoo.com

can also analyze the behavior of the crowds during searching and use all the collected information to improve the rankings of the results. Although this approach seems to be promising for popular topics where many users share the same search intents, effective automatic annotation of multimedia data represents an intensively investigated topic [44].

Modern web-based technologies fostered initiatives for the collection of large annotated collections, especially image datasets [45, 52, 177]. As a prominent example, a large-scale hierarchical image database ImageNet [52] was introduced collecting a large number of example images for WordNet [121] synsets (synonym sets). In connection with highly parallel hardware and new machine learning approaches [64], significant breakthroughs towards automatic content-based annotation were achieved. Especially new deep convolutional and recurrent neural network models significantly improved keyword-based accessibility of annotation-free multimedia data. The ImageNet Large Scale Visual Recognition Challenge [149] (ILSVRC) was at the center of rapid development of deep convolutional neural network architectures for image classification or object detection. In 2012, Krizhevsky et al. [90] proposed a deep network architecture called AlexNet that not only won the image classification task, but the victory was so significant that it completely changed the course of computer vision research[6]. The AlexNet architecture considered rectified linear units, dropout regularization and combined convolutional, max pooling, normalization and fully connected layers. In the following years, various deep architectures were developed and gradually enhanced by many research teams. Within just three years, deeper models implementing new features like inception modules [174] or residual connections [74] brought the ILSVRC classification performance to the human level. The convolutional architecture was adopted also for object detection, resulting in fast networks predicting directly both concepts and corresponding bounding boxes (e.g., YOLO [141] or SSD [98] models). Please note that the bounding box detection significantly enriches the annotation, as users can issue textual queries with spatial localization. Image captioning represents another investigated direction of automatic annotation [188, 79], focusing on complex textual image descriptions comprising present objects and their relations. In video annotation, a fundamental and difficult task is action classification [63] as videos exhibit a higher variability than still images, while the

[6]During the last seven years, the paper received more than thirty eight thousand citations at google scholar.

available training sets of video actions are small. All the presented content-based methods can be used to improve findability of searched concepts using convenient query specification with keywords. However, the effectiveness of the annotation models is not yet sufficient for all types of data and tasks (especially in general video domain [9, 102]) and also the users might face problems to express their search needs by a suitable textual query.

### 2.3.2 Content-based search

Many retrieval systems complement keyword search with the query-by-example paradigm, where users provide a query object (e.g., image or shot) to enter their search intents. The systems transform the content of the query object into a suitable form for employed representation universe $\mathbb{U}$. Subsequently, the transformed representation is employed in the similarity search retrieval process. There are several sources, where users can take a query object representing their search intents. In the systems combining keyword-based and query by example retrieval, the query object can be obtained using previous keyword-based search. The query object can be also painted as a sketch, or uploaded by a user from his/her camera. In some retrieval scenarios, the query object does not have to necessarily represent a clear search intents of a user. For example, in multimedia exploration [24] users can just browse and investigate an unknown data collection. In such scenarios, the query object can represent just a mediator (or link) to quickly browse to another view of the data. In specific retrieval tasks, the query can be formulated directly using descriptor features. Especially descriptors that are designed as sets of features with clear semantics for users can be used to formulate query sketches. For example, users may specify simple color regions in a sketch canvas to search for video frames (or a sequence of frames) represented by position-color feature signatures [103, 35, 34].

During last decades, a vast number of content representation approaches have been proposed [51, 196]. A good representation has to comprise features for effective similarity modeling and should facilitate efficient evaluation. Recent survey by Zhou et al. [196] categorizes content-based image representation issues to feature extraction, visual codebook learning, spatial context embedding, feature quantization and feature aggregation. Feature extraction methods map the original object to a set of characteristic features, where the mapping function can be either hand-crafted or learned. Visual codebook learning aims at detection of a representative set of visual words for the

extracted features. A popular learning approach is unsupervised clustering (e.g., k-means or its variants [5]). The codebook is used to transform a set of extracted features to a fixed-length vector [161]. Spatial context embedding approaches try to enhance codebooks with contextual information, while feature quantization methods deal with efficient assignment of visual words to features. Finally, feature aggregation focuses on ways to accumulate assigned features and produce the final representation. Popular approaches for effective retrieval with codebooks comprise re-ranking using spatial information [140], Hamming embedding [84], compressed Fisher vectors [139] or vectors of locally aggregated descriptors [86].

With the achievements of deep convolutional neural networks in classification tasks [90], new descriptor extraction approaches were revealed as well. Donahue et al. [54] have demonstrated that activation features from a selected layer of a trained deep convolutional neural network can be employed for novel generic tasks. A vector of extracted activations from a selected (retrained) network has become a popular representation for generic similarity search. Features from deep convolutional neural networks proved to be also effective for particular object retrieval [178].

Retrieval effectiveness can be further enhanced with additional strategies. Multi-modal fusion [8, 38] usually provides performance gains as different types of features provide additional view of data (including semantic concepts). The features are usually combined using early or late fusion strategies. Another approach to improve effectiveness is query expansion that has been introduced for image retrieval by Chum et al. [46]. Given an automatically refined result of an initial query (e.g., by spatial verification [140]), an expanded query is re-issued to boost recall. Retrieval effectiveness can be also improved by relevance feedback approaches [143, 116, 57], incorporating implicitly/explicitly formulated feedback from users. A classical approach is the Rocchio algorithm [143] developed for the vector space model. The algorithm employs provided sets of relevant and non-relevant objects/vectors to update the query vector. Based on balancing weights, the query is navigated away from non-relevant objects towards relevant objects. Another example of a relevance feedback model is a statistical Bayesian framework [57] for class search, where in each iteration users provide feedback by selecting one image from an appropriately selected display. The feedback is used to update the estimated probability that a database image belongs to the searched class. Recently, it has been demonstrated by the Blackthorn approach [192] that a relevance feedback method can be incorporated also in interactive multi-

modal learning frameworks considering huge datasets (up to one hundred million images).

## 2.4 Indexing data structures for similarity search

Given a large multimedia database, indexing data structures aim at efficient data organization and query processing. A suitable indexing strategy depends on several factors like the considered representation universe, type of query, or available hardware. In the following, only a brief overview of selected successful ideas for efficient distance-based similarity search is presented, considering range or k-NN queries defined in Section 2.2.3. For a broader summary of various indexing approaches and types of queries, we refer the reader to comprehensive surveys and monographs (e.g., [36, 42, 193, 151]).

For range and k-NN queries, users ask only for a set of the most similar database objects with respect to a given query object, distance measure and query constraint. The constraint represents an important clue to process the queries more efficiently. For example, the range query has a distance threshold parameter that can be employed in filtering rules safely discarding groups of objects outside the "query ball". Generally, the filtering rules depend on a particular data organization and indexing method. Please note that it is necessary to distinguish non-relevant objects with respect to information needs and non-relevant objects with respect to the currently issued type of query (e.g., database objects outside the "query ball").

The form of the representation universe $\mathbb{U}$ significantly affects the design options for data indexing structures. Usually, the multimedia objects are modeled as descriptors from a multi-dimensional space $\mathbb{U} = \mathbb{R}^n$ or more general distance space $(\mathbb{U}, \delta)$. For the multi-dimensional spaces, indexing methods can employ tools of vector spaces to organize data and design filtering rules for efficient query processing. Popular indexes for low-dimensional data organization comprise a regular grid partitioning, k-d trees [30] relying on binary disjoint partitioning by splitting hyperplanes, or disk oriented balanced R-trees [70] organizing data to a hierarchy of (potentially overlapping) minimum bounding rectangles/boxes. However, for high-dimensional spaces the so-called *curse of dimensionality* problem [36] prevents from the design of efficient filtering rules and so approximate search strategies are often employed. For example, locality-sensitive hashing approaches [50, 137]

or locality preserving mappings to one dimensional domain (e.g., based on Z-curve [191]) have been proposed to efficiently find a good approximation of searched nearest neighbors. To scale-up search engines to billions of descriptors, distributed approaches were proposed to handle huge volumes of data and speed up query processing (e.g., for Map-Reduce frameworks [122]). Large volumes of high-dimensional data can be managed also with a memory index employing short binary codes (e.g., based on product quantization [85]) enabling estimation of the original distances for filtering of non-relevant objects. Another way to face a high number of dimensions is to consider feature selection or dimension reduction approaches [71]. For example, the principal component analysis [138] represents a popular dimension reduction approach trying to minimize information loss by a linear mapping of centered data to a new coordinate system determined by eigendecomposition of the corresponding data covariance matrix. The organization of high-dimensional data with respect to a considered distance function is another option to face the curse of dimensionality, where the so-called *intrinsic dimensionality* becomes an indexability indicator of the distance space [42].

General distance based indexing approaches consider the set of properties of a distance function $\delta$ and often treat the universe $\mathbb{U}$ as a black-box. The motivation for relying just on the properties is to develop universal indexing approaches for a provided arbitrary data representation and distance function satisfying the properties. Metric space indexing (see Section 3.3.2) is an exemplary approach, where metric axioms are employed to define sound filtering rules for groups/partitions of objects. Please note that for costly distance measures, avoiding evaluations of $\delta(q, o_i)$ between the query object $q \in \mathbb{U}$ and database objects $o_i \in \mathbb{S}$ speeds up the search. The key axiom is the triangle inequality enabling an efficient estimation of the lower bound of the distances between a query and database objects. Popular metric indexing/access methods comprise memory based pivot tables (e.g. LAESA [120]), disk based dynamic structures (e.g., M-tree [47], PM-tree [166]), structures designed for effective approximate search [2, 41] or distributed structures (e.g., M-index [131]). Beside the triangle inequality, there are classes of distances satisfying additional properties enabling efficient retrieval. For example, ptolemaic distances enable estimation of the lower bound based on two reference points as discussed in Section 3.3.3. Another example is a subclass of metric spaces satisfying the four-point property, enabling to embed each four objects into the three dimensional Euclidean space (i.e., all six distances are preserved between the embedded objects). As investigated by Connor et

al. [49], the property can be used to derive additional geometric guarantees for more efficient filtering.

For some specific distances and additional conditions, there were proposed highly efficient retrieval mechanisms and structures. For example, the cosine distance between a query object and the whole dataset can be evaluated efficiently using inverted files, if the query vector is very sparse. Only for the query non-zero dimensions, corresponding lists of database objects are visited and processed. Another example considers Hamming space representations. With the introduction of hardware support for counting bits (*popcnt* instruction), the Hamming distance can be evaluated with a few CPU instructions. Hence, efficient similarity search can be achieved by training a descriptor extraction model designed to provide binary hash codes for data objects (e.g., as demonstrated for images [96]). Binary codes (so-called *sketches*) can be obtained also for objects represented in a general metric space $(\mathbb{U}, \delta)$ using appropriately selected pairs of reference objects. Míč et al. [119] have demonstrated in a set of experiments that the codes are effective for secondary filtering of non-relevant candidate objects.

# Part I

# Models based on feature signatures

# Chapter 3

# Commentary for Part I

## 3.1 Motivation

The first part of this thesis focuses on traditional unsupervised content-based retrieval models, where object descriptors are based on a hand-crafted *feature space* $\mathbb{F}$ (usually $\mathbb{R}^n$) designed for a particular retrieval task. For example, a feature space for images can be modeled as an n-dimensional Euclidean space where dimensions can correspond to location, color information in pixels or more complex features based on statistics from the neighborhood of pixels (e.g., texture [175], SIFT [114], or SURF [16]). The author of the thesis has joined also two projects, where the feature space was defined for different domains of objects. The first project focused on malware detection in encrypted communication considering the HTTPS protocol [89]. Each message was represented as a real vector in a feature space based on a limited set of attributes usually logged by web proxies (e.g., bytes sent, bytes received, duration, and inter-arrival time). The second project focused on scalable 3D shape retrieval using robust local features [160]. The feature space was based on so-called heat kernel signatures. The heat kernel signature for a point $x$ on a compact Riemannian manifold is a function over the time domain $HKS(x,t) = k_t(x,x)$, where $k_t(x,x)$ is a special restricted case of the family of heat kernels $\{k_t(x,\cdot)\}_{t>0}$ [172]. Sampling the time for a fixed period, $k_t(x,x)$ can be used to obtain a vector representation that is robust with respect to manifold perturbations.

During last decades, there have been designed and even standardized many types of similarity models based on various features present in database

objects (e.g., the MPEG-7 standard [40, 125]). The models usually consider a descriptor modeling the distribution of selected features and a similarity/dissimilarity measure evaluating the relevance score for two descriptors. In order to aggregate the features, various methods can be employed. For example, a feature space partitioning (e.g., a regular grid) or a probabilistic model (e.g., a Gaussian mixture model) can be utilized. In this part, we consider just traditional extraction approaches based on an unsupervised partitioning of the feature space. For more advanced techniques like feature selection [71] or feature learning [87], we refer readers to the corresponding literature.

With a higher dimensionality of a feature space, the number of required bins for a regular grid based partitioning of the feature space can be too high. Therefore, adaptive binning is considered to assign extracted features $F^o \subset \mathbb{F}$ of an object $o$ to a limited number of so-called *representatives* $r_i \in \mathbb{F}$. A shared set of representatives $\mathbb{X} \subset \mathbb{F}$ is usually detected in a preprocessing step by a clustering method (e.g., k-means). Given a set of shared representatives, extracted features of a modeled object can be aggregated into a fixed-length *adaptive histogram* $h \in \mathbb{R}_{\geq 0}^{|\mathbb{X}|} = \mathbb{U}$, where each histogram bin aggregates features aligned to the corresponding shared representative [147]. Adaptive histograms enable efficient similarity evaluation, because corresponding histogram bins have the same semantics and thus cheap bin-to-bin distance measures can be considered. Furthermore, if the query histogram is sparse, highly efficient retrieval can be achieved using inverted files [197] and bin-to-bin distance measures.

