UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Towards Interactive, Intelligent, and Integrated Multimedia Analytics

Zahálka, J.; Worring, M.

DOI
10.1109/VAST.2014.7042476

Publication date
2014

Document Version
Submitted manuscript

Published in
2014 IEEE Conference on Visual Analytics Science and Technology: proceedings: Paris, France, 9-14 October [i.e. November] 2014

Link to publication

Citation for published version (APA):
Zahálka, J., & Worring, M. (2014). Towards Interactive, Intelligent, and Integrated Multimedia Analytics. In M. Chen, D. Ebert, & C. North (Eds.), *2014 IEEE Conference on Visual Analytics Science and Technology: proceedings: Paris, France, 9-14 October [i.e. November] 2014* (pp. 3-12). Piscataway, NJ. https://doi.org/10.1109/VAST.2014.7042476

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:12 Oct 2026

Towards Interactive, Intelligent, and Integrated Multimedia Analytics

Jan Zahálka and Marcel Worring, *Senior Member, IEEE*

Abstract—The size and importance of visual multimedia collections grew rapidly over the last years, creating a need for sophisticated multimedia analytics systems enabling large-scale, interactive, and insightful analysis. These systems need to integrate the human's natural expertise in analyzing multimedia with the machine's ability to process large-scale data. The paper starts off with a comprehensive overview of representation, learning, and interaction techniques from both the human's and the machine's point of view. To this end, hundreds of references from the related disciplines (visual analytics, information visualization, computer vision, multimedia information retrieval) have been surveyed. Based on the survey, a novel general multimedia analytics model is synthesized. In the model, the need for semantic navigation of the collection is emphasized and multimedia analytics tasks are placed on the exploration-search axis. The axis is composed of both exploration and search in a certain proportion which changes as the analyst progresses towards insight. Categorization is proposed as a suitable umbrella task realizing the exploration-search axis in the model. Finally, the pragmatic gap, defined as the difference between the tight machine categorization model and the flexible human categorization model is identified as a crucial multimedia analytics topic.

Index Terms—Multimedia (image/video/music) visualization, machine learning.

1 Introduction

Recent years marked a rapid growth of importance and volume of multimedia data. An increasing number of scientific fields, such as physics, biology, and astronomy, utilize scientific imagery to advance the state of the art [41]. Modern medical science also relies increasingly on multimedia datasets, which greatly assist physicians in diagnosis. In the non-scientific world, the advent of smartphones made devices with good multimedia recording capabilities ubiquitous, resulting in the general public contributing large amounts of social media shared on immensely popular sites like Flickr or Facebook. Such social media bring resources for social sciences and media companies. The abundance of public multimedia data also provides police, intelligence services, and forensics with a major information source for investigation of felonies like child abuse or terrorism. As shown by these examples, multimedia datasets provide a wealth of resources and tremendous potential for knowledge gain in a wide spectrum of application areas. Being able to gain insight into multimedia datasets is thus of paramount importance. However, with current multimedia collections easily comprising millions of images and months of video, it is no longer feasible to have multimedia datasets analyzed by humans only. Sophisticated systems to assist the human analyst in assessing multimedia datasets are needed, and despite the increasing importance of multimedia in our society, few such systems exist.

Multimedia analytics, an emerging field combining visual analytics and multimedia analysis, focuses on creating systems for large-scale multimedia analysis [13]. Visual analytics, the science of analytical reasoning facilitated by interactive visual interfaces [78], has been successfully applied in diverse fields since its inception in 2005. Multimedia analysis, the other component of multimedia analytics, is an umbrella term for many different automatic multimedia analysis techniques. In this paper, we focus on visual multimedia collections (images/videos) with associated data sources like text annotations and metadata, making the fields of computer vision, image retrieval, and video retrieval our focus prism [73][75]. Multimedia analytics aims to guide the analyst to deep insight. They aim to combine the analyst's natural expertise in analyzing multimedia information (humans master this skill shortly after birth) with the memory and processing power of the machines, which are able to process contemporary large-scale collections. True to the visual analytics spirit embodied by the visual analytics process diagram by Keim et al. [40][41], this integration needs to be **interactive**. Multimedia analytics also need to be able to utilize the heterogeneous data sources within a collection, combining the visual content with text annotations and string/numeric metadata. The aims of multimedia analytics are ambitious, and the integration of relevant techniques to fulfill these aims is far from trivial.

The difficulty of fulfilling the multimedia analystics ambitions stems from multimedia collections being quite a specific data source. While humans perceive multimedia chiefly through the semantic properties of the visual content, machines need a mathematical representation, which does not provide comparable semantic richness and is unintuitive for a human. The heterogeneity of data sources within a multimedia collection also provides a challenge, both for the visualization and the model. Moreover, any one multimedia data instance has a much higher information bandwidth and also size than an instance in a classic dataset. This results in a larger computational load on the machine, and techniques related to multimedia analytics need to be carefully examined before adaptation.

In this paper, we provide a comprehensive overview of related state-of-the-art techniques and theory: in Section 2, we focus on the human perception of multimedia; Section 3 reviews machine processing of multimedia data. To this end, we processed the relevant literature in the following four steps:

1. Exhaustive search on articles since 2003 in leading journals and conference proceedings → **thousands of references**
2. Filtering abstracts and titles based on topical relevance, i.e., papers concerning multimedia analytics, visual analytics theory, multimedia visualization or multimedia analysis → ∼**800 references**
3. Topical-relevance filtering based on the key sections of the content → ∼**370 references** available online [96]
4. Final filtering based on topical relevance of the complete content and on the impact of the article in the respective community (measured by Google Scholar citations) → ∼**100 references** in this paper

In Section 4, we synthesize the relevant techniques into a general multimedia analytics model, which is, to the best of our knowledge, still largely missing in the literature. Section 5 concludes the paper.

- *Jan Zahálka is with the University of Amsterdam. E-mail: j.zahalka@uva.nl.*
- *Marcel Worring is with the University of Amsterdam. E-mail: m.worring@uva.nl.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

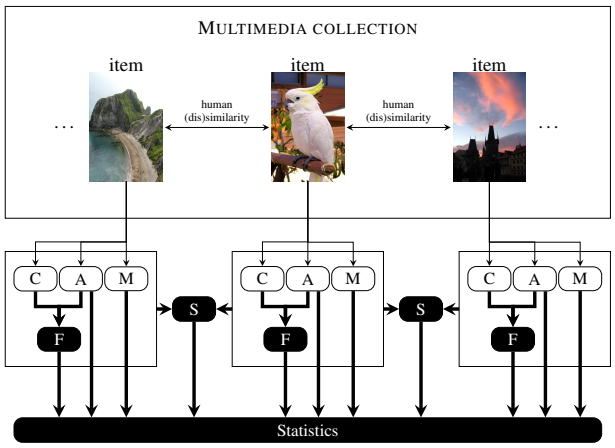

Fig. 1. Representing multimedia: a multimedia collection comprises individual items, which are composed of content (C), annotations (A), and metadata (M). Thick arrows depict data transformations into derived values (white text, black background): features (F), (dis)similarity (S) and statistics.

2 HUMAN PERCEPTION OF MULTIMEDIA

The human's excellent capability to analyze multimedia heavily revolves around being able to see the multimedia items in question. Hence, this section is focused on visualization techniques and analytic interfaces. Issues related to the human processing of multimedia data are outlined in Section 2.1. The text focuses on issues related to data representation and the technical aspects of multimedia visualization. Other important issues, whose thorough description is beyond the scope of this paper, also exist: examples include the cognitive aspects of image understanding [4] and relevance judgment [30] by the brain, quality of the data [28], or performance issues. Multimedia browsers, the interactive interfaces used to access multimedia collections, are surveyed in Section 2.2.

2.1 Representation and learning

When working with multimedia data, we are initially provided with a **multimedia collection**, structurally depicted in Figure 1. Multimedia collections comprise individual raw multimedia **items**, i.e., single images and/or video clips. Individual items consist of three elements — content, annotations, and metadata. The visual information present in the item and conveyed by it is the item's **content**. Those individual pieces of visual information carrying an objective semantic meaning, for example "cat" or "dog," are called **semantic concepts**. **Annotations** are text descriptors such as labels, captions, or tags assigned to individual items by a human which relate to the content in an objective or subjective manner. **Metadata** are string or numeric descriptors related to the technical parameters of the item, such as Exif information or GPS localization.

Humans are excellent multimedia analysts by nature, trained to process high-bandwidth visual information through the visual cortex since their early days. Humans can extract semantic information directly from raw multimedia. Then, they synthesize it into concepts and similarity measures with complexity ranging from purely visual-based through visually-semantic (i.e., representing semantics grounded in visual characteristics) to completely abstract [88]. This entire process is remarkably fast. The reasoning about the data and its structure is heavily context- and intent-based. To illustrate, consider a toy image collection containing a picture of a British phone booth, a picture of Little Red Riding Hood, and a picture of Queen Elizabeth II. The notion of similarity (and thus structure) can be based on the "amount of red colour," "person," and "United Kingdom" characteristics, corresponding to a visual-based, visually-semantic, and abstract similarity notion, respectively. Each conveys different structure and none of them is inherently wrong. This learning and reasoning process shows that humans do not perceive multimedia content in a mathematical manner. Mathematical summaries and statistics are thus less prominent than in the case of classic visual analytics approaches. Indeed, Santini and Jain show that modelling multimedia similarities mathematically is rather treacherous [70]. However, they are still useful for annotations, metadata, and simple visual statistics like the amount of a certain colour. The main limitation of humans is limited cognitive capacity [29], barring any attempt at brute-force analysis of a large-scale collection containing million images or more. Analytic interfaces thus need to be **semantically navigable** to support the human's reasoning process.

Visualization	Efficiency	Navigability	Heterogeneous
Basic grid	+	−	−
Similarity space	−	++	−
Similarity-based	+	+	−
Spreadsheet	∗	∗	++
Thread	−	+	+

Table 1. Summary of analytic capabilities of multimedia visualizations: screen space **efficiency**, semantic **navigability**, and integration of **heterogeneous** information channels, i.e., content, annotations and metadata. The values range from − (poor) through + (good) to ++ (excellent), ∗ denotes a value dependent on the task at hand (explained in the text).

Since a human needs to see the items in question to derive information from them, the most natural visualization paradigm is directly displaying the items or their respective thumbnails on the screen. Using primitives and conventional charts for visual content is in principle possible, but quite unorthodox: the visual content, a powerful information channel, is lost, diminishing the ability to discriminate between content classes [24]. Direct visualization has large demands for screen space. The human needs to see the content, so each thumbnail needs to be big enough and occlusion should be minimized. As many items as possible should be displayed on the screen. Analytic interfaces thus need to be **screen-space-efficient**, minimizing unused screen space. Annotations and metadata might also provide a valuable information gain. They should be visualized in an **integrated manner** within an analytic interface.

2.2 Multimedia visualization

In the current big data era, even the most extravagantly large screens cannot contain entire large-scale collections, and even if they could, the displayed data exceeds the human's cognitive limit. **Interactivity** is thus crucial for any analytic interface. The classic approach towards visualizing multimedia is the **multimedia browser**, a prime example of a casual information visualization tool [67]. Multimedia browsers are used daily by a large number of computer users to examine multimedia collections. Examples include Internet image search interfaces like Google Images or Bing Images and social sites like Flickr or Instagram. Based on the conducted survey, we grouped the state of the art into five groups based on the visual paradigm used. While this grouping is necessarily non-exhaustive, it helps identifying the suitability of state-of-the-art multimedia browsers for different tasks. The main multimedia visualization techniques discussed in this section are conceptually depicted in Figure 2, their analytic aspects are summarized in Table 1.

The most prevailing multimedia collection visualization is a two-dimensional **grid** of thumbnails. The grid is navigated by scrolling, usually one-dimensional and vertical. The user interacts with the thumbnails, usually magnifying the image and revealing annotations and/or metadata in a side panel. A grid is quite efficient in utilizing screen space: except for a side panel (if present), all of it is devoted to displaying individual items with zero overlap. A grid browser may also feature sorting, filtering and/or hierarchical display. In a basic grid, these interactions rely only on non-semantic attributes like file name or time. This makes the semantic exploration capability of a basic grid difficult. This problem is tackled by approaches which order the items in the grid according to a similarity measure. These ap-

proaches are discussed below as "similarity-based." Overall, a grid is certainly a very familiar, screen-space efficient, and usable interface. Its strong suit are tasks where viewing the full content of the item is imperative: high screen-space efficiency ensures high visibility of the thumbnails, and the user can typically magnify the item of interest and inspect it. One of the issues are the lack of integration of annotations and metadata, which are displayed separately in the side panel. Also, in the case of the basic, non-similarity based grid, semantic navigation of large collections is limited. A basic grid is thus unsuitable for tasks where the overall structure of the collection is of importance.

Another approach is the **similarity space** browser. The items' position on the screen is determined by a projection from the high-dimensional feature space to the 2-D screen space. The projection should preserve the similarity and the resulting structure present in the feature space as accurately as possible [60][69]. Examples of similarity space browsers are numerous, including the browser by Liu et al. [51], the semantic image browser by Yang et al. [92], the news video databases browser by Luo et al. [54], or the Flickr summarization browser by Fan et al. [23]. Similarity space browsers convey a notion of structure, making them excellently navigable. Scalability and screen space efficiency are an issue, however: some parts of the screen space are empty, and thus wasted; some parts may be cluttered with overlapping thumbnails. Yang et al. tackled this problem using miniature thumbnails [92]. These display the distribution of colour in the collection quite well, but lack the expressiveness to convey other visual content. Fan et al. use representative images to represent a semantic similarity neighbourhood [23]. Annotations and metadata are typically displayed in a side panel, making the integration rather poor. However, in some systems, annotations and metadata are used as a basis for the entire similarity structure, for example in Chronosphere by Worring et al. [89]. Overall, similarity space browsers are useful as a semantics-conveying paradigm due to them directly showing data structure. This makes them an excellent choice for applications focused on uncovering the structure of the collection. However, the screen space and scalability issues make it rather difficult to inspect individual items of interest.