Minkowski metrics $L_p(x, y) = (\sum_{i=1}^{d} |x_i - y_i|^p)^{\frac{1}{p}}$ represent a frequently used class of cheap bin-to-bin distances for d-dimensional vectors $x, y$, given $p \in \mathbb{R}, p \geq 1$. The Minkowski metrics can perform well as long as features from similar multimedia objects are aggregated to the same histogram bins. However, if the features are aggregated among several neighboring cells of the shared feature space partitioning, it may turn out that two similar objects can have dissimilar adaptive histograms considering just bin-to-bin distances. In such situations, a soft assignment coding [97] can be considered, aggregating a feature $f \in F^o$ in multiple partitions. Another option is to employ the quadratic form distance $QFD_A(x, y) = \sqrt{(x-y)A(x-y)^T}$ [72]. The distance uses a $d \times d$ positive-definite correlation matrix $A$ that can be employed to straighten the ambiguity of the descriptor extraction process on homogeneous domains. If the quadratic form distance is utilized just to model fixed correlations between the histogram bins (i.e., matrix $A$ is fixed),

the costly retrieval model employing quadratic form distance can be transformed to an equivalent but much cheaper retrieval model employing the euclidean distance [163]. The quadratic form distance can be also used to model user preferences changing over time [83], however, such a dynamic model can be efficiently indexed just using methods that can partition the descriptor space independently of a distance measure. Another approach to model similarity between two normalized feature histograms with correlated bins is the Earth Mover's Distance [148] that interprets the similarity as a transportation problem.

As an alternative to adaptive histograms considering a given set of shared representatives, models based on so-called *feature signatures* have been investigated [147, 18] as a more flexible approach to represent contents of objects. A feature signature models a multimedia object $o$ using a finite set of object-specific representatives $r_i^o \in \mathbb{F}$ with weights $w_i^o \in \mathbb{R}^+$ corresponding to the mass of extracted features assigned to $r_i^o$.

The difference between shared and object-specific representatives gets remarkable in high-dimensional feature spaces. In Figure 3.1, there are depicted three ways to represent an image using 7-dimensional position-color-texture feature space – a feature signature (Figure 3.1a), compared to histograms based on 10000 and 1000 shared representatives (Figure 3.1bc). Whereas feature signatures employing object-specific representatives can flexibly aggregate the contents of the original image, the expressiveness of histograms can suffer from the usage of shared representatives, especially for unique images in the database. In Figure 3.1b, the expressiveness of the histogram is

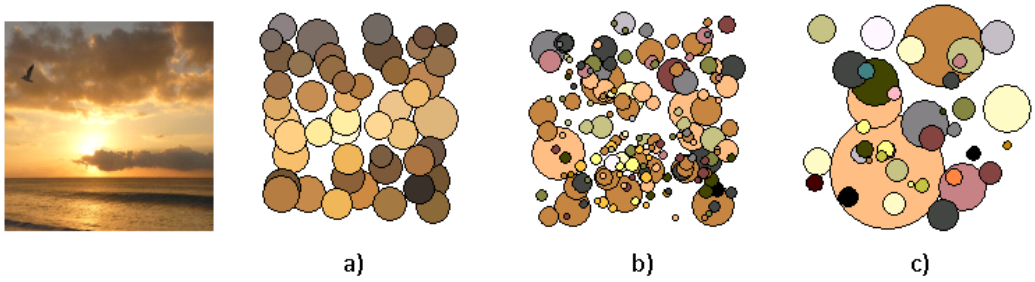



Figure 3.1: An image represented using position-color-texture feature space and **a)** feature signature, compared to adaptive histograms based on **b)** 10000 and **c)** 1000 shared representatives (only representatives with a non-zero weight are displayed).

improved at the cost of the high number of shared representatives, while in Figure 3.1c, the original image contents is just roughly approximated using the small number of shared representatives not corresponding to the modeled image.

Both approaches have pros and cons. So which descriptor is a better choice? Generally, the prior knowledge of the database matters. If a learned set of shared representatives is sufficient to create discriminative descriptors for all database objects, then adaptive histograms are the preferred efficient choice. If the dataset is dynamically changing over time, then the feature signatures can be used to flexibly represent the contents of the objects.

## 3.2 Feature signatures

Feature signatures [147, 18] enable to flexibly represent the contents of modeled objects using object specific representatives. Formally, the feature signature of a multimedia object $o$ can be defined as a finite set of tuples $\langle r_i^o, w_i^o \rangle$, where $w_i^o \in \mathbb{R}^+$ represents the importance of the representative $r_i^o$ in object $o$. For a more general unifying concept of feature representations modeled as functions $\mathbb{F} \to \mathbb{R}$ we refer readers to [18].

**Definition 6 (Feature Signature)** *Given a feature space $\mathbb{F}$, the feature signature $S^o$ of a multimedia object $o$ is defined as a finite set of tuples $\{\langle r_i^o, w_i^o \rangle\}_{i=1}^n$ from $\mathbb{F} \times \mathbb{R}^+$, consisting of representatives $r_i^o \in \mathbb{F}$ and their weights $w_i^o \in \mathbb{R}^+$.*

Feature signatures for a given feature space can be obtained using various extraction techniques. For example, a position-color feature signature for an image can be created using an image resize operation. In such case, each representative with a constant weight corresponds to one pixel of the small image thumbnail, where the resolution of the thumbnail has to be fixed in advance (see Figure 3.2c). Another approach to extract a feature signature is an adaptive k-means clustering of features of sampled/selected pixels from the image, where found centroids are used as representatives $r_i^o$. In Figure 3.2ab, the extracted position-color-texture features $F^o \subset \mathbb{R}^7$ from the pixels are assigned to representatives $r_i^o \in \mathbb{R}^7$ depicted as colored circles and weight $w_i^o \in \mathbb{R}_0^+$ (proportional to the circle radius) corresponds to $|S_i^o \cap F^o|$, where $S_i^o$ is a feature space partition determined by $r_i^o$. Although the adaptive k-means clustering can flexibly aggregate the content of the image, the

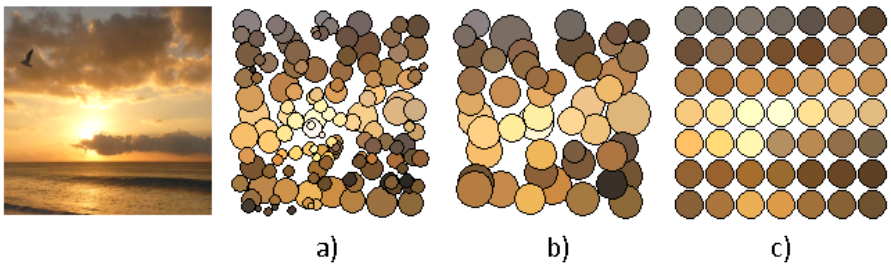



Figure 3.2: Examples of feature signature extraction techniques of an image – **a, b)** two position-color-texture feature signatures obtained by two variants of adaptive k-means clustering of the same set of sampled points and **c)** position-color feature signature obtained by image resize operation.

clustering is costly operation and thus an efficient parallel implementation of feature signature extraction is necessary for huge multimedia collections. Our technique enabling extraction of thousands of position-color-texture feature signatures per second was proposed in [91].

In order to define a similarity model based on feature signatures, adaptive similarity/distance measures capable to compare two feature signatures with different sets of representatives can be employed. The adaptive measures employ a ground distance for the representatives of two compared feature signatures [25]. During last decades, there have been proposed several adaptive distance measures like Hausdorff Distance [82], Earth Mover's Distance [148], Perceptually Modified Hausdorff Distance [135], Signature Quadratic Form Distance [28], or recently introduced Signature Matching Distance [19], and there have been also evaluated several studies comparing the distances [25, 23, 19]. The distances employ all pairwise ground distances between representatives of two features signatures, which leads to at least quadratic time complexity of the similarity evaluation.

In our work, we have mainly focused on distance spaces based on the Signature Quadratic Form Distance, because the distance spaces are effective and provide properties for efficient indexing. The Signature Quadratic Form Distance is defined [28] as follows:

**Definition 7 (Signature Quadratic Form Distance)** *Given two feature signatures* $S^o = \{\langle r_i^o, w_i^o \rangle\}_{i=1}^n$ *and* $S^p = \{\langle r_i^p, w_i^p \rangle\}_{i=1}^m$ *and a similarity function* $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ *over a feature space* $\mathbb{F}$*, the signature quadratic form distance* $\mathrm{SQFD}_{f_s}$ *between* $S^o$ *and* $S^p$ *is defined as:*

$$\mathrm{SQFD}_{f_s}(S^o, S^p) = \sqrt{(w_o \mid -w_p) \cdot A_{f_s} \cdot (w_o \mid -w_p)^T},$$

*where* $A_{f_s} \in \mathbb{R}^{(n+m)\times(n+m)}$ *is the similarity matrix arising from applying the similarity function* $f_s$ *to the corresponding feature representatives, i.e.,* $a_{ij} = f_s(r_i, r_j)$*. Furthermore,* $w_o = (w_1^o, \ldots, w_n^o)$ *and* $w_p = (w_1^p, \ldots, w_m^p)$ *form weight vectors, and* $(w_o \mid -w_p) = (w_1^o, \ldots, w_n^o, -w_1^p, \ldots, -w_m^p)$ *denotes the concatenation of weight vectors* $w_o$ *and* $-w_p$*.*

As an example of the similarity function $f_s$, the Gaussian similarity function $f_{gauss}(r_i, r_j) = e^{-\alpha L_2^2(r_i, r_j)}$ or more efficient Heuristic similarity function $f_{heuristic}(r_i, r_j) = 1/(\alpha + L_2(r_i, r_j))$ were suggested [28]. The $\alpha$ parameter can be used to fine-tune the precision, and $L_2$ denotes the Euclidean distance.

The Signature Quadratic Form Distance has not only proved to be an effective distance measure in various domains (e.g., for 3D object retrieval [160]), but also a distance suitable for efficient retrieval. Although the distance has quadratic time complexity, the distance can be used for efficient distance-based indexing. Under certain conditions [18], the distance satisfies metric/ptolemaic postulates necessary for efficient metric/ptolemaic indexing [77]. Furthermore, the $\alpha$ parameter of the distance affects not only effectiveness, but also the intrinsic dimensionality property[1] of the corresponding distance space [21]. Last but not least, we have also demonstrated that Signature Quadratic Form Distance represents a suitable task for GPU devices [92].

In the following section, we overview approaches for large-scale multimedia retrieval using models based on feature signatures and present our contributions.

[1]The intrinsic dimensionality is a crucial property for efficiency of the metric/ptolemaic indexing, for more details see Section 3.3.

## 3.3 Efficient retrieval using adaptive distance measures

Given a retrieval model $(\mathbb{U}, \delta)$ based on feature signatures and a similarity search task defined by a query object $q \in \mathbb{U}$ and a query constraint $\phi$, there can be utilized several orthogonal approaches to process such task more efficiently than just simple sequential search using the original expensive model. The approaches differ in assumptions about the similarity model and the database, whether the model is static or dynamic (i.e., descriptors and the distance can be changed), and whether the database is static or dynamic. Also the number of query objects affects the choice of the optimal solution. Despite their differences, most of the techniques share one principle for efficient filtering of non-relevant objects – *lower-bounding principle*, where a lower-bound distance $LB(\delta(q, o)) \leq \delta(q, o)$ between a query object $q \in \mathbb{U}$ and a database object $o \in \mathbb{U}$ is expected to be much cheaper than the original distance $\delta(q, o)$. Using the lower-bound distance, the query can be processed using a filter and refine approach, where the original distance is evaluated only on a fraction of the database. The lower-bound distance can be approximated, determined for a specific domains rigorously or determined using general properties of the distance measure (e.g., metric/ptolemaic properties).

### 3.3.1 Distance-specific approaches

During the last decade, there have been presented many attempts to find rigorously efficient lower-bound distances for various adaptive distance measures. However, many of the methods are usually restricted to feature histograms or the lower-bound is not too tight.

For example, a simple lower-bound for the Earth Mover's Distance is the Rubner filter [148] that evaluates the ground distance between centers of mass of two compared feature signatures $S^o, S^q$. The Rubner filter holds only if the sum of weights is the same for both feature signatures. In [6, 7, 190], the authors have presented novel dimensionality reduction techniques for the Earth Mover's Distance for filter-and-refine architectures enabling efficient exact search. The techniques are restricted just to feature histograms. In [158], the authors presented an algorithm approximating the Earth Mover's Distance in linear time. The algorithm considers the sum of absolute values

of weighted wavelet transform coefficients computed for the difference of two histograms. Recently, a new lower-bound called Independent Minimization for Signatures [184, 183] has been presented for more efficient retrieval using the Earth Mover's Distance.

There have been also attempts to find cheap distances approximating the Signature Quadratic Form Distance. In [29], the authors have demonstrated that if similarity function $f_{L_2}(r_i, r_j) = -L_2^2(r_i, r_j)/2$ is utilized, then the Signature Quadratic Form Distance becomes $L_2$-Signature Quadratic Form Distance computable in linear time, but providing worse effectiveness. In [27], the authors proposed a simple feature signature reduction technique considering removal of tuples with small weights and defined a signature quadratic form filter distance for approximate filter and refine retrieval. The filter distance computes the signature quadratic form distance using reduced feature signatures. However, according to our experiments the filter distances do not provide too tight approximations of the lower-bounds. One of the reasons is that removing a lot of tuples with small weights may significantly deteriorate the original feature signatures. On the other hand, the idea of feature signature reduction is a general approach that enables retrieval using an arbitrary adaptive distance measure.

Therefore, we have investigated advanced feature signature reduction techniques that can significantly improve the efficiency of the retrieval[2] (see Chapter 4). In [101], we have presented scalable feature signatures, a class of feature signature reduction techniques based on agglomerative hierarchical clustering [66], [56]. We have experimentally demonstrated that feature signatures can be significantly reduced at the cost of just a small loss of quality. In Figure 3.3, we may observe an example of a feature signature for the image of a sunrise and corresponding reduced feature signatures. Whereas the original feature signature flexibly approximates the contents of the original image, the reduced feature signatures at least preserve the general color layout of the image. Furthermore, the reduced feature signatures can be compared with an arbitrary adaptive distance measure, in other words, this approach is not restricted just to metric/ptolemaic distances discussed in the following sections. In [99], we have proposed a journal extension with a more comprehensive evaluation of various popular agglomerative reduction techniques implemented in the framework of scalable feature signatures.