Similarity-based approaches utilize a space-filling view (typically a grid or a treemap) where the items are arranged based on a similarity measure. Rodden et al. showed that arranging similar images together in the grid indeed helps to determine the clusters in the collection, but also somewhat hampers serendipitous discovery, due to interesting images standing out less [69]. Bederson's PhotoMesa adapts the treemap algorithm to establish similarity [3]. PhotoMesa's interface is zoomable, which along with the grouping provides overview of the collection and its structure. Zavesky et al. map image features of interest into a 2D abstraction layer based on similarity and then snap the images to a grid, forming visual islands [97]. Visual islands can be further used for guided navigation: items of interest can be used as probes to rearranging the visual islands. Quadrianto et al. introduce a multiple-tier semantic hierarchy [68], with the grid in each successive tier showing representative images for an increasingly detailed neighbourhood of interest. The approach by Brivio et al. utilizes Voronoi diagrams to fill the display, with the focus item placed in the center and other items being represented on the screen based on their distance to the focus with respect to an ordering of the collection [7]. Overall, similarity-based approaches combine screen space efficiency with an enhanced notion of structure through visually contiguous regions. This makes them easily navigable and suitable for both structure overview and inspection of individual items. Annotations and metadata are typically still relegated to a side panel, lacking true integration.

In the **spreadsheet** paradigm, rows and columns represent different dimensions of annotations and metadata, and cells display individual items. The first browser based on the spreadsheet paradigm is PhotoSpread by Kandel et al. [39] MediaTable by de Rooij et al. visualizes each item in one row, with its content, annotations, and metadata forming individual columns [20]. Multimedia Pivot Tables by Worring and Koelma allow the analyst to define the rows and columns of the table herself [90]. Navigability and screen space efficiency are dependent on the role of annotations and metadata in the task at hand. If those

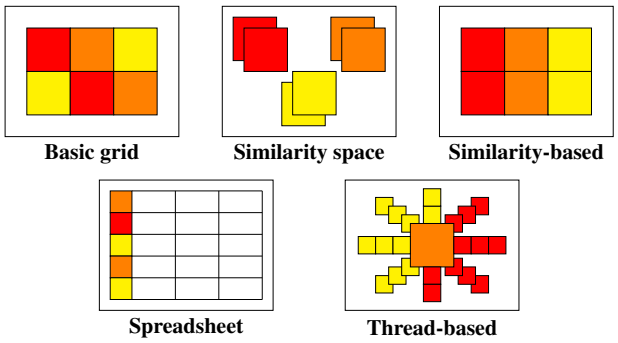

Fig. 2. Multimedia visualizations (conceptual depiction).

are missing or their role is marginal, then a spreadsheet browser is a poor choice, since only few columns will be filled and informative. When annotations and metadata are crucial, however, the spreadsheet provides a very screen-space-efficient way to explore the collection. Other approaches centered on annotations and metadata are driven by specific types of metadata, especially geographic location. Examples include the browser summarizing photographs based on their location by Jaffe et al. [36] or building a 3D scene of the location from the corresponding photos, as conceived by Szeliski et al. [76] Out of all the groups mentioned in this paper, spreadsheets emphasize and integrate annotations and metadata the best, making it a great choice for tasks with heavy involvement of annotations and metadata. Moreover, the individual items are easily inspectable. The spreadsheet paradigm, however, falls short when the focus of the task is purely or mostly visual content.

Another approach is the **thread** browser, where the collection can be navigated along threads, i.e., sequences of items based on a certain criterion. Originally, this paradigm was used by de Rooij et al. for navigating large video collections [17], where threads were temporal, semantic or based on annotations and metadata ordering. In a thread browser, the interface displays the focus item at the center of the screen and a number of threads relevant to the focus. The user can then navigate along a thread of choice, shifting the focus to the item along the chosen thread. The user is thus able to navigate the collection without requiring her to overview large number of items at once or search manually. Multiple browsers implement this paradigm, the difference lies in the number of threads displayed at once: the CrossBrowser displays two, the ForkBrowser five [18], the RotorBrowser up to eight [17]. The thread browser's design is centered around semantic navigation, which makes it excellent for tasks where inspecting individual items based on semantic dimensions is imperative. Annotation- and metadata-based threads also integrate the heterogeneous data channels seamlessly. Screen-space efficiency is the main issue of the thread browser: only a limited number of threads can be displayed in order not to overwhelm the user [17], therefore a part of the screen is empty. This makes the thread browser lacking in cases where an overview of the structure of the collection is needed.

There is a great variety of multimedia visualizations, with our text covering the main approaches and being by no means exhaustive. From the analytic perspective, there is no strongly dominant solution, each is strong in different areas and suitable for different tasks. The quest for visualizations and metaphors suitable for multimedia analytics thus remains open.

3 Machine Perception of Multimedia

Unlike humans, machines have an excellent capability to process large-scale data, thanks to their large memory and processing power. This makes them excellent assistants to the human analyst, who struggles when faced with large collections. In this section, techniques enabling the machine to provide such assistance are reviewed: Section 3.1 focuses on machine representations of multimedia and machine learning in multimedia analysis, chiefly focusing on feature extraction,

supervised and unsupervised learning [31], and ranking. Interactivity in machine learning is surveyed in Section 3.2.

3.1 Representation and learning

The workflow of machine multimedia analysis algorithms is depicted in Figure 3. Machines need to extract an explicit and mathematical intermediate representation of the data in order to extract semantics. Building the representation starts with **features**, i.e., numeric values derived from pixels in a single item mimicking the low-level representation used by the human's perceptual system. There are many kinds of features, each being a different abstraction from pixel data. Typically, multiple features per item are computed, resulting in a **feature vector** representing that item. Feature vectors are in turn aggregated into a **feature representation** of the entire collection. Annotations can be treated either as string data and represented directly, or converted to text features or conceptual representations if semantics extraction is needed. Metadata are machine-readable per se and require no conversion. Extracting semantics from feature representations is then performed by machine learning.

The semantics extraction process faces numerous challenges. The main one is the **semantic gap** defined by Smeulders et al., i.e., the disproportion between the information extractable from a multimedia item's content by a human and the information extractable from the feature representation of the same item by a machine [73]. Simply put, machines are essential to multimedia analytics due to their capability to handle large multimedia collection much better than a human, but their understanding of multimedia is worse than a human's. In addition, not only is the human representation difficult for machines, but also vice versa. The human's semantic, non-numeric perception of multimedia renders the vast majority of features unintuitive. Since most features carry no meaning to a human, the usefulness of modelling the data using basic feature statistics is limited. All in all, the semantic gap remains a major research challenge prohibiting easy high-level extraction of semantics from multimedia data.

In the current state of the art, there are essentially two pipelines for semantic multimedia analysis: **explicit feature extraction followed by classification**, and **deep learning**. The former is the more classic one. The most used features are local, each of them corresponding to a certain region within the image. Their advantage is a number of invariances, for example positional invariance, i.e., a concept is detected irregardless of its position in an item. The dominant one is the scale-invariant feature transform (SIFT) by Lowe [53]. The existing variations of SIFT, each focused on a different visual aspect, are surveyed by Van de Sande et al. [80]. A faster option, achieving similar properties to SIFT, are the speeded-up robust features (SURF) by Bay et al. [2]. GIST by Oliva and Torralba is an example of a global feature, which computes scene characteristics directly from the data [63]. Feature representations involve grouping similar feature vectors together (typically using quantization and/or histograms). The dominant representations in the state of the art are the histogram of oriented gradients (HOG) by Dalal and Triggs [15], bag of visual words (BoW) by Sivic and Zisserman [72], Fisher vectors by Perronnin et al. [64], and the vector of locally aggregated descriptors (VLAD) by Jégou et al. [37]. Feature representations can be further enhanced by incorporating spatial information, using spatial pyramids by Lazebnik et al. [45], part-based models by Felzenszwalb et al. [26] or codemaps by Li et al. [50]. Hashing, surveyed by Zhang and Rui [98], is often used in order to reduce dimensionality. The second step of the pipeline, classification, is in almost all cases carried out by a support vector machine (SVM) by Cortes and Vapnik [14], although other classifiers such as nearest neighbour [6] or random forests [22] have also been used. The second pipeline, deep learning, revolves around using deep neural nets, which extract features and semantics using one model. In the visual multimedia domain, convolutional neural networks (CNN) by LeCun and Bengio are used [46]. While the original idea of a convolutional neural network can be traced back to the 80s, efficient algorithms for training have not existed until fairly recently: the first efficient algorithm for a deep net was conceived in 2006 by Hinton et al. [34]. Since then, deep nets have successfuly been used both in a narrow domain as shown by Tang et al. in their face recognition experiments [77], and a broad domain such as the ImageNet classification performed by Krizhevsky et al. [43] The ImageNet paper is considered a breakthrough in computer vision, establishing deep learning as the currently dominant approach. Both pipelines remain viable and actively researched though, and overall, the quality of semantics extraction in the state of the art has been steadily rising in recent years [74].

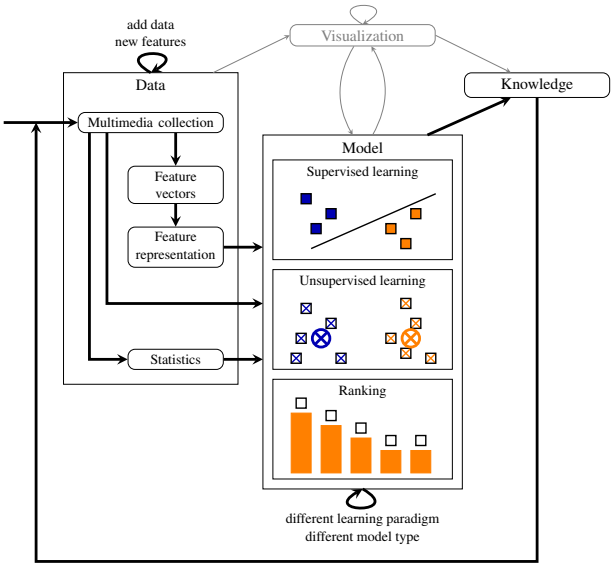

Fig. 3. Workflow of machine-centered approaches related to the visual analytics process diagram by Keim et al. [40][41]. Elements of the visual analytics process not utilized in machine processing are greyed out.

The typical case of supervised learning in semantics extraction involves classic classification and using only one information channel of the data (typically visual). In multimedia analysis, more and more effort is being made to make the task and the data more flexible. Discovery of new classes and transfer of knowledge between them is the domain of zero-shot learning, which is predominantly realized by attribute learning, i.e., representing each class by a set of attributes [44]. New classes are characterized by attributes learned from previously available classes. The three heterogeneous data sources (content, annotations, metadata) are widely utilized for learning in the video domain [75]. In the image domain, surprisingly enough, comparatively little attention has been given to this phenomenon, even though annotations and metadata have been shown to improve the quality of the machine learning model in several studies [9][11][49][87][91]. This situation is changing due to the advent of social media, which provide vast annotated collections at the cost of high noise. Learning to annotate, i.e., assign annotations based on the content, has received much attention in the recent years. Annotation as a task is essentially a special case of classification, and the main approaches exploit this [10][27][84][86]. The classic annotation approaches rely on training on reliable expert annotations, whose reliability is not a given in the social media era. The emerging field of social image retrieval caters for the low quality and high noise of social tag annotations by incorporating tag relevance learning [48]. To summarize, supervised learning is very well-studied in the multimedia domain, with an array of well-performing approaches with increasingly flexible and heterogeneous models.

Unsupervised learning in the multimedia domain, useful for structuring the collection, is wholly dependent on the feature representation and similarity measure used. Since neither the representation nor the similarity measure is canonical, clusters of multimedia items convey *a* structure of the data, rather than *the* structure of the data, as shown on the "Queen-Little Red Riding Hood-phone booth" example in Section 2.1. Technically, unsupervised learning algorithms do not differ

from those used on conventional datasets, since once the collection is converted to features, an unsupervised algorithm does not distinguish whether it is operating on feature representation or on a conventional dataset [16]. Mainstream algorithms thus include k-means [52], the EM algorithm [21], and hierarchical clustering, reflecting the application invariance of unsupervised learning.

Ranking is crucially important for semantic search, with search results being a reranking of the collection based on relevance to the search query. There are two dominant approaches towards semantic search: text-based search and content-based search. **Text-based search** relies on performing text search on annotations, while **content-based search** performs the search based on the semantics extracted from visual content [98]. Text-based search is the more mature field, but depends heavily on accuracy and relevance of annotations, a handicap in the current social tagging era. Content-based search, on the other hand, suffers from the semantic gap. Semantic search in multimedia collections does not have a canonical query scheme. Dominant query methods, surveyed by Snoek and Worring [75], comprise query by keyword (matching the query against annotations/metadata), query by example (the user provides an example item and wants similar results), and query by concept (results are based on relevance to the specified concept). A comprehensive survey of ranking methods in semantic search has been provided by Mei et al. [57]. Similarly to supervised learning, ranking has also received much research attention, with the respective state of the art providing respectable performance.

3.2 Interaction

Interactive machine learning techniques operate in a semi-supervised setting where labels exist only for a small fraction of the otherwise unlabeled dataset [12]. This is precisely the situation we face in multimedia analytics. The flow of interactive machine learning is depicted in Figure 4. The quality of the results of an interactive machine learning technique is determined by three factors: relevance, speed of convergence, and speed of response. Relevance of items in each round and speed of convergence are competing principles. The trade-off between them is known in the literature as the **precision-recall trade-off** [25]. Increasing **precision**, i.e., the proportion of relevant items in each round, harms **recall**, the proportion of the relevant items in the whole dataset returned during all interaction rounds, and vice versa. Each system thus needs to select the position on the precision-recall curve carefully. Search favours precision, since the highest number of relevant items in the early round(s) is imperative; while thorough exploration leans towards recall, since we need to see as many relevant items as possible in a small number of rounds. The speed of response is a hard constraint: interactivity imposes the time between the user input and the system response not exceeding the order of seconds. Relevance of results, speed of convergence, and response are all key to interactive machine learning employed in multimedia analytics.