[2]The time complexity of adaptive distance measures depends quadratically on the size of the feature signature vocabulary, i.e., the number of tuples.

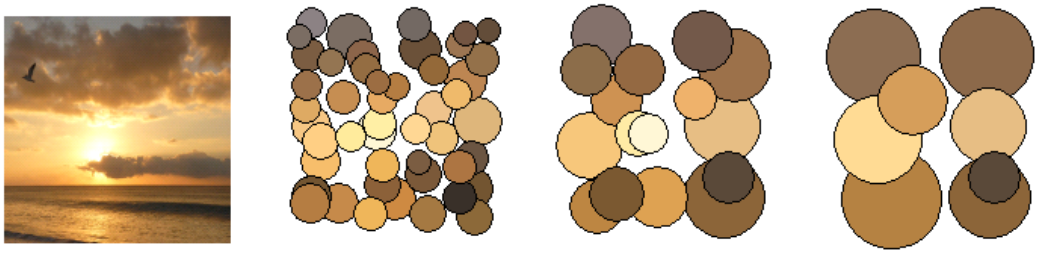


Figure 3.3: An example of a feature signature reduction, starting from left – original image, original feature signature, and reduced feature signatures comprising 32 and 16 tuples.

### 3.3.2 Metric indexing

The metric space approach [42, 193, 151] can be used to efficiently process similarity queries employing models based on feature signatures and metric adaptive distance measures. In order to process the queries efficiently, the metric space approach utilizes lower-bounding techniques that employ precomputed distances between database objects and a set of reference points $p_i \in \mathbb{P} \subset \mathbb{U}$, so-called pivots. The metric space approach assumes that the utilized distance function satisfies reflexivity, non-negativity, symmetry and triangle inequality axioms. Especially the triangle inequality axiom $(\forall x, y, z \in \mathbb{U} : \delta(x, y) \leq \delta(x, z) + \delta(y, z))$ is necessary for the correctness of the lower-bounding based on precomputed distances. More precisely, given a query object $q \in \mathbb{U}$, a database object $o \in \mathbb{U}$ and a pivot $p \in \mathbb{U}$, the lower-bound distance between $o$ and $q$ can be directly derived from the triangle inequality using precomputed distances as $LB_{\triangle}(\delta(q, o)) = |\delta(o, p) - \delta(q, p)|$, where $\delta(q, p)$ is evaluated just once before query processing and $\delta(o, p)$ is the precomputed distance stored in a metric index (see Figure 3.4).

Furthermore, the metric space approach provides also partitioning mechanisms enabling grouping of similar objects into partitions so the whole groups of objects can be filtered during query processing. For example, given $p \in \mathbb{U}, r \in \mathbb{R}_0^+$, $Ball(p, r) = \{o | o \in \mathbb{U} \wedge \delta(p, o) \leq r\}$ represents a popular *Ball-region* set used for metric space partitioning and organization of database $\mathbb{S} \subset \mathbb{U}$. During the last three decades, a lot of indexing techniques have been designed for metric spaces, so-called metric access methods [42, 193, 151]. The techniques differ in the way they partition database objects, store precomputed distances and process similarity queries

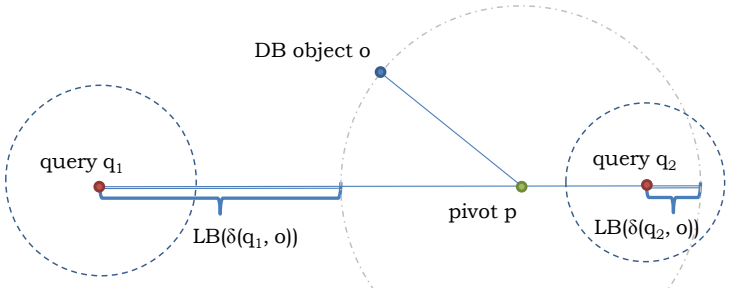



Figure 3.4: Pivot-based lower-bounding using triangle inequality.

[120, 47, 166, 181, 41, 164, 124, 4]. Furthermore, there still appear new approaches that can be used to enhance many of the well-established metric access methods.

For example, we have presented *Cut-regions* [113, 109] that represent compact metric regions suitable for more efficient indexing and retrieval (see Chapter 5). The idea of Cut-regions was initially introduced in the PM-Tree [166] for more efficient query processing in a hierarchy of Ball-regions constructed as the M-Tree [47]. In our work, we have separated the Cut-regions from the PM-Tree and defined an additional general set of operations suitable for index construction. We have also demonstrated that Cut-regions can improve the performance of other metric access methods relying on Ball-regions.

However, not only the metric axioms but also the distribution of the distances between database objects plays a significant role in the efficiency of the metric space indexing. In [42], the authors have proposed the intrinsic dimensionality measure that indicates whether the data can be efficiently indexed using a given distance space. The lower values of intrinsic dimensionality indicate that the data form clusters in the distance space and thus metric indexes based on metric space partitioning can be utilized [130, 131, 113]. On the other hand, high values of intrinsic dimensionality indicate no clusters, which means only the sequential processing using just simple query-to-object lower-bounding can be utilized to filter at least some costly distance computations [120]. The problem of high intrinsic dimensionality can be also addressed by approximate search strategies that can provide interesting precision-speedup trade-offs [136, 162, 2, 132]. For example, the Signature

Quadratic Form Distance trained for maximal effectiveness often suffers from high intrinsic dimensionality. Nevertheless, we have experimentally demonstrated [104] that using approximate k-NN search strategies designed for M-Index [130, 131], the retrieval effectiveness can be still competitive even if just a fraction of the database is visited.

Although metric indexes can significantly speed up query processing, the methods still assume that the number of query objects is high enough to compensate the indexing costs. Furthermore, the techniques assume that indexed data are not changed too often, so once data are indexed, they are often queried. These assumptions are not always satisfied, considering for example multimedia streams, where just a few queries can be issued for data stored in a search window. In such cases, the cost of updating the index could overcome the benefits of the indexing, while a sequential scan using a multi-query processing strategy [37] could be more efficient. If the queries are issued independently and no delays for query collection are allowed, the D-Cache structure [165] can be used to efficiently process a few number of independently issued similarity queries (see Chapter 6).

The structure of the D-Cache is simple – it is just a simple block of memory where distances evaluated during previous queries are hashed and stored. As in other cache types (disk, processor), the space for cached distances is limited and thus the new distances can replace the original ones. In order to filter non-relevant objects, each actually processed query object $q_i$ considers several previously issued query objects $q_j, j < i$ as pivots, and thus using $\delta(q_i, q_j)$ and $\delta(o_k, q_j)$ potentially stored in D-Cache, the lower-bound $LB_{\triangle}(\delta(o_k, q_i)) = |\delta(q_i, q_j) - \delta(o_k, q_j)| \leq \delta(o_k, q_i)$ can be evaluated and used for filtering. If $\delta(o_k, q_j)$ was not stored in the D-Cache or was already replaced, distance $\delta(o_k, q_i)$ has to be evaluated. Although the lower-bounding is the same as the one used by metric access methods, the D-cache does not have to create an index structure in advance, thus it can be used instantly and starting from the second query object the metric filtering can be employed. The D-Cache can be employed also for dynamically changing similarity models, for example, if the alpha parameter of the Signature Quadratic Form Distance is changed to improve efficiency of the filtering. We have also demonstrated that standard metric access methods can be enhanced by D-Cache for more efficient indexing and retrieval [165].

### 3.3.3 Ptolemaic indexing

The metric space approach is not the only way to efficiently index adaptive distance measures. Recently, Hetland et al. [77] proved that based on specific assumptions [18] the Signature Quadratic Form Distance is a ptolemaic metric (see Chapter 7), which means it satisfies metric properties and also the ptolemaic inequality stating that for any quadrilateral, the pairwise products of opposing sides sum to more than the product of the diagonals. Formally, for any four points $x$, $y$, $u$, $v \in \mathbb{U}$, the following inequality holds:

$$\delta(x,v) \cdot \delta(y,u) \leq \delta(x,y) \cdot \delta(u,v) + \delta(x,u) \cdot \delta(y,v) \tag{3.1}$$

As for the triangle inequality, the ptolemaic inequality can be used for distance-based indexing to construct a pivot-based lower bound. Given a query $q$, object $o$, and pivots $p$ and $s$, the Ptolemaic bound can be evaluated as:

$$\delta_{\mathrm{C}}(q,o,p,s) = \frac{|\delta(q,p) \cdot \delta(o,s) - \delta(q,s) \cdot \delta(o,p)|}{\delta(p,s)}, \tag{3.2}$$

where for $\delta(p,s) = 0$, the bound is defined to be zero $\delta_{\mathrm{C}}(q,o,p,s) = 0$. For a set of pivots $\mathbb{P}$, an optimal Ptolemaic bound [76, 105] considers all pairs of distinct pivots from $\mathbb{P}$:

$$\delta(q,o) \geq \mathrm{LB}_{\mathrm{ptol}}(\delta(q,o)) = \max_{p,s \in \mathbb{P}} \delta_{\mathrm{C}}(q,o,p,s) \tag{3.3}$$

However, the optimal Ptolemaic bound has the quadratic time complexity (based on $|\mathbb{P}|$) and so its estimation can be too costly with respect to a lower-bounded cheap distance. Therefore, only a specific set of pairs can be considered by a pair selection heuristic. Given a ptolemaic metric, another question is whether to use lower-bounding based on triangle inequality or ptolemaic inequality. In other words, whether the ptolemaic lower bounding can improve the triangle lower bounding, and vice versa.

In Figure 3.5, we have visualized points in 2D euclidean space, that can be filtered just by $LB_{\triangle}$ (blue points), $LB_{\mathrm{ptol}}$ (green points), by both lower bounding techniques (gray points) and points that cannot be filtered by any of the two techniques (white points). We may observe that both filtering techniques can contribute to the filtering, where the filtering power of each technique depends on the query radius and also on the constellation of pivots and the query object. The filtering power of cheap triangle lower bounding

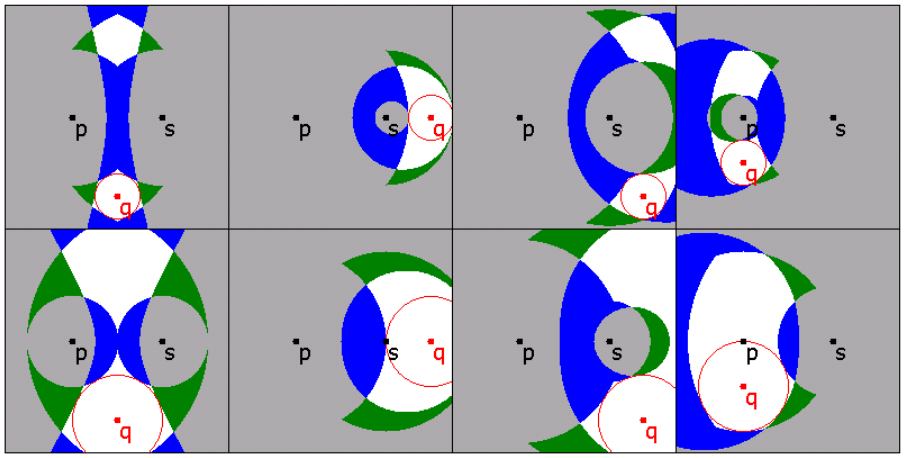



Figure 3.5: Triangle versus ptolemaic filtering in 2D euclidean space using two pivots $p, s$ and range query $(q, r_q)$, where the blue points can be filtered just using $LB_{\triangle}$, green points can be filtered just using $LB_{\text{ptol}}$, gray points can be filtered by both techniques, while white points cannot be filtered by neither of the two techniques.

could be increased by higher number of pivots. On the other hand, the ptolemaic lower bounding can improve the filtering power of a given pivot set. Both techniques can be also combined, where first the cheap triangle lower bound is evaluated and used for filtering. If the triangle lower bound is not sufficient, more expensive ptolemaic lower bound is employed for a given object. As for triangle lower-bounding, the ptolemaic lower bound can be evaluated also for whole regions and thus, given a ptolemaic metric, the filtering rules of many metric access methods can be simply extended [77].

### 3.3.4 Parallel computing

Although domain specific approaches, feature signature reduction techniques, and distance based indexing methods can significantly improve the efficiency of multimedia retrieval using feature signatures and adaptive distance measures, the techniques alone cannot make the model applicable for immense databases comprising billions of multimedia objects. In such cases, approaches like distributed computing and/or massively parallel computing have to be employed as well [15, 62, 118, 194]. For distributed computing, there have been already developed several approaches that can be directly

applied for models based on feature signatures and metric adaptive distance measures. For example, in [131] the authors propose a distributed metric index (M-Index) that can organize the database into a large number of nodes according to a metric distance. The index is suitable both for exact and approximate search, where especially the approximate search strategies can prune a significant part of the searched database. Furthermore, each node can utilize a centralized index structure and/or massively parallel computing to improve the efficiency of the distributed index. Inspired by the M-index, we have proposed an distributed approximate similarity join approach for Hadoop that enables efficient evaluation of k-NN graphs [186].

The parallel computing (e.g., GPU architectures [134, 133]) represents another promising approach for evaluation of costly adaptive distance measures that constitute a serious bottleneck of a multimedia retrieval system based on feature signatures. In recent days, novel many-core devices with specific hardware architectures are designed for various computing tasks. Hence, one of the goals of the research in this area is to find suitable computation tasks for existing many-core devices and to adapt the existing algorithms to better utilize properties of the devices. The typical example are GPU cards that provide thousands of cores. However, their hardware architecture and programming model significantly differ from traditional CPUs. Especially different memory organization and thread execution in GPU cards require different algorithms. The CPU and GPU approaches can be also efficiently combined in hybrid systems that better utilize available hardware.