The precision-recall trade-off is most reflected in the query strategy of interactive machine learners, i.e., the selection of items whose relevance is to be judged by the user. The original, and to date still used interactive machine learning paradigm is **relevance feedback**, surveyed by Zhou and Huang [100]. Relevance feedback approaches maximize precision, showing the items the learner is *most* certain about being relevant. Presenting those items means a positive effect on the user, but little gain for the learner, since it is already certain about them. Another approach, currently the dominant one, is **active learning**, which queries items the learner is *least* certain about, i.e., those with the biggest information gain. Query strategies in active learning are extensively surveyed by Settles [71]. Maximizing the information gain for the learner results in better recall and results overall, but requires a patient, cooperative user. The queried items that are difficult for the learner might also be difficult for the user. This phenomenon is called variable labelling cost, with several works devoted to predicting the cost [82] and conducting active learning on a labelling budget [83]. For multimedia analytics, both relevance feedback and active learning have their merit: we need to convince the analyst that the machine is learning by showing more relevant items each round, but we also need to maximize the information gain per interaction to retrieve as many relevant items in total as possible.

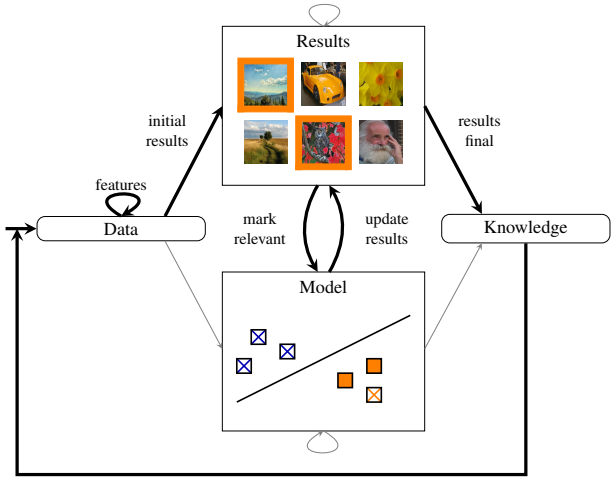

Fig. 4. Interactive machine learning workflow, incorporated into the visual analytics process diagram by Keim et al. [40][41]. Elements of the visual analytics process not utilized in interactive machine learning are greyed out.

The technical execution of interactive machine learning boils down to two key aspects: the interactions and the algorithms. The most dominant interaction is marking relevant items, possibly with degrees of certainty or explicit marking of irrelevant items [100]. A newer trend is relative feedback, where the analyst states what attributes the queried item lacks in comparison to the relevant one, enabling pruning items which are more lacking in the particular attribute than the queried item. Examples of approaches successfully using this paradigm include WhittleSearch by Kovashka and Grauman [42] and the weighted online attribute learner by Biswas and Parikh [5]. State-of-the-art active learning approaches are thus shifting from simpler relevance indications to more complex, informative ones. Algorithm-wise, the three dominant approaches in active learning, as surveyed and explained in detail by Huang et al. [35], are active SVM by Tong and Chang [79], biased discriminant analysis (BDA) by Zhou and Huang [99], and rank-based approaches surveyed by Mei et al. in their broader ranking survey [57]. Approaches with a more flexible model also exist. Weak class labels have been considered by Mitra et al. in their probabilistic active SVM [59], as well as in the conditional active learning setting by Li and Sethi [47]. Evolving classes and new class discovery are considered for example by the binary feedback framework by Joshi et al. [38], and are surveyed by Settles [71]. Overall, interactive machine learning provides a wide array of techniques with an increasingly elaborate interaction structure, showing promise with respect to incorporation into multimedia analytics.

4 Multimedia Analytics

Now that the individual human-centered and machine-centered components and techniques are reviewed, the multimedia analytics model can be proposed. This is done in Section 4.1. Section 4.2 reviews pioneer multimedia analytics systems in view of the desired capabilities. Section 4.3 presents a research agenda for multimedia analytics.

4.1 Model

The previous sections structurally revised the preliminaries of a multimedia analytics system, with emphasis on interactivity, joint utilization of the heterogeneous data sources in the collection (content, annotations, and metadata) and semantic navigability. The strong need for semantic navigability and the related semantic gap phenomenon are especially important considerations and the main difference between multimedia analytics and the closely-related classic visual analytics.

The proposed multimedia analytics model, which we explain in this section, is depicted in Figure 5 and divided into four tiers. The first

and lowest tier, **methods**, represents atomic techniques of visualization, interaction and data analysis in a multimedia analytics setting. The second tier, **intents**, represents combinations of methods which express individual intentions of the analyst. The third tier, **procedure**, corresponds to a high-level model of activities and intents taking the analyst from the beginning to the completion of her objective. **Objectives**, the fourth and top tier, are the master goals of the analyst. To illustrate, consider a medical scientist using a multimedia analytics system on a collection of medical scans. Her objective is to find the incidence of cancer in the population. The procedure to take the analyst to her objective involves exploring the scans, searching for symptoms of cancer and determining the distribution of cancer within the population. Each of the steps taken in the procedure exhibits a certain intent, e.g., "sort the patients based on liver abnormality." This intent is accomplished for example by the following methods: computing the ranking, visualizing thumbnails of the scans in a grid ordered by the ranking, the analyst panning over the grid, and magnifying the thumbnails. The four-tiered model thus provides a structured overview of all cornerstones of multimedia analytics.

Let us first examine the objectives. The main, high-level multimedia analytics objective is to guide the analyst through large and complex multimedia collections to knowledge. This knowledge ranges from abstract to particular. Abstract knowledge means that the analyst knows something about the data she did not before, understands the data, or grasps their structure. Abstract knowledge gain models are well known in visual analytics literature. Pirolli and Card coined the term sensemaking [66]. Insight is maybe the most used term for abstract knowledge gain, appearing both in information visualization [62][95] and visual analytics [40][41][78]. Insight is a notoriously fickle term evading the boundaries of precise definition. Rather, North enumerates the characteristics of insight: insight is complex, deep, qualitative, unexpected, and relevant [62]. Particular knowledge gain means completing a complex, high-level task, like the medical scientist determining the incidence of cancer. Gaining particular knowledge involves gaining abstract knowledge: in any domain, the analyst still needs to explore and understand the data. Hence, from the model point of view, we consider the different terms for knowledge gain largely equivalent. In the rest of the paper, we will use the term **insight** to indicate the main objective.