Considering GPU architectures in the time of conducting our research [94], we have focused on parallel processing of adaptive distances and compared the efficiency of the retrieval when using two parallel environments with different architectures and also prices. More precisely, we have compared a cheap desktop PC comprising two GPU cards with CUDA architecture and an expensive high-end NUMA server. As parallel computing tasks, we have investigated the efficiency of the retrieval when using parallel computing for batches of distance computations, or even parallel processing using a simple metric index structure [94]. More specifically, we have designed two algorithms considering utilization balance between CPU and many-core GPUs for efficient similarity search with the Signature Quadratic Form Distance. We have shown how to process multiple distance computations and other parts of the search procedure (e.g., lower-bound estimation) in parallel, achieving maximal performance of the combined CPU/GPU system. We have experimentally demonstrated that using GPU cards for models based

on feature signatures represents an order of magnitude faster and cheaper solution than a high-end many core NUMA server, despite the memory organization and thread execution specifics of the GPU architectures. Similar results have been achieved also for position-color-texture feature signature extraction process, where we have reached the throughput of approximately 8000 extracted feature signatures per second [93] (given images pre-cached in RAM).

# Chapter 4

# Approximating the Signature Quadratic Form Distance Using Scalable Feature Signatures

Jakub Lokoč

Published in the proceedings of the 20th International Conference on Multi-Media Modeling MMM 2014, LNCS, ISSN 0302-9743.
dx.doi.org/10.1007/978-3-319-04114-8_8

The extended version of this paper [99] was published in Multimedia Tools and Applications journal (IF in 2014: 1.346).

# Chapter 5

# On Indexing Metric Spaces Using Cut-regions

Jakub Lokoč
Juraj Moško
Přemysl Čech
Tomáš Skopal

Published in the *Information Systems* journal, volume 43, pages 1–19. Elsevier, July 2014. ISSN 0306-4379.
dx.doi.org/10.1016/j.is.2014.01.007

Impact Factor in 2014: 1.456
5-Year Impact Factor in 2014: 1.618

# Chapter 6

# D-Cache: Universal Distance Cache for Metric Access Methods

Tomáš Skopal
Jakub Lokoč
Benjamin Bustos

Published in the *IEEE Transactions on Knowledge and Data Engineering* journal, volume 24, Number 5, pages 868-881. IEEE, May 2012. ISSN 1041-4347.
dx.doi.org/10.1109/TKDE.2011.19

Impact Factor in 2012: 1.892

# Chapter 7

# Ptolemaic Access Methods: Challenging the Reign of the Metric Space Model

Magnus Lie Hetland
Tomáš Skopal
Jakub Lokoč
Christian Beecks

Published in the *Information Systems* journal, volume 38, Issue 7, pages 989–1006. Elsevier, October 2013. ISSN 0306-4379.
dx.doi.org/10.1016/j.is.2012.05.011

Impact Factor in 2013: 1.235
5-Year Impact Factor in 2013: 1.435

# Part II

# Interactive video retrieval

# Chapter 8

# Commentary for Part II

## 8.1 Motivation

Every day, users search for videos to get informed, educated, entertained, or to inspect specific video collections (e.g., a personal archive or medical repository). The particular information need is usually in the mind of the user and might be a subject of evolution/change during the search process. Given an information need, users enter queries to modern video retrieval systems by means of a user interface (e.g., a textbox input) and browse a ranked result set. The effectiveness of the retrieval depends on many factors. On the systems site, effective video retrieval approaches have to be incorporated, including video analysis, automatic annotation, feature extraction and relevance scoring models. Another factor is that users have to be able to formulate an expressive query describing the target scenes or to use the system interface/models effectively to solve a given task. For example, if users want to find a video with a funny cat on youtube using keyword search, the first page of the result set would be probably sufficient for most users. However, if the user would like to search for one particular video with the funny cat (e.g., observed long time ago), the searched video would be probably hard to find due to too many candidates matching a simple keyword query or potential problems with automatic annotation for more detailed query specification. Generally, with the rise of deep learning models it becomes easier to find instances of a more general class of scenes on the first page (i.e., precision at K). However, finding all of the instances of the class or finding one particular scene is still a challenge in many types of video retrieval tasks in general

video domains. For narrow domains with a lot of available training data, the options to design effective retrieval models are better, but the querying might require domain knowledge.

Two types of video retrieval tasks are considered throughout this part of the thesis – known-item and ad-hoc search. Both types of tasks assume that the task description does not change. Known-item search tasks represent situations, where users search for one particular scene in a given video collection. Either the users saw the scene before or they have a specific textual description of the scene. Ad-hoc search tasks represent situations, where users search a collection for all scenes corresponding to a given short textual description. Even though the description of searched scenes does not change for a given ad-hoc search task, the opinion of which scene is relevant may change based on voting (or calibration) among live judges. For example, there was a task to find people queuing at the Video Browser Showdown 2019. After a first few submissions, the judges had to decided whether one person in a queue can be already considered as people queuing. As presented in TRECVID [9] or the Video Browser Showdown [102] reports, both types of tasks still represent a challenge for current retrieval models. Especially in cases, when an ideal query object is not available.

The difficulty of the tasks can be illustrated by the following image retrieval simulation using 499 randomly selected ImageNet classes. Each class consists of 100 randomly selected images represented as neuron activations from the last pooling layer of the NasNet Mobile network [198]. Hence, the dataset $\mathbb{S}$ has 49900 images represented as vectors $v \in \mathbb{R}^n$. Let $p_i = \{k_i, q_i\}, k_i, q_i \in \mathbb{R}^n$ be a pair of a sampled known-item $k_i$ and a query $q_i$ of the same class as $k_i$. The pair simulates that a user tries to employ $q_i$ to find $k_i$. For each sampled pair $p_i$ and considering the cosine similarity, the query $q_i$ induces ranking on $\mathbb{S} - \{q_i\}$ which determines the rank of the searched item $k_i$ and also the rank of the first/last item of the query object class. Considering 20 sampled pairs for each class, Figure 8.1 shows the empirical cumulative percentage of found items up to a given rank.

Whereas the first found image of the query class has a low rank for most queries, searching for one particular item or all items of the query class is more difficult even for images. For example, to find 80% of simulated known-items with a given query, users would need to (effectively) browse results up to rank about 6500. Hence, interactive search is still considered to boost the effectiveness of video retrieval systems. Instead of sequential browsing of one large result set, the search process comprise iterative query reformulation,

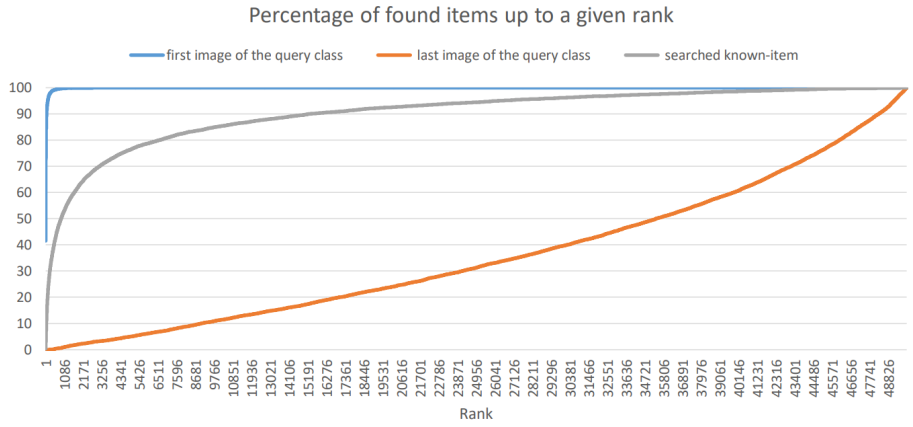



Figure 8.1: Simulations of three types of retrieval tasks.

results visualization and exploration/browsing phases. Given more control over the search process, users can provide valuable high-level decisions based on inspection of the results of automatized subtasks. The corresponding more complex interfaces should consider special design efforts to minimize a usability gap (e.g., the eight golden rules [159]).

Figure 8.2 presents an example of known-item search interactions in our interactive video retrieval tool VIRET [107] during a visual known-item search task. The x-axis represents the actual task time (first 4 minutes), while y-axis shows the current position of the top ranked frame from the searched video (gray color) and shot (red color). As there were applied presentation filters for the maximal number of displayed top-ranked frames from a video/shot, the orange line shows the position of the top ranked frame from the searched shot without the filters. Below the x-axis, currently used queries are presented. The red color represents keyword-search, the blue color represents querying by an example image (E denotes that the image was taken from an external image search engine). The model used for sorting is highlighted on the x-axis. The heat map below queries shows a browsing activity (e.g., video or temporal context inspection). We may observe that the user employed multiple modalities and query images. Once a frame from the searched scene appeared on the first page, the user started also with browsing and then selected the correct frame/shot.

In the following, popular approaches incorporated by state-of-the-art interactive video retrieval systems are presented in connection with performance evaluation challenges.

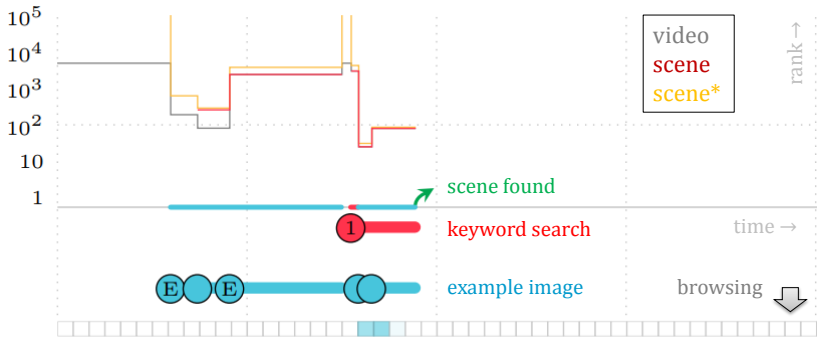



Figure 8.2: User interactions and the actual position of the top ranked frame from the searched video/scene in a visual known-item search task. The orange color is used to show the position without presentation filters.

## 8.2 Methods for interactive video retrieval

Effective and user friendly approaches to interactively solve difficult video retrieval tasks are still the subject of intensive research. Various interactive video retrieval frameworks (e.g., vitrivr [145], VIREO [128], VISIONE [1], diveXplore [157], VERGE [3], VIBRO [14], or our system VIRET [107, 108]) are designed and confronted at international evaluation campaigns. Please note that interactive video retrieval systems participating at the Video Browser Showdown competition are summarized in Chapters 9 and 10. Based on our experience with task-oriented evaluations at the Video Browser Showdown [102], this section briefly summarizes a list of general approaches identified for a successful known-item and ad-hoc search interactive video retrieval system. We also highlight methods used by our VIRET system (summarized in Chapter 11) and several additional contributions. For a broader overview of various interactive retrieval ideas, paradigms and trends, we refer the readers to the following surveys [176, 156].

### 8.2.1 Video analysis and preprocessing

Video is represented as a temporally ordered sequence of frames. In addition, videos from professional productions are often organized to a hierarchy of scenes and shots, where the contents of the frames in one temporal unit is usually highly similar both visually and semantically. Since information

about scenes and shots is not available in raw video data, an automatic detection of the temporal units represents a traditional step of video analysis. The knowledge of the video structure is beneficial for video presentation, browsing, filtering and retrieval. In an comprehensive survey on content-based video retrieval, Hu et al. [80] considers three problems of video structure analysis – shot boundary detection, key frame extraction and scene segmentation. The methods for shot boundary detection try to identify various types of transitions, relying either on threshold based approaches or statistical learning models. Key frame extraction methods try to reduce redundant consecutive frames and select the set of salient representative frames for each shot. The quality of selected frames has to be assessed by users, which makes the key frame selection a difficult problem. Popular approaches employ heuristics based on comparison between frames, clustering or even simple uniform segmentation. Please note that representative frame selection is an important step as the result sets are mostly presented in the form of a collection of images/frames. Scene segmentation approaches [80] rely either on the analysis of frames, audio-visual fusion or similar background assumption.

In order to simplify the development of interactive video retrieval systems for the Video Browser Showdown participants, the currently employed V3C dataset [146] (shared with TRECVID) is already provided with a precomputed temporal segmentation of videos. For the considered video fragments (e.g., segments or selected frames), the teams can focus on descriptor extraction functions to prepare suitable representations for video retrieval and browsing models. Currently, our system employs an own temporal segmentation approach focusing on shot transition detection using an own 3D deep convolutional neural network. Since the retrieval models operate on the set of selected frames, a frame selection method is used considering clustering of consecutive uniformly sampled frames in detected shots. The shots can be used also for approximate presentation filtering of results (e.g., show just the top ranked frame from a shot), assuming that selected frames from a shot are sufficiently visually similar and users can recognize the searched shot.

### 8.2.2 Search initialization with a query

Human computing in connection with a well-designed sequential browsing interface [81] proved to be surprisingly effective approach up to one hundred hours of video in textual known-item search tasks [48]. However, systems

employing only sequential browsing are generally exhaustive for users and not scalable. For large datasets, an initialization of the search with a query is an important step to get a smaller candidate set of top ranked results for further inspection and refinement. Users can inspect the results to find the searched scene on first few pages, find a suitable query object, or exploit a returned frame as a link to a region in a precomputed exploration/navigation structure (e.g., ImageMap [12] or another suitable browsing model [75]). The following query initialization approaches are frequently supported.

Keyword search represents a popular and intuitive form of search initialization for users. Given automatically detected annotations for videos, (interactive) video retrieval systems can integrate various keyword search approaches [150, 11]. The participating video retrieval systems at the Video Browser Showdown often rely on automatic annotation of frames or scenes using deep learning approaches that can extract concepts, captions, texts and speech with the state-of-the-art precision. For example, the recent version of the *vitrivr* system [145] relies on various deep learning models for semantic segmentation [43], extraction of spatio-temporal features [179] for action classification, or captioning of selected frames [188]. The tool successfully incorporated also annotations from OCR/ASR models. A popular option used by many VBS teams are also (retrained) deep networks for image classification. In the current version of our VIRET system, we consider a retrained NasNet large network [198] with an own set of 1243 classes/labels. In the preprocessing phase, for each selected frame the output (softmax) vector $x \in \mathbb{R}^{1243}_{\geq 0}$ is extracted by the network. Users can query the selected frames only with the supported 1243 labels and additional 765 hypernyms prompted by the input control. Since users might face problems with expressing their needs (e.g., small vocabulary), advanced prompting is integrated to our system linking the entered words also with keyword descriptions. Generally, users do not have to rely just on the supported set of labels. For example, the *VERGE* system [3] employs a module that translates the original textual query to a set of supported visual concepts, considering in some cases also query expansion with semantically similar concepts.

Sketch based retrieval is useful in scenarios, where users remember visual appearance of the searched scene. Sketching is supported by most of the mentioned interactive video retrieval systems, as it enables to express memorized colors, edges or motion. However, for humans the ability to reproduce previously observed features is limited (e.g., due to a limited visual working memory capacity [115]) and so the retrieval models have to consider

incomplete and potentially noisy sketches. For example, it is necessary to assume that a color sketch only approximates the real colors in the searched shot as users face problems with memorizing a rich palette of colors and the reproduced colors tend to be biased [10]. Furthermore, if the retrieval system operates only on selected frames, it does not have to be guaranteed that the frame containing the memorized color distribution is selected from the shot. Hence, the sketch drawing component can support interactive editing operations for frequent sketch modifications. For example, our system uses editable color query regions with ALL/ANY specification [112]. Another strategy for interactive color sketching was presented by the *VIREO* system [127] that uses grid-based color sketching with informative visualizations. Given a selected color and a set of already placed colors, remaining empty places in the grid show a density estimate of the corresponding color combination in the dataset. With the improvements in object detection and semantic segmentation, systems start to incorporate also semantic sketches. The semantic concepts are entered either as positioned boxes relying on an object detector (e.g., as used by the *VISIONE* system [1]) or as free form shapes relying on semantic segmentation (e.g., as supported by the *vitrivr* system [145]).

An ideal example image represents a strong clue to find a searched scene. However, finding such example might be a challenge on its own. Users can try an external image search engine with effective keyword search, or the image can be discovered and selected from the current results set. Selection of the query example has to be convenient (e.g., drag and drop from the external engine, as implemented by our system [107]) and the history of query examples should be accessible. In order to model the contents of a selected frame, representations based on neuron activations of deep convolutional neural networks [174, 74, 65, 198] are often employed. For example, our system models the similarity of two images using the cosine similarity of the corresponding neuron activation vectors obtained from the retrained NasNet mobile network [198].

Videos provide different useful modalities for retrieval purposes. In order to improve retrieval effectiveness, the modalities are used in fusion schemes [168]. Various fusion strategies were investigated by interactive video retrieval systems. For example, the *VERGE* system [123] employed a hybrid approach combining non-linear and graph-based late fusion [61], given top K relevant shots returned by a dominant feature. In the most recent version of our system [107, 108], we considered a temporal fusion for a sequence of

queries of one modality, each targeting a different frame in the searched sequence. For a multi-modal query, a simple late fusion strategy is utilized. The intersection of top ranked results of each employed modality is evaluated and the overall result is sorted by a selected modality. The number of considered top ranked results for a query modality can be interactively controlled by users (except semantic sketches serving as filters). Implicitly, the thresholds are set to larger values, which decreases the chance to filter out the searched frame by the fusion for the price of larger candidate result sets.

### 8.2.3 Visualization and browsing

For human visual system, data visualization represents a powerful tool to present and communicate information. One of the tasks where human vision abilities might help is inspection of candidate sets. Therefore, data visualization approaches complement querying to speed up the time to solve a video retrieval task. Video interaction systems [156] consider visualization of result sets, interactive navigation in advanced exploration structures [75] and video browsing approaches for fast video inspection (augmented navigation bars, video summaries and ergonomic controls) on various types of devices.

The very basic visualization component adopted by many systems is an image grid presenting representative video frames (e.g., top ranked items). The grid has no overlaps and fills the whole designated area. Since human perception is limited, about 20-50 images per page [12] are usually presented on a display. Since the searched item might reside much deeper in the current ranked candidate set, additional visualization/browsing strategies have to be considered. For example, a popular use case implemented by popular web search engines and also by many interactive video retrieval tools is that users scroll down a list of results revealing additional items. Another strategy is to show many small thumbnails at once and use a method organizing similar images close to each other on a display (e.g., Self-Organizing Maps [88], t-SNE [185] or Self-Sorting Maps [169]). These approaches enable users to quickly overview a larger portion of the result set, look for the searched item or identify promising query examples. The sorted maps can be also easily organized into hierarchies. As demonstrated by the Vibro system, hierarchical image maps [13] and graph-based browsing [14] in a grid can be effectively employed in known-item search scenarios. Typical browsing scenarios in a static image map are inspired by classical map services. Users can navigate

in a hierarchy (zoom in/out) and inspect image regions on the same level with panning. In order to preserve image relations and support dynamically changing datasets, a hierarchical similarity graph approach was proposed with a method projecting requested subgraphs to a 2D grid [14]. Given the projection mechanism, the same navigation interface can be designed as for the static image map variant. Browsing in sorted image maps was successfully implemented also by the diveXplore system [157], integrating advanced cooperation functionality for multiple users operating different instances of the tool. For a broader survey of content-based browsing approaches in image collections, we refer to Heesch [75].

Once users find a good query example (e.g., a searched TV studio), the temporal context becomes important for fast visual inspection of promising candidates. Hence, instead of a single top matching frame, stripes depicting also a temporal context are beneficial for the users. Switching from the grid of images to a list of stripes corresponds to a shift from the exploration to exploitation phase of the search. The temporal context can be also presented interactively for a selected displayed frame, "playing" a sequence of preceding or following frames on mouse wheel over the frame. Examples of interfaces that were used in our interactive video retrieval tool at the Lifelog Search Challenge are presented in [112]. Please note that a simple grid was often employed for the "exploration" phase, where users entered a query and a larger number of frames was presented for inspection. However, showing more top ranked items at once leads also to situations where users overlook a searched frame (see Chapter 11). Specific visualization techniques can be designed also for continuously evolving videos. For example, we have investigated a hierarchical visualization method for selected frames from an endoscopic video [111] to aid with after-inspection of endoscopic surgeries.

Alternatively to 2D visualization methods, various 3D interfaces have been considered in multimedia retrieval [153, 154]. Please note that approaches for 3D layouts can rely on additional visualization and browsing options (e.g., perspective or camera movements). In our work [117], we have examined efficient 3D visualization approaches based on the particle physics model and a precomputed graph with edges corresponding to similarities between database objects. We have demonstrated that the model can be configured to efficiently produce various types of layouts.

### 8.2.4 Relevance feedback

In order to narrow the gap between search needs and effectiveness of retrieval models for a given task, a relevance feedback approach can be incorporated into the search loop. The feedback can be provided explicitly (binary or multivalued relevance), where users pick positive and/or negative examples that are used to adapt the retrieval system for the next iteration [143, 192]. Implicit feedback is more convenient for users, as the system tries to collect and analyze user behavior automatically. Let us note that interactive video retrieval interfaces provide a rich set of elements that can be tracked for implicit feedback. A relevance feedback has been considered in a Bayesian framework for image class search based on a mental picture [57]. In the framework, the user observes in each iteration an appropriately selected display of images and picks the most relevant image with respect to the searched concept. An update model is then used to estimate posterior probabilities based on the feedback. The model has been extended to support multiple selected images [17], scalable retrieval with HEAT approach [170] and to consider exploration/exploitation phases [171]. Inspired by the Bayesian framework, an interactive video retrieval system for mental visual browsing [73] was presented at the Video Browser Showdown 2016.

## 8.3 Interactive video retrieval evaluation

Assessing the performance of an interactive video retrieval system is a difficult task as both the effectiveness of retrieval models and usability of interfaces have to be tested. Since every user is different, evaluations of a large number of information retrieval tasks with many users is necessary which makes the testing unwieldy and time demanding [189]. The evaluations for particular retrieval models are often simplified and automatized with respect to a benchmark collection with predefined tasks and expected results. This ensures experiment repeatability and enables automatic evaluations of precision and recall based measures. The automation provides clues to select suitable models for a given type of data and optimize involved model parameters. For example, in our recent work [100] we have investigated options for an automatic performance analysis of a simple color sketching retrieval model (used in precedent versions of our video retrieval tool [103, 35]). For randomly selected "known" images $o \in \mathcal{S}$ rep-

resented as position-color feature signatures $FS^o = \{\langle r_i^o, w_i^o \rangle\}_{i=1}^n$, known-item search simulations were considered using automatically generated color sketch queries $FS^q = \{\langle \epsilon(r_i^o), w_i^o \rangle : \langle r_i^o, w_i^o \rangle \in \Pi(FS^o)\}$. Several functions $\Pi$ projecting the signature to a subset of tuples $\Pi(FS^o) \subset FS^o$ were investigated, while $\epsilon : \mathbb{F} \to \mathbb{F}$ represents a simplified user error model considering Gaussian white noise. The simulation framework can be used as a starting analytical tool for investigation of potentially effective search strategies and preliminary insights of the performance of employed ranking models. However, even more complex interactive search simulations [195] and evaluations based on benchmark collections do not cover all possible open world search scenarios and so provide only a tentative estimate of the general performance. Hence evaluations with real users should be performed as well.

Ideally, the evaluations with real users should incorporate both qualitative and quantitative usability studies. Qualitative usability studies investigate whether users face problems with certain system/interface features and components. The users are asked to use the analyzed system for a short period of time, while a usability evaluation method is performed. For example, in the think-aloud protocol [95] users are encouraged to communicate aloud their thoughts during a set of tasks with a tested system. Besides a valuable feedback, the protocol provides also an insight into cognitive processes of each user when learning to use the system. Another popular evaluation method is based on appropriate post-task and post-test usability questionnaires. Post-task questionnaires focus on user experience with particular tasks (e.g., was the tested system helpful to solve the task?), while post-test questionnaires collect information about the overall impression from the tested system. Quantitative analysis can be connected with task-based testing to measures the effectiveness of solving a certain set of tasks (e.g., by means of the success rate and time to solve a task). The analysis can be further supported with interaction logging and eye/mouse tracking, which provides an additional valuable data source for interface and usability analysis. Both qualitative and quantitative evaluations have to be performed with many users and trials, in order to support the overall results with significance tests. Whereas all the presented evaluations are manageable for a single system or its components, comparative evaluations of many different interactive video retrieval systems pose a difficult challenge for a single research team. In order to provide a task-based comparative evaluation platform and foster research in the (interactive) video retrieval area, evaluation campaigns are organized for different types of tasks and data sources. The organizers of

the campaigns define video retrieval tasks, settings, datasets, ground truth, evaluation metrics and scoring.

Video retrieval evaluation campaigns started to appear since 2000 as a by-product of traditional text-based retrieval campaigns. One of the most famous campaigns is TRECVID [9], which was organized since 2001 as a Text Retrieval Conference (TREC) track and since 2003 it has become an independent event. TRECVID focuses on realistic tasks and automatic content-based video analysis for retrieval and event detection. The primary focus is fully automatic search based on a provided query, however, some settings enable user interactions (e.g., manually change the query, or reformulate query based on top few results). The systems evaluate the query and return top 1000 items to TRECVID organizers, who focus both on precision and recall levels. In order to foster interactive search, additional campaigns have been started. VideOlympics [167] were organized at CIVR in years 2007-2009, where research teams could present their systems and solve tasks directly in front of the audience. Inspired by this exciting event, the Video Browser Showdown campaign started in 2012. The Video Browser Showdown [102] focuses on task-based comparative evaluation of interactive video retrieval systems, given the same set of known-item and ad-hoc search tasks, settings and conditions. All the participating teams try to solve the tasks in a concurrent way, given the same time limit and in front of the audience. Recently, a similar Lifelog Search Challenge campaign was established to foster interactive search in related lifelog data [68, 67].

## 8.4 Video Browser Showdown participation

The author of this thesis is involved in the Video Browser Showdown (VBS) evaluation campaign both as a co-organizer and regular participant with a video retrieval tool developed by his team. Already the first participation was successful in 2014, where the presented tool called the Signature Based Video Browser [103, 35] achieved the first position. The tool focused on visual known-item search using simple interactive color sketches and underlaying ranking models, which was a successful strategy also the next year at the Video Browser Showdown in 2015 (with an enhanced version [34]). With the growing video collection and new textual tasks, the new version of the tool [32, 33] incorporated also edge based sketches, query by an example image and keyword search using ILSVRC classes extended by hypernyms.

After being twice at the third place (at the Video Browser Showdown in 2016 and 2017), a significantly revisited version of the tool called VIRET [112] achieved again the first place at the Video Browser Showdown 2018. The VIRET tool used its own set of supported concepts, own color sketching model, deep features from GoogleNet, and supported multi-modal search. Please note that the VIRET tool version employed at VBS 2018 was used also at the Lifelog Search Challenge 2018 [67], where the tool achieved the third place even without considering Lifelog specific modalities (e.g., location, date, heart rate). At the Video Browser Showdown 2019, the VIRET tool considered a retrained NasNet deep neural network [198], a new neural network architecture for shot boundary detection *TransNet*, temporal queries and updated result presentation panels [107]. A more detailed description of the most recent version of our system is presented in Chapter 11. Our system achieved the overall second position after the winning vitrivr system [145]. It was an interesting comparison of two frameworks, both able to solve 18 out of 23 known-item search tasks in the currently used V3C dataset with 1000 hours of video [146]. Whereas our system relied only on visual information and general concepts, the vitrivr system demonstrated that automatic speech recognition and optical character recognition are very effective for the current visual known-item search settings at VBS.