Several works develop the methods and intents as incorporated in the model. Their grouping in the model depicted in Figure 5 is inspired by the work of Pike et al. [65]. For analytic purposes, it is highly desirable that they operate on the semantics of the content. Thus, they need to be realized by employing machine learning techniques, and as such are affected by the semantic gap. **Structural methods**, focusing on uncovering the structure of the data, correspond to the interactions by Amar et al. [1]. Their usefulness in the multimedia domain is limited: since the human perception of multimedia is non-numeric, retrieve value, compute derived value, find extremum, determine range, and characterize distribution are all meaningless in their purely numeric sense. Find anomalies, cluster, and correlate are viable, but the analyst needs to see the items and perform them in a semantic manner. **Visualization methods** (particular visualization techniques like charts or grids) and **interaction methods** (atomic interactions like panning or zooming), both adapted from the work of Pike et al. [65] are in their technical sense unaffected by the semantic gap. **Interaction intents**, corresponding to the 7 interactions by Yi et al. [94], represent individual steps towards insight. Four are impacted by the semantic gap: explore, reconfigure, filter, and abstract/elaborate. Indeed, the analyst might want to navigate further based on a semantic dimension (exploration, "show me other dogs of this colour"); sort based on a semantic concept (reconfigure, "sort dresses from least formal to most formal"); filter semantically ("show me only pictures of people hiking"); and visualize item subsets based on concept hierarchies, if present (abstract/elaborate, "show me all animals → all mammals → all foxes"). The usefulness of "encode" in the visual domain is limited due to the items being displayed directly. Finally, the **visualization intents**, conceived by Pike et al. [65], represent visualization capabilities leading to insight. In the multimedia domain, "differentiate,"

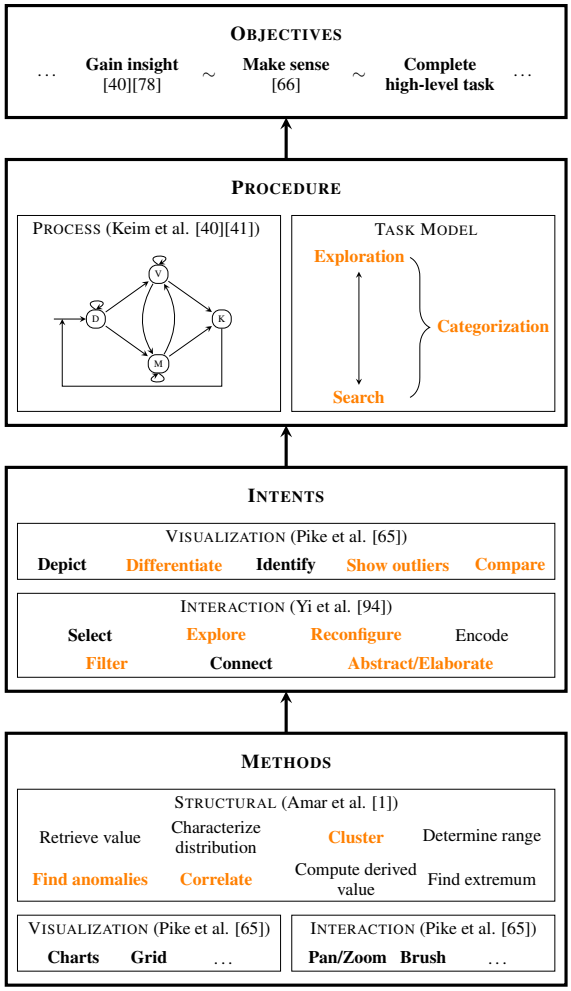

Fig. 5. The four-tiered multimedia analytics model. The tiers progress from low-level methods (bottom) to high-level objectives (top). **Bold** text indicates components critical for multimedia analytics tasks, orange text denotes components desired to operate on content semantics.

"show outliers," and "compare" are expected to be semantic. All intents and methods affected by the semantic gap are precisely those that are actually related to the analysis of the collection as a data resource. Indeed, if we select only those methods and intents unaffected by the semantic gap, we end up with the capabilities of a generic multimedia browser as discussed in Section 2.2. These are certainly useful for browsing and visualizing data, but might not be very suitable for the analysis of large-scale multimedia resources. This emphasizes the need to maintain a semantic model of the data to facilitate semantic methods and intents.

Now that the high-level objectives and the lower-level methods and intents are in place, the model of the procedure actually bringing the analyst to her objectives using the methods and intents needs to be discussed. In visual analytics, the process of attaining insight in visual analytics has been modelled by Keim et al. [40][41]. This process, which takes the analyst from data processing through iteratively updated visualization and model to knowledge, is also highly relevant in the multimedia analytics domain and it is highly desirable its flow gets fully adapted. Its **iterative** flow captures the nature of building insight, which takes time, non-trivial interactions, and reasoning. The visual analytics process also advocates tight integration of the visualization with the model, a notion equally important for multimedia analytics due to the emphasis on semantics. Bearing in mind the nature of multimedia analytics objectives, there are two basic approaches towards attaining them [55]:

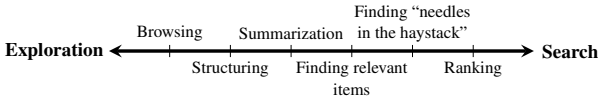

Fig. 6. Exploration-search axis with example tasks.

- **Exploration**, applicable when the analyst is faced with a collection she does not know much about beforehand, and wants to discover what is inside and/or how the data are structured. An exploratory session typically takes time and involves a very dynamic model of the data, continuously refined as the analyst iteratively gains knowledge.

- **Search**, applicable when the analyst has a clear idea what she is looking for and queries the system for items relevant to certain attributes. A search session is then a sequence of query-response pairs, and the analyst expects fast response. The data model is static, since the analyst knows exactly what she is looking for, and communicated to the system through a query.

A typical multimedia analytics task has elements of both. For example, a forensics expert might not only want to judge if the multimedia content of a suspect's seized computer contains incriminating content (search), but also determine the full extent of the suspect's illegal activities (exploration). Based on this observation, we model multimedia analytics tasks as lying on an **exploration-search axis** as depicted in Figure 6. During the analytic session, the proportion of exploration and search in a task at hand changes over time. Typically, the task is more exploration-oriented in the beginning, as the analyst does not know much about the collection yet, and progress more towards search when the analyst already has a good grasp of the collection's content and knows what to look for. The exploration-search axis is an umbrella model for most multimedia analytics tasks. The depiction of some archetypal tasks in the multimedia domain is included in Figure 6. In order to help the analyst achieve insight, multimedia analytics systems need to enable the user to alternate between exploration and search.