Regarding the Video Browser Showdown organization, we have contributed to new task presentation settings, scoring formulas, interaction logging and led two extensive Video Browser Showdown evaluation reports included in this thesis. Similar as for TRECVID, the Video Browser Showdown focuses on simulations of realistic tasks which bring several challenges. For example, is it better to simulate visual KIS tasks by presenting the "known scene" only once or playing it in the loop? The first case is more realistic, however, users may forget the "implanted" searched scene. This topic is discussed in more detail in Chapter 9. The scoring represents another challenge as the teams try to maximize their overall score to win the competition. We have revisited scoring formulas for all types of tasks, incorporating time and the number of correct/incorrect submissions.

In order to provide more advanced comparative reports and insights for used tool features, we defined a simple interaction logging format for currently used interactive video retrieval tools and integrated collected interaction logs to the submission process. The proposed methodology has led to the first successful attempt to collect interaction logs from eight of nine teams at the Video Browser Showdown 2018. Although the teams did not

implement the logging in the full extent, the logs (at least partially) connected successful submissions with used tool features. This information was available for the majority of the participating systems for the first time in the Video Browser Showdown history. The logs also showed expected elementary search patterns and raised logging challenges for the next installments of the Video Browser Showdown. A survey paper revisiting and summarizing the Video Browser Showdown in years 2015-2017 is included as Chapter 9, while a detailed analysis of the Video Browser Showdown 2018 is presented in Chapter 10.

# Chapter 9

# On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015-2017

Jakub Lokoč
Werner Bailer
Klaus Schoeffmann
Bernd Muenzer
George Awad

Published in the *IEEE Transactions on Multimedia* journal, volume 20, Number 12, pages 3361-3376. IEEE, April 2018. ISSN 1520-9210.
doi.org/10.1109/TMM.2018.2830110

2017 Journal Impact Factor: 3.977

# Chapter 10

# Interactive Search or Sequential Browsing? A Detailed Analysis of the Video Browser Showdown 2018

Jakub Lokoč
Gregor Kovalčík
Bernd Muenzer
Klaus Schoeffmann
Werner Bailer
Ralph Gasser
Stefanos Vrochidis
Phuong Anh Nguyen
Sitapa Rujikietgumjorn
Kai Uwe Barthel

Published in *ACM Transactions on Multimedia Computing, Communications, and Applications* journal, volume 15, Number 1, pages 29:1-29:26. ACM, February 2019. ISSN 1551-6857. doi.org/10.1145/3295663
2017 Journal Impact Factor: 2.019

# Chapter 11

# VIRET: A video retrieval tool for interactive known-item search

Jakub Lokoč
Gregor Kovalčík
Tomáš Souček
Jaroslav Moravec
Přemysl Čech

Accepted to International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.
doi.org/10.1145/3323873.3325034

# Chapter 12

# Conclusions and discussion

During years 2011-2016, we have focused on efficient content-based similarity search using models based on feature signatures. The feature signatures enable unsupervised flexible representation of multimedia objects. However, the efficiency of the retrieval using feature signatures represents a challenge for large multimedia collections. Therefore, we have investigated parallel feature extraction techniques to speed up multimedia indexing phase and several approaches for efficient retrieval. More specifically, we have demonstrated that metric/ptolemaic access methods can be employed for efficient retrieval using metric/ptolemaic adaptive distance measures, feature signatures can be scaled for effective approximate search, and also that the distances represent a suitable task for parallel processing. Furthermore, many of the presented techniques have been designed as general approaches not restricted just to feature signatures and adaptive distance measures. Generally, any similarity model with expensive distance measure satisfying metric/ptolemaic properties could benefit from many of our new techniques.

In years 2014-2019, we have focused also on interactive video retrieval. In the early installments of the Video Browser Showdown, we have demonstrated that position-color feature signatures representing video frames can be considered for the design of interactive color sketch interfaces and effective ranking models for visual known-item search in small video datasets (winning the Video Browser Showdown in 2014 and 2015). The original Signature-based video browser was incrementally extended by keyword and semantic search models for increasing size of the utilized dataset. In 2018, the major revision of the tool (named VIRET) won the Video Browser Showdown again, providing access to 600 hours of video using multi-modal fusion

of selected models for keyword search, color sketching and query by example image. In 2019, the VIRET tool was extended by temporal queries, new temporal segmentation using a new deep neural network architecture and considered a retrained state-of-the-art NasNet architecture [198] for automatic annotation of selected frames. The VIRET team (including novices) solved 18 out of 23 known-item search tasks (in 1000 hours of video!) at the Video Browser Showdown 2019 and achieved the overall second place.

In 2016, the author of the thesis has joined the Video Browser Showdown organization committee and worked on the revision of task presentation settings, scoring functions, proposed interaction logging methodology, and led two extensive Video Browser Showdown summary papers. The first paper surveyed, revisited and summarized the Video Browser Showdown evaluation settings and results for years 2015-2017. The results of ad-hoc search tasks were also compared with TRECVID. The second paper proposed a detailed analysis of the Video Browser Showdown 2018, including a first successful attempt to collect interaction logs from participating teams. Based on our experience with video evaluation campaigns, we have also presented a half-day tutorial at ACM MM 2018 [155]. The tutorial motivated for interactive video retrieval in the age of deep learning, summarized observed successful approaches and tools, detailed problems related to the organization of selected video retrieval evaluation campaigns, and discussed the observed results from recently organized events.

At the end of the thesis, we provide a short discussion on the future of models based on features signatures and challenges of interactive video retrieval.

## 12.1 Where are you heading models based on feature signatures?

Object specific representations in connection with adaptive distance measures constitute a strong handcrafted formal framework [147, 18], yet simple enough to prove many properties analytically. In recent years, the framework has demonstrated its effectiveness in several benchmarks across various domains [110, 19, 20, 59, 182, 22, 160]. However, newly developed representation models are continuously pushing the effectiveness towards better and better results. It is highly remarkable, how developments in the machine

learning area consistently outperform many traditional approaches in various classification or retrieval tasks. Nevertheless, as long as machine learning approaches requiring just a few training examples (e.g., one shot learning [187]) do not reach superior performance, we believe that retrieval systems can still consider flexible frameworks for unsupervised similarity search in domains without large collections of annotated training data. Especially in connection with interactive retrieval scenarios, where users switch between retrieval models based on the currently observed model performance for a given query object and task.

Regarding the efficiency of the retrieval with models based on feature signatures, Beecks et al. [26] recently proposed so-called gradient-based signatures relying on a generative model with a finite set of parameters $\theta$ to aggregate the representatives of a feature signature. The authors proposed a likelihood function $\mathcal{L}(\theta|FS)$ of the generative model parameters $\theta$ with respect to a feature signature $FS$. Given the log-likelihood function $\log(\mathcal{L}(\theta|FS))$, the gradient-based signature corresponds to the change of the function parameters $\theta$ to better fit $FS$. Considering the same generative model for all feature signatures in the database, classical cheap bin-to-bin measures can be employed. Based on the employed set of image retrieval benchmark datasets, the authors presented that the effectiveness of the gradient-based signatures can outperform models based on feature signatures and adaptive distance measures. On the other hand, in the endoscopic domain Beecks et al. [20] presented that models based on feature signatures are more effective than gradient-based signatures for the task of linking images to video segments. Hence, it seems that the effectiveness of the models depends on the particular data distribution and task type. As the gradient-based model focuses on highly efficient retrieval, a comparative study involving proposed metric/ptolemaic indexing approaches could provide a deeper insight of effectiveness/efficiency trade-off for both models. Generally, a comprehensive comparative evaluation in various domains including also popular bag of features models represents an interesting task for future investigation. Especially for scenarios, where issued queries do not fit trained codebooks or employed generative models.

## 12.2 Interactive video retrieval challenges

Information retrieval tasks focusing on high recall in general video data represent still a difficult challenge for video retrieval systems. In order to find a searched scene (or set of scenes) based on memories, users need search clues that satisfy three properties – the clues have to be discriminative, reproducible and automatically detectable for the searched scene. For example, the vitrivr system [145] has demonstrated for visual known-item search tasks at the Video Browser Showdown 2019 that a suitable search clue can be speech or text present in the scene (considering the V3C collection [146]). For a heard sequence of several words or observed unique text label, the employed ASR/OCR models already provided sufficiently effective automatic annotation. Nevertheless, in web-scale databases these clues do not have to be unique. Furthermore, after some time users may remember only concepts or general visual clues, or the searched scene does not have to contain speech/label clues. In these cases, at least one of the three properties does not have to be satisfied and so the performance of the employed ranking models could be lower.

We believe that machine learning approaches and large-scale data annotation initiatives will continue to improve effectiveness of automatic video annotation models. With more precisely detected objects/concepts (even small items), relations between objects and temporal actions, keyword search effectiveness will be significantly enhanced. Given a more accurate region proposal for arbitrary objects, the systems could rely also on similarity search in promising image subregions. However, it is questionable whether it is feasible to provide sufficient training data for all possible "open world" search tasks. In addition, provided that users are able to remember only a limited number of objects or details in the scene, it will be probably always necessary to assume incomplete queries. Furthermore, the provided queries can be noisy as users may mistakenly remember some concepts/features (e.g., wrong color hue or its position). Hence, methods for effective and efficient inspection of larger candidate sets are necessary. For example, advanced result set exploration approaches or adaptive models incorporating explicit/implicit relevance feedback represent a promising direction to improve effectiveness of known-item and ad-hoc search processes. Another problem arises in textual tasks where users often face problems with good imagination of the scene. Clinging to a misleading idea of the scene may result in highly inefficient search. Therefore, visualizations providing diversification of potentially

searched candidates have to be employed to help to resolve such discrepancies sooner in the retrieval process.

Given a fixed dataset, approaches to automatically analyze and set up parameters of complex interactive retrieval systems could further aid with the development of the systems. We believe that analytical frameworks could provide promising configuration candidates for evaluations with real users. Automatic configuration approaches for interactive systems require artificial users for query specification and browsing. With suitable artificial users, simulation frameworks could be designed to obtain useful insights and estimated limits of considered models. Ideally, the simulation framework could propose a minimal set of recommended retrieval models for a given dataset and render a prototype application with corresponding standard user interfaces.

Regarding the Video Browser Showdown, it is necessary to keep focus on realistic simulations of tasks and revisit methodology for interactive ad-hoc search evaluations. Especially a suitable form of calibration of the search intents among judges and participants represents an open challenge. For known-item search evaluation, we plan to collect a limited prefix of each computed ranked result set from participating tools during search sessions. As the searched scene is known, the effectiveness of employed relevance score models could be compared. In future installments of the Video Browser Showdown, we also plan to start cooperations with experts from the human-computer interaction community and incorporate usability evaluations. For example, we would like to design and incorporate questionnaires for novice users to address usability issues of the compared interactive video retrieval tools.

# Bibliography


[1] G. Amato, P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, and C. Vairo. VISIONE at VBS2019. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, pages 591–596, 2019.

[2] G. Amato and P. Savino. Approximate similarity search in metric spaces using inverted files. In *Proceedings of the 3rd International Conference on Scalable Information Systems*, InfoScale '08, pages 28:1–28:10, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[3] S. Andreadis, A. Moumtzidou, D. Galanopoulos, F. Markatopoulou, K. Apostolidis, T. Mavropoulos, I. Gialampoukidis, S. Vrochidis, V. Mezaris, I. Kompatsiaris, and I. Patras. VERGE in VBS 2019. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, pages 602–608, 2019.

[4] L. Ares, N. Brisaboa, A. Ordóñez Pereira, and O. Pedreira. Efficient similarity search in metric spaces with cluster reduction. In G. Navarro and V. Pestov, editors, *Similarity Search and Applications*, volume 7404 of *Lecture Notes in Computer Science*, pages 70–84. Springer Berlin Heidelberg, 2012.

[5] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[6] I. Assent, A. Wenning, and T. Seidl. Approximation techniques for indexing the earth mover's distance in multimedia databases. In *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on*, pages 11–22. IEEE, 2006.

[7] I. Assent, M. Wichterich, T. Meisen, and T. Seidl. Efficient similarity search using the earth mover's distance for large multimedia databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 307–316. IEEE, 2008.

[8] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, Nov 2010.

[9] G. Awad, A. Butt, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet. Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[10] G.-Y. Bae, M. Olkkonen, S. R. Allred, and J. I. Flombaum. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of experimental psychology: General*, 144(4):744–763, 2015.

[11] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[12] K. U. Barthel and N. Hezel. Visually exploring millions of images using image maps and graphs. In B. Huet, S. Vrochidis, and E. Chang, editors, *Big Data Analytics for Large-scale Multimedia Search*, pages 251–275. John Wiley and Sons Inc., 2019.

[13] K. U. Barthel, N. Hezel, and R. Mackowiak. Imagemap - visually browsing millions of images. In X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. A. Hasan, editors, *MultiMedia Modeling*, pages 287–290, Cham, 2015. Springer International Publishing.

[14] K. U. Barthel, N. Hezel, and R. Mackowiak. Navigating a graph of scenes for exploring large video collections. In Q. Tian, N. Sebe, G.-J.

Qi, B. Huet, R. Hong, and X. Liu, editors, *MultiMedia Modeling*, pages 418–423, Cham, 2016. Springer International Publishing.

[15] M. Batko, D. Novak, F. Falchi, and P. Zezula. On scalability of the similarity search in the world of peers. In *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006*, New York, NY, USA, 2006. ACM.

[16] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[17] T. Bdiri, N. Bouguila, and D. Ziou. *A Statistical Framework for Mental Targets Search Using Mixture Models*, pages 99–118. Springer International Publishing, Cham, 2015.

[18] C. Beecks. Distance-based similarity models for content-based multimedia retrieval. In *Dissertation, Fakultat fur Mathematik, Informatik und Naturwissenschaften, RWTH Aachen University.*, 2013.

[19] C. Beecks, S. Kirchhoff, and T. Seidl. Signature matching distance for content-based image retrieval. In *Proc. ACM International Conference on Multimedia Retrieval (ICMR 2013), Dallas, Texas, USA*, pages 41–48, New York, NY, USA, 2013. ACM.