Even though exploration and search have different, sometimes antagonistic properties, a multimedia analytics model needs to incorporate exploration and search integrally. Otherwise, the analyst cannot alternate between the two and the system breaks down into disjoint exploratory and search components. The analyst perceives multimedia content through attributes which are chiefly semantic, optionally also involving annotation and metadata and the related statistics. Pure search is then simply filtering or reranking based on those attributes. When building a mental model of the data, the analyst structures her interest into attribute aggregates, e.g., "hiking in the Alps" (combining the "person" and "outdoors" semantic concepts with geo coordinates matching the Alps region). The attributes are categorical; the aggregations of the attributes into meaningful concepts, if we disregard fringe cases where the semantics of the content play no role, are thus **categories**. The analyst then assigns, often implicitly, its labels to individual items. This applies, for example, to the nowadays typical tasks where we select images based on some notion of relevance or interest, like picking out nice photos from the collection shot on vacation or searching for pictures of a favourite celebrity. Even though we might not explicitly assign labels, by picking relevant images we actually categorize them as relevant, while the non-picked ones are effectively categorized as irrelevant. Exploration is then an instance of categorization with few dynamically evolving categories, while search could be instantiated as categorization with fixed categories corresponding to the degree of relevance of items to the search query. Indeed, the vast majority of multimedia analysis techniques concerning semantic tasks is a variation on supervised learning, and classification (i.e., categorization) in particular. Categorization is thus a useful umbrella task for the exploration-search axis, and thus, by proxy, able to accomodate most multimedia analytics tasks:

System capabilities	**Limited**	**Intermediate**	**Advanced**
Semantic gap	Non-semantic model, only metadata and basic visual characteristics (e.g., color)	Model uses objective semantic concepts (e.g., "person")	Model uses high-level, complex semantics (e.g., "this patient has cancer")
Pragmatic gap	Non-adaptive, rigid model (classic classification)	Model adaptive to interactions, categories static (= fixed classes)	Model fully adapts to user intent, dynamic categories

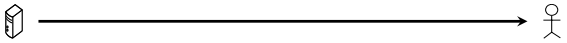

Fig. 7. Semantic and pragmatic gaps and their effect on multimedia analytics models.

- **Categorization** — the task of assigning individual items into categories from a category model. The category model is defined by the analyst based on attribute aggregates.

As discussed extensively in the previous text and shown by the majority of the model components involving semantics in Figure 5, categorization in multimedia analytics is heavily revolving around semantics. In addition, visual multimedia categorization is much less dependent on statistics (semantic statistics are unintuitive and hard to obtain) and has much lower tolerance for error from the users (since the individual errors are spotted instantly) than categorization on other types of multimedia collections. Hence, it is doubly imperative that the learned categorical model very closely follows the mental model of the analyst. There is, however, a difference between analytic categorization as a model of human reasoning and categorization or classification in the classic statistics and machine learning sense. The human's notion of categories is quite flexible: she can adapt new information into her knowledge schema and conversely, adapt the knowledge schema to fit the newly acquired information [29]. Categorization in the machine world, i.e., classification, is in its classic form much more rigid: the class schema is defined beforehand and it cannot be changed without retraining the model from scratch. The meaning and parameters of categorization thus depend on whether it is performed by a human or a machine. In linguistics, the phenomenon of the meaning of words being dependent on context is studied in pragmatics [58]. Hence, we call this phenomenon the pragmatic gap:

- **Pragmatic gap** — the gap between the parameters of a categorization task as performed by the human and the parameters of a categorization task as performed by the machine.

In other words, the pragmatic gap is related to the adaptability of the model as the analyst progresses towards insight. To support the multimedia analyst, the machine models need to be able to mimic her flexible mental model as closely as possible. That involves at least three aspects:

- **New categories on the fly** — For example, when a forensics expert who has initially worked with "child abuse" and "harmless" categories encounters evidence of terrorism, she adapts her cognitive model to include a new "terrorism" category. The machine should be able to do the same.

- **Non-exclusive categories** — If one category is "people," and another is "Rome," then the analyst should not be forced to choose where to put a photo of a couple in front of Fontana di Trevi.

- **Dynamic category semantics** — Suppose an analyst is looking for evidence of arms trafficking and his "suspicious" category contains firearms. Further exploration reveals photos involving explosives, so they are added into the model of the "suspicious" category. Later, the firearms photos are deemed no longer suspicious and removed from the model of the category (e.g., due to the suspect's firearm being properly licensed), which now concerns explosives only. The machine's category model should follow all these steps.

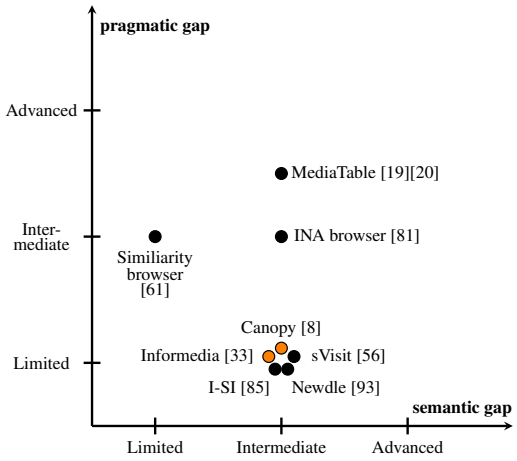

Fig. 8. Comparison of pioneer multimedia analytics systems based on their position with respect to the semantic and pragmatic gaps (axis ticks correspond to system capabilities described in Figure 7). **Orange** points represent systems whose model involves heterogeneous data (visual content, annotations, metadata), **black** points correspond to systems without a heterogeneous model.

The pragmatic gap is orthogonal to the semantic gap: the former's concern is the adaptability of the model, the latter is related to the semantic expressiveness of the model. Figure 7 highlights the impact of the gaps on multimedia analytics models. Bridging the gaps will foster better overall "understanding" between the machine and the human, allowing the machine to create models more closely resembling the analyst's actual model. Bridging the gaps in the context of exploration and search is thus one of the core challenges in multimedia analytics.

4.2 Pioneer systems

To the best of our knowledge, no multimedia analytics system in existence can handle both the semantic and the pragmatic gaps fully. In this section, we review pioneer multimedia analytics systems that have so far paved the way towards narrowing the gaps. Figure 8 depicts the discussed systems in the semantic gap-pragmatic gap space, mapping the current state of the art in multimedia analytics.

One of the first pioneers is the Informedia system used for semantic navigation of the eponymous online digital library, conceived by Hauptmann and Smith in 1995 [32]. Being continuously updated since then, Informedia has received a relevance feedback component in 2008 [33] (albeit one refining search results, rather than maintaining an adaptive model). Another early pioneer system is the similarity manipulation browser by Nguyen et al. [61]. This approach employs a similarity space browser through which the user directly manipulates the similarity space, with the machine recomputing the used similarity and rearranging the items based on the interactions. The last example of an early adopter is the multimedia browser developed for the French Audiovisual Institute (INA) by Viaud et al. [81] This approach combines a similarity space browser, visual summary techniques, and active learning for interactive exploration of the French TV archives.