[20] C. Beecks, S. Kletz, and K. Schoeffmann. Large-scale endoscopic image and video linking with gradient-based signatures. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 17–21, April 2017.

[21] C. Beecks, J. Lokoč, T. Seidl, and T. Skopal. Indexing the signature quadratic form distance for efficient content-based multimedia retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 24:1–24:8, New York, NY, USA, 2011. ACM.

[22] C. Beecks, K. Schoeffmann, M. Lux, M. S. Uysal, and T. Seidl. Endoscopic video retrieval: A signature-based approach for linking endoscopic images with video segments. In *2015 IEEE International Symposium on Multimedia (ISM)*, pages 33–38, Dec 2015.

[23] C. Beecks and T. Seidl. On stability of adaptive similarity measures for content-based image retrieval. In K. Schoeffmann, B. Merialdo, A. G. Hauptmann, C.-W. Ngo, Y. Andreopoulos, and C. Breiteneder, editors, *Advances in Multimedia Modeling*, pages 346–357, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[24] C. Beecks, T. Skopal, K. Schoeffmann, and T. Seidl. Towards large-scale multimedia exploration. In *Proc. 5th International Workshop on Ranking in Databases (DBRank 2011), Seattle, WA, USA*, pages 31–33, 2011.

[25] C. Beecks, M. Uysal, and T. Seidl. A comparative study of similarity measures for content-based multimedia retrieval. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1552–1557. IEEE, July 2010.

[26] C. Beecks, M. S. Uysal, J. Hermanns, and T. Seidl. Gradient-based signatures for efficient similarity search in large-scale multimedia databases. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1241–1250, New York, NY, USA, 2015. ACM.

[27] C. Beecks, M. S. Uysal, and T. Seidl. Efficient k-nearest neighbor queries with the signature quadratic form distance. In *Proc. 4th International Workshop on Ranking in Databases (DBRank 2010) in conjunction with IEEE 26th International Conference on Data Engineering (ICDE 2010), Long Beach, California, USA*, pages 10 – 15, Washington, USA, 2010. IEEE.

[28] C. Beecks, M. S. Uysal, and T. Seidl. Signature quadratic form distance. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 438–445, New York, NY, USA, 2010. ACM.

[29] C. Beecks, M. S. Uysal, and T. Seidl. L2-signature quadratic form distance for efficient query processing in very large multimedia databases. In K.-T. Lee, W.-H. Tsai, H.-Y. Liao, T. Chen, J.-W. Hsieh, and C.-C. Tseng, editors, *Advances in Multimedia Modeling*, volume 6523 of *Lecture Notes in Computer Science*, pages 381–391. Springer Berlin Heidelberg, 2011.

[30] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sept. 1975.

[31] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pages 359–370. AAAI Press, 1994.

[32] A. Blažek, D. Kuboň, and J. Lokoč. Known-item search in video databases with textual queries. In *Similarity Search and Applications - 9th International Conference, SISAP 2016, Tokyo, Japan, October 24-26, 2016. Proceedings*, pages 117–124, 2016.

[33] A. Blažek, J. Lokoč, and D. Kuboň. Video hunter at VBS 2017. In *MultiMedia Modeling - 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II*, pages 493–498, 2017.

[34] A. Blažek, J. Lokoč, F. Matzner, and T. Skopal. Enhanced signature-based video browser. In *MultiMedia Modeling - 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II*, volume 8936 of *Lecture Notes in Computer Science*, pages 243–248. Springer International Publishing, 2015.

[35] A. Blažek, J. Lokoč, and T. Skopal. Video retrieval with feature signature sketches. In *Similarity Search and Applications - 7th International Conference, SISAP 2014, Los Cabos, Mexico, October 29-31, 2014. Proceedings*, volume 8821 of *Lecture Notes in Computer Science*, pages 25–36. Springer International Publishing, 2014.

[36] C. Böhm, S. Berchtold, and D. A. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Comput. Surv.*, 33(3):322–373, Sept. 2001.

[37] B. Braunmüller, M. Ester, H. Kriegel, and J. Sander. Multiple similarity queries: A basic DBMS operation for mining in metric databases. *IEEE Trans. Knowl. Data Eng.*, 13(1):79–95, 2001.

[38] P. Budikova, M. Batko, and P. Zezula. *Fusion Strategies for Large-Scale Multi-modal Image Retrieval*, pages 146–184. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.

[39] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6, Aug 2017.

[40] S.-F. Chang, T. Sikora, and A. Purl. Overview of the mpeg-7 standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):688–695, Jun 2001.

[41] E. Chávez, K. Figueroa, and G. Navarro. Effective proximity retrieval by ordering permutations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1647–1658, Sept. 2008.

[42] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, Sept. 2001.

[43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing.

[44] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li. A survey and analysis on automatic image annotation. *Pattern Recognition*, 79:242 – 259, 2018.

[45] T.-S. 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 ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 48:1–48:9, New York, NY, USA, 2009. ACM.

[46] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE 11th International Conference on Computer Vision*,

*ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[47] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proceedings of the 23rd International Conference on Very Large Data Bases*, VLDB '97, pages 426–435, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[48] C. Cobârzan, K. Schoeffmann, W. Bailer, W. Hürst, A. Blažek, J. Lokoč, S. Vrochidis, K. U. Barthel, and L. Rossetto. Interactive video search tools: a detailed analysis of the video browser showdown 2015. *Multimedia Tools and Applications*, 76(4):5539–5571, Feb 2017.

[49] R. C. H. Connor, L. Vadicamo, F. A. Cardillo, and F. Rabitti. Supermetric search. *Inf. Syst.*, 80:108–123, 2019.

[50] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, SCG '04, pages 253–262, New York, NY, USA, 2004. ACM.

[51] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008.

[52] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[53] E. Deza and M.-M. Deza. Chapter 1 - general definitions. In E. Deza and M.-M. Deza, editors, *Dictionary of Distances*, pages 2 – 30. Elsevier, Amsterdam, 2006.

[54] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages 647–655. JMLR.org, 2014.

[55] C. Eickhoff. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 53–56, New York, NY, USA, 2014. ACM.

[56] B. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster analysis*. Wiley, 5th edition, 2011.

[57] M. Ferecatu and D. Geman. A statistical framework for image category search from a mental picture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):1087–1101, 2009.

[58] A. Foncubierta-Rodríguez, H. Müller, and A. Depeursinge. Retrieval of high-dimensional visual data: current state, trends and challenges ahead. *Multimedia Tools and Applications*, 69(2):539–567, Mar 2014.

[59] P. Galuščáková, M. Kruliš, J. Lokoč, and P. Pecina. CUNI at mediaeval 2014 search and hyperlinking task: Visual and prosodic features in hyperlinking. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014*. CEUR-WS.org, 2014.

[60] J. Gantz and D. Reinsel. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, 2012.

[61] I. Gialampoukidis, A. Moumtzidou, D. Liparas, S. Vrochidis, and I. Kompatsiaris. A hybrid graph-based and non-linear late fusion approach for multimedia retrieval. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2016.

[62] V. Gil-Costa and M. Marin. Approximate distributed metric-space search. In *Proceedings of the 9th Workshop on Large-scale and Distributed Informational Retrieval*, LSDS-IR '11, pages 15–20, New York, NY, USA, 2011. ACM.

[63] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3174, July 2017.

[64] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[65] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 241–257, Cham, 2016. Springer International Publishing.

[66] A. D. Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (General)*, 150(2):pp. 119–137, 1987.

[67] C. Gurrin, K. Schoeffmann, H. Joho, A. Leibetseder, L. Zhou, A. Duane, D.-T. Dang-Nguyen, M. Riegler, L. Piras, M.-T. Tran, J. Lokoč, and W. Hürst. [invited papers] comparing approaches to interactive lifelog search at the lifelog search challenge (lsc2018). *ITE Transactions on Media Technology and Applications*, 7(2):46–59, 2019.

[68] C. Gurrin, K. Schoeffmann, H. Joho, B. Munzer, R. Albatal, F. Hopfgartner, L. Zhou, and D.-T. Dang-Nguyen. A test collection for interactive lifelog retrieval. In *MultiMedia Modeling*, pages 312–324, Cham, 2019. Springer International Publishing.

[69] C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.

[70] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, SIGMOD '84, pages 47–57, New York, NY, USA, 1984. ACM.

[71] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[72] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:729–736, 1995.

[73] J. He, X. Shang, H. Zhang, and T.-S. Chua. Mental visual browsing. In Q. Tian, N. Sebe, G.-J. Qi, B. Huet, R. Hong, and X. Liu, editors, *MultiMedia Modeling*, pages 424–428, Cham, 2016. Springer International Publishing.

[74] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[75] D. Heesch. A survey of browsing models for content based image retrieval. *Multimedia Tools and Applications*, 40(2):261–284, Nov 2008.

[76] M. L. Hetland. Ptolemaic indexing. `arXiv:0911.4384 [cs.DS]`, 2009.

[77] M. L. Hetland, T. Skopal, J. Lokoč, and C. Beecks. Ptolemaic access methods: Challenging the reign of the metric space model. *Inf. Syst.*, 38(7):989–1006, 2013.

[78] M. Hladík. *Lineární algebra (nejen) pro informatiky*. 2018.

[79] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6):118:1–118:36, Feb. 2019.

[80] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):797–819, 2011.

[81] W. Hürst, R. van de Werken, and M. Hoet. A storyboard-based interface for mobile video browsing. In X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. A. Hasan, editors, *MultiMedia Modeling*, pages 261–265, Cham, 2015. Springer International Publishing.

[82] D. P. Huttenlocher, G. A. Klanderman, G. A. Kl, and W. J. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:850–863, 1993.

[83] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. In *VLDB'98, Proceedings of*

*24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 218–227. Morgan Kaufmann, 1998.

[84] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[85] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, Jan 2011.

[86] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1704–1716, Sept. 2012.

[87] K. Kavukcuoglu, P. Sermanet, Y. lan Boureau, K. Gregor, M. Mathieu, and Y. L. Cun. Learning convolutional feature hierarchies for visual recognition. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1090–1098. Curran Associates, Inc., 2010.

[88] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1-3):1–6, 1998.

[89] J. Kohout, T. Komárek, P. Čech, J. Bodnár, and J. Lokoč. Learning communication patterns for malware discovery in https data. *Expert Systems with Applications*, 101:129 – 142, 2018.

[90] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1097–1105. Curran Associates, Inc., 2012.

[91] M. Kruliš, J. Lokoč, and T. Skopal. Efficient extraction of clustering-based feature signatures using gpu architectures. *Multimedia Tools and Applications*, 75(13):8071–8103, Jul 2016.

[92] M. Kruliš, J. Lokoč, C. Beecks, T. Skopal, and T. Seidl. Processing the signature quadratic form distance on many-core gpu architectures. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2373–2376, New York, NY, USA, 2011. ACM.

[93] M. Kruliš, J. Lokoč, and T. Skopal. Efficient extraction of feature signatures using multi-gpu architecture. In S. Li, A. El Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, and C. Gurrin, editors, *Advances in Multimedia Modeling*, volume 7733 of *Lecture Notes in Computer Science*, pages 446–456. Springer Berlin Heidelberg, 2013.

[94] M. Kruliš, T. Skopal, J. Lokoč, and C. Beecks. Combining CPU and GPU architectures for fast similarity search. *Distributed and Parallel Databases*, 30(3-4):179–207, 2012.

[95] C. Lewis. *Using the "thinking Aloud" Method in Cognitive Interface Design*. Research report. IBM T.J. Watson Research Center.

[96] K. Lin, H. Yang, J. Hsiao, and C. Chen. Deep learning of binary hash codes for fast image retrieval. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 27–35, June 2015.

[97] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2486–2493. IEEE, Nov 2011.

[98] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 21–37, 2016.

[99] J. Lokoč. Approximating adaptive distance measures using scalable feature signatures. *Multimedia Tools and Applications*, 74(24):11569–11594, Dec 2015.

[100] J. Lokoč, A. N. Phuong, M. Vomlelová, and C.-W. Ngo. Color-sketch simulator: A guide for color-based visual known-item search. In G. Cong, W.-C. Peng, W. E. Zhang, C. Li, and A. Sun, editors,

*Advanced Data Mining and Applications*, pages 754–763, Cham, 2017. Springer International Publishing.

[101] J. Lokoč. Approximating the signature quadratic form distance using scalable feature signatures. In C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, editors, *MultiMedia Modeling*, volume 8325 of *Lecture Notes in Computer Science*, pages 86–97. Springer International Publishing, 2014.

[102] J. Lokoč, W. Bailer, K. Schoeffmann, B. Münzer, and G. Awad. On influential trends in interactive video retrieval: Video browser showdown 2015-2017. *IEEE Trans. Multimedia*, 20(12):3361–3376, 2018.

[103] J. Lokoč, A. Blažek, and T. Skopal. Signature-based video browser. In C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, editors, *MultiMedia Modeling*, volume 8326 of *Lecture Notes in Computer Science*, pages 415–418. Springer International Publishing, 2014.

[104] J. Lokoč, T. Grošup, and T. Skopal. On scalable approximate search with the signature quadratic form distance. In N. Brisaboa, O. Pedreira, and P. Zezula, editors, *Similarity Search and Applications*, volume 8199 of *Lecture Notes in Computer Science*, pages 312–318. Springer Berlin Heidelberg, 2013.

[105] J. Lokoč, M. L. Hetland, T. Skopal, and C. Beecks. Ptolemaic indexing of the signature quadratic form distance. In *Proceedings of the Fourth International Conference on SImilarity Search and APplications*, SISAP '11, pages 9–16, New York, NY, USA, 2011. ACM.

[106] J. Lokoč, G. Kovalčík, B. Münzer, K. Schöffmann, W. Bailer, R. Gasser, S. Vrochidis, P. A. Nguyen, S. Rujikietgumjorn, and K. U. Barthel. Interactive search or sequential browsing? a detailed analysis of the video browser showdown 2018. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(1):29:1–29:18, Feb. 2019.

[107] J. Lokoč, G. Kovalčík, T. Souček, J. Moravec, J. Bodnár, and P. Čech. VIRET tool meets nasnet. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, pages 597–601, 2019.

[108] J. Lokoč, G. Kovalčík, T. Souček, J. Moravec, and P. Čech. Viret: A video retrieval tool for interactive known-item search. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada*, pages 1–5, 2019.

[109] J. Lokoč, J. Moško, P. Čech, and T. Skopal. On indexing metric spaces using cut-regions. *Information Systems*, 43(0):1 – 19, 2014.

[110] J. Lokoč, D. Novák, M. Batko, and T. Skopal. Visual image search: feature signatures or/and global descriptors. In *Proceedings of the 5th international conference on Similarity Search and Applications*, volume 7404 of *Lecture Notes in Computer Science*, pages 177–191. Springer Berlin Heidelberg, 2012.

[111] J. Lokoč, K. Schoeffmann, and M. del Fabro. Dynamic hierarchical visualization of keyframes in endoscopic video. In *MultiMedia Modeling - 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II*, pages 291–294, 2015.

[112] J. Lokoč, T. Souček, and G. Kovalčík. Using an interactive video retrieval tool for lifelog data. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC@ICMR 2018, Yokohama, Japan, June 11, 2018*, pages 15–19, 2018.

[113] J. Lokoč, P. Čech, J. Novák, and T. Skopal. Cut-region: A compact building block for hierarchical metric indexing. In G. Navarro and V. Pestov, editors, *Similarity Search and Applications*, volume 7404 of *Lecture Notes in Computer Science*, pages 85–100. Springer Berlin Heidelberg, 2012.

[114] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2. IEEE, 1999.

[115] S. J. Luck and E. K. Vogel. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, 17(8):391 – 400, 2013.

[116] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *Proceedings of the 18th ACM Conference on Information*

*and Knowledge Management*, CIKM '09, pages 255–264, New York, NY, USA, 2009. ACM.

[117] M. Macík, J. Lokoč, P. Čech, and T. Skopal. Particle physics model for content-based 3d exploration. In *14th International Workshop on Content-Based Multimedia Indexing, CBMI 2016, Bucharest, Romania, June 15-17, 2016*, pages 1–6, 2016.

[118] Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov. Scalable distributed algorithm for approximate nearest neighbor search problem in high dimensional general metric spaces. In G. Navarro and V. Pestov, editors, *Similarity Search and Applications*, volume 7404 of *Lecture Notes in Computer Science*, pages 132–147. Springer Berlin Heidelberg, 2012.

[119] V. Mic, D. Novak, and P. Zezula. Binary sketches for secondary filtering. *ACM Trans. Inf. Syst.*, 37(1):1:1–1:28, 2019.

[120] M. L. Mico, J. Oncina, and E. Vidal. A new version of the nearest-neighbour approximating and eliminating search algorithm (aesa) with linear preprocessing time and memory requirements. *Pattern Recogn. Lett.*, 15(1):9–17, 1994.

[121] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[122] D. Moise, D. Shestakov, G. Gudmundsson, and L. Amsaleg. Indexing and searching 100m images with map-reduce. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ICMR '13, pages 17–24, New York, NY, USA, 2013. ACM.

[123] A. Moumtzidou, T. Mironidis, F. Markatopoulou, S. Andreadis, I. Gialampoukidis, D. Galanopoulos, A. Ioannidou, S. Vrochidis, V. Mezaris, I. Kompatsiaris, and I. Patras. Verge in vbs 2017. In *MultiMedia Modeling*, pages 486–492, Cham, 2017. Springer International Publishing.

[124] J. Moško, J. Lokoč, and T. Skopal. Clustered pivot tables for i/o-optimized similarity search. In *Fourth International Conference on Similarity Search and Applications, SISAP 2011, Lipari Island, Italy, June 30 - July 01, 2011*, pages 17–24, 2011.

[125] MPEG-7. Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002, 2002.

[126] H. Müller, P. Clough, B. Hersh, and A. Geissbühler. Variation of relevance assessments for medical image retrieval. In S. Marchand-Maillet, E. Bruno, A. Nürnberger, and M. Detyniecki, editors, *Adaptive Multimedia Retrieval: User, Context, and Feedback*, pages 232–246, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[127] P. A. Nguyen, Y.-J. Lu, H. Zhang, and C.-W. Ngo. Enhanced vireo kis at vbs 2018. In K. Schoeffmann, T. H. Chalidabhongse, C. W. Ngo, S. Aramvith, N. E. O'Connor, Y.-S. Ho, M. Gabbouj, and A. Elgammal, editors, *MultiMedia Modeling*, pages 407–412, Cham, 2018. Springer International Publishing.

[128] P. A. Nguyen, C. Ngo, D. Francis, and B. Huet. VIREO @ video browser showdown 2019. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, pages 609–615, 2019.

[129] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[130] D. Novák, M. Batko, and P. Zezula. Metric index: An efficient and scalable solution for precise and approximate similarity search. *Information Systems*, 36(4):721–733, 2011.

[131] D. Novák, M. Batko, and P. Zezula. Large-scale similarity data management with distributed metric index. *Inf. Process. Manage.*, 48(5):855–872, Sept. 2012.

[132] D. Novák and P. Zezula. Rank aggregation of candidate sets for efficient similarity search. In *Database and Expert Systems Applications*, volume 8645 of *Lecture Notes in Computer Science*, pages 42–58. Springer International Publishing, 2014.

[133] NVIDIA. Kepler GPU Architecture `http://www.nvidia.com/object/nvidia-kepler.html`.

[134] NVIDIA. *Maxwell GPU Architecture* `http://developer.nvidia.com/maxwell-compute-architecture`.

[135] B. Park, K. Lee, and S. Lee. A new similarity measure for random signatures: Perceptually modified hausdorff distance. In J. Blanc-Talon, W. Philips, D. Popescu, and P. Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, volume 4179 of *Lecture Notes in Computer Science*, pages 990–1001. Springer Berlin Heidelberg, 2006.

[136] M. Patella and P. Ciaccia. Approximate similarity search: A multi-faceted problem. *J. Discrete Algorithms*, 7(1):36–48, 2009.

[137] L. Paulevé, H. Jégou, and L. Amsaleg. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11):1348–1358, 2010.

[138] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[139] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3384–3391. IEEE, 2010.

[140] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, pages 1–8. IEEE Computer Society, 2007.

[141] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6517–6525, 2017.

[142] D. Reinsel, J. Gantz, and J. Rydning. DATA AGE 2025: The Digitization of the World, From Edge to Core, 2018.

[143] J. J. Rocchio. *Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall Inc., 1971.

[144] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, New York, NY, USA, 2010. ACM.

[145] L. Rossetto, M. A. Parian, R. Gasser, I. Giangreco, S. Heller, and H. Schuldt. Deep learning-based concept detection in vitrivr. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, pages 616–621, 2019.

[146] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt. V3C - A research video collection. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I*, pages 349–360, 2019.

[147] Y. Rubner and C. Tomasi. *Perceptual Metrics for Image Database Navigation*. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[148] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, Nov. 2000.

[149] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[150] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, Aug. 1988.

[151] H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, 2006.

[152] S. Santini and R. Jain. Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, Sep. 1999.

[153] G. Schaefer. A next generation browsing environment for large image repositories. *Multimedia Tools and Applications*, 47(1):105–120, Mar 2010.

[154] K. Schoeffmann, D. Ahlström, and L. Böszörményi. 3d storyboards for interactive visual search. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME 2012, Melbourne, Australia, July 9-13, 2012*, pages 848–853, 2012.

[155] K. Schoeffmann, W. Bailer, C. Gurrin, G. Awad, and J. Lokoč. Interactive video search: Where is the user in the age of deep learning? In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 2101–2103, 2018.

[156] K. Schoeffmann, M. A. Hudelist, and J. Huber. Video interaction tools: A survey of recent work. *ACM Comput. Surv.*, 48(1):14:1–14:34, Sept. 2015.

[157] K. Schoeffmann, B. Münzer, A. Leibetseder, J. Primus, and S. Kletz. Autopiloting feature maps: The deep interactive video exploration (divexplore) system at VBS2019. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, pages 585–590, 2019.

[158] S. Shirdhonkar and D. Jacobs. Approximate earth movers distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8. IEEE, june 2008.

[159] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, and N. Elmqvist. *Designing the User Interface - Strategies for Effective Human-Computer Interaction, 6th Edition*. Pearson, 2016.

[160] I. Sipiran, J. Lokoč, B. Bustos, and T. Skopal. Scalable 3d shape retrieval using local features and the signature quadratic form distance. *The Visual Computer*, 33(12):1571–1585, 2017.

[161] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 1470–, Washington, DC, USA, 2003. IEEE Computer Society.

[162] T. Skopal. Unified framework for fast exact and approximate search in dissimilarity spaces. *ACM Trans. Database Syst.*, 32(4), 2007.

[163] T. Skopal, T. Bartoš, and J. Lokoč. On (not) indexing quadratic form distance by metric access methods. In *Proceedings of the 14th International Conference on Extending Database Technology*, EDBT/ICDT '11, pages 249–258, New York, NY, USA, 2011. ACM.

[164] T. Skopal and J. Lokoč. New dynamic construction techniques for m-tree. *J. Discrete Algorithms*, 7(1):62–77, 2009.

[165] T. Skopal, J. Lokoč, and B. Bustos. D-cache: Universal distance cache for metric access methods. *Knowledge and Data Engineering, IEEE Transactions on*, 24(5):868–881, May 2012.

[166] T. Skopal, J. Pokorný, and V. Snášel. Nearest Neighbours Search Using the PM-Tree. In L. Zhou, B. Ooi, and X. Meng, editors, *Database Systems for Advanced Applications*, volume 3453 of *Lecture Notes in Computer Science*, pages 803–815. Springer Berlin Heidelberg, 2005.

[167] C. G. M. Snoek, M. Worring, O. d. Rooij, K. E. A. van de Sande, R. Yan, and A. G. Hauptmann. Videolympics: Real-time evaluation of multimedia retrieval systems. *IEEE MultiMedia*, 15(1):86–91, Jan 2008.

[168] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 399–402, New York, NY, USA, 2005. ACM.

[169] G. Strong and M. Gong. Self-sorting map: An efficient algorithm for presenting multimedia data in structured layouts. *IEEE Transactions on Multimedia*, 16(4):1045–1058, June 2014.

[170] N. Suditu and F. Fleuret. Heat: Iterative relevance feedback with one million images. In *2011 International Conference on Computer Vision*, pages 2118–2125, Nov 2011.

[171] N. Suditu and F. Fleuret. Iterative relevance feedback with adaptive exploration/exploitation trade-off. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1323–1331, New York, NY, USA, 2012. ACM.

[172] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, pages 1383–1392, Aire-la-Ville, Switzerland, Switzerland, 2009. Eurographics Association.

[173] M. Swan. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *International journal of environmental research and public health*, 6(2):492–525, 2009.

[174] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1–9. IEEE, June 2015.

[175] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, June 1978.

[176] B. Thomee and M. S. Lew. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, 1(2):71–86, Jul 2012.

[177] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.

[178] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. *CoRR*, abs/1511.05879, 2015.

[179] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society.

[180] A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, 1977.

[181] R. Uribe, G. Navarro, R. Barrientos, and M. Marín. An index data structure for searching in metric space databases. In V. Alexandrov,

G. van Albada, P. Sloot, and J. Dongarra, editors, *Computational Science ICCS 2006*, volume 3991 of *Lecture Notes in Computer Science*, pages 611–617. Springer Berlin Heidelberg, 2006.

[182] M. Uysal, C. Beecks, and T. Seidl. On efficient content-based near-duplicate video detection. In *Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on*, pages 1–6. IEEE, June 2015.

[183] M. S. Uysal, C. Beecks, J. Schmücking, and T. Seidl. Efficient filter approximation using the earth mover's distance in very large multimedia databases with feature signatures. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 979–988, New York, NY, USA, 2014. ACM.

[184] M. S. Uysal, C. Beecks, and T. Seidl. On efficient query processing with the earth mover's distance. In *Proceedings of the 7th Workshop on Ph.D Students*, PIKM '14, pages 25–32, New York, NY, USA, 2014. ACM.

[185] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9: 25792605, Nov 2008.

[186] P. Čech, J. Maroušek, J. Lokoč, Y. N. Silva, and J. Starks. Comparing mapreduce-based k-nn similarity joins on hadoop for high-dimensional data. In *Advanced Data Mining and Applications - 13th International Conference, ADMA 2017, Singapore, November 5-6, 2017, Proceedings*, pages 63–75, 2017.

[187] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3637–3645, USA, 2016. Curran Associates Inc.

[188] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):652–663, 2017.

[189] E. M. Voorhees. The philosophy of information retrieval evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[190] M. Wichterich, I. Assent, P. Kranen, and T. Seidl. Efficient emd-based similarity search in multimedia databases via flexible dimensionality reduction. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 199–212, New York, NY, USA, 2008. ACM.

[191] B. Yao, F. Li, and P. Kumar. K nearest neighbor queries and knn-joins in large relational databases (almost) for free. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 4–15, March 2010.

[192] J. Zahálka, S. Rudinac, B. Jónsson, D. C. Koelma, and M. Worring. Blackthorn: Large-scale interactive multimodal learning. *IEEE Transactions on Multimedia*, 20(3):687–698, March 2018.

[193] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer US, 2006.

[194] P. Zezula, V. Dohnal, and D. Novák. *Towards Scalability of Similarity Searching*, pages 277–300. IOS Press, Amsterdam, The Netherlands, 2006.

[195] Y. Zhang, X. Liu, and C. Zhai. Information retrieval evaluation as search simulation: A general formal framework for ir evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 193–200, New York, NY, USA, 2017. ACM.

[196] W. Zhou, H. Li, and Q. Tian. Recent advance in content-based image retrieval: A literature survey. *CoRR*, abs/1706.06064, 2017.

[197] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 38(2), July 2006.

[198] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012, 2017.