The field of multimedia analytics has been defined in 2010 by Chinchor et al. [13] and since then, the first systems bearing the multimedia analytics label have started appearing. One group of approaches, including Newdle by Yang et al. [93] and I-SI by Wang et al. [85], targets news and social media, bringing interactive exploration of topic trends in news archives and on social networks. MediaTable by de Rooij et al. [20] facilitates categorization of large image collections using the spreadsheet visual metaphor, making it the first approach in multimedia analytics devoted to categorization. MediaTable has been extended with the active buckets framework maintaining an adaptive model of the data [19]. Meghdadi and Irani adapted the multimedia analytics spirit to the surveillance domain: their sViSIT system allows security experts to search for objects of interest within long video segments and track their trajectory [56]. Canopy by Burtner et al. [8] has a strong focus on integrating content with annotations and metadata, involving the three data sources both in learning and visualization. However, the machine model is not adaptive. These examples show that the field is steadily growing and that multimedia analytics systems are increasingly capable to fill their respective niches. There is no perfect solution which covers all the aspects yet, opening exciting opportunities for multimedia analytics research.

4.3 Research agenda

The previous section concluded that so far, no perfect solution covering both gaps exists. In order to advance multimedia analytics, numerous research questions need to be answered. In this section, we propose several key research questions based on the insight gained from the survey process, establishing a multimedia analytics research agenda:

1. Multimedia visualizations and interfaces
   (a) Are the existing multimedia visualizations and interfaces described in Section 2.2 suitable for categorization as defined in this paper? If not, how can such an interface be created?
   (b) How can the heterogeneous data in multimedia collections be presented in a truly integrated manner?
   (c) What is the best way to present large-scale (>1M items) collections to the user?
   (d) How can multimedia analytics systems be evaluated?
2. Semantic gap
   (a) How can increasingly higher-level semantics be extracted from raw multimedia?
   (b) Can high-level semantics be extracted in a manner not prohibiting an interactive multimedia analytics experience?
   (c) How can the heterogeneous data in multimedia collections be leveraged to improve the semantic quality of the model?
3. Pragmatic gap
   (a) How can the performance and learning speed of active learning be improved, especially in view of the large scale of the data?
   (b) Do the currently used interactions with the model described in Section 3.2 have sufficient information bandwidth for the model to improve? How can they be improved?
   (c) What is the most efficient way to introduce new categories on the fly? Is it attribute-based zero-shot learning [44], or is there a better way?
   (d) What is the best way to introduce non-exclusive categories? Is there a better way than training n 1-vs-all classifiers as done for example in MediaTable [19]?
   (e) What is the best way to model and introduce fully dynamic, human-like categories whose semantics and boundaries evolve over time?

5 Conclusion

The mission of multimedia analytics, facilitating understanding and insight into large-scale multimedia, is very important and ambitious. In this paper, we have surveyed a large body of work related to multimedia analytics and proposed a novel general multimedia analytics model. This model brings two major benefits: it provides a structured overview of all levels of multimedia analytics and establishes a clear agenda for the future of multimedia analytics. Both benefits are based

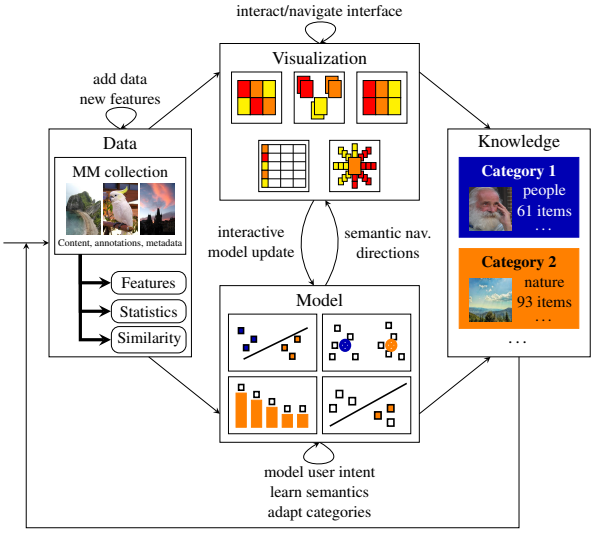

Fig. 9. The proposed multimedia analytics process, expanding upon the diagram by Keim et al. [40][41].

on the extensive survey, therefore grounded in the established scientific theory and bearing in mind the possibilities and limitations of the state-of-the-art techniques in the related fields. This paper thus paves the way towards interactive, intelligent, and integrated multimedia analytics systems of the future, schematically depicted in Figure 9. We believe that those systems will play an increasingly important role in our increasingly digital and multimedia society.

Acknowledgments

The authors thank Jack van Wijk for his insightful comments. This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

References

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *INFOVIS*, pages 111–117, 2005.

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

[3] B. B. Bederson. PhotoMesa: A zoomable image browser using quantum treemaps and bubblemaps. In *UIST*, pages 71–80, 2001.

[4] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 97(2):115–147, 1987.

[5] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, pages 644–651, 2013.

[6] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[7] P. Brivio, M. Tarini, and P. Cignoni. Browsing large image datasets through Voronoi diagrams. *TVCG*, 16(6):1261–1270, 2010.

[8] R. Burtner, S. Bohn, and D. Payne. Interactive visual comparison of multimedia data through type-specific views. In *SPIE VDA*, 2013.

[9] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Annotating collections of photos using hierarchical event and scene models. In *CVPR*, 2008.

[10] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *TPAMI*, 29(3):394–410, 2007.

[11] P. Chandrika and C. V. Jawahar. Multi modal semantic indexing for image retrieval. In *CIVR*, pages 342–349, 2010.

[12] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[13] N. A. Chinchor, J. J. Thomas, P. C. Wong, M. G. Christel, and W. Ribarsky. Multimedia analysis + visual analytics = multimedia analytics. *TCGA*, 30(5):52–60, 2010.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

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

[17] O. de Rooij and M. Worring. Browsing video along multiple threads. *TMM*, 12(2):121–130, 2010.

[18] O. de Rooij and M. Worring. Efficient targeted search using a focus and context video browser. *ACM TOMCCAP*, 8(4):51, 2012.

[19] O. de Rooij and M. Worring. Active bucket categorization for high recall video retrieval. *TMM*, 15(4):898–907, 2013.

[20] O. de Rooij, M. Worring, and J. J. van Wijk. MediaTable: Interactive categorization of multimedia collections. *TCGA*, 30(5):42–51, 2010.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Statistical Society, Ser. B*, 39(1):1–38, 1977.

[22] R. Ewerth, K. Ballafkir, M. Mühling, D. Seiler, and B. Freisleben. Long-term incremental web-supervised learning of visual concepts via random savannas. *TMM*, 14(4):1008–1020, 2012.

[23] J. Fan, Y. Gao, H. Luo, D. A. Keim, and Z. Li. A novel approach to enable semantic and visual image summarization for exploratory image search. In *ACM MIR*, pages 358–365, 2008.

[24] H. Fang, G. K. Tam, R. Borgo, A. J. Aubrey, P. W. Grant, P. L. Rosin, C. Wallraven, D. Cunningham, D. Marshall, and M. Chen. Visualizing natural image statistics. *TVCG*, 19(7):1228–1241, 2013.

[25] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2005.

[26] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[27] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *CVPR*, pages II–1002–II–1009, 2004.

[28] G. Ghinea and J. P. Thomas. Quality of perception: User quality of service in multimedia presentations. *TMM*, 7(4):786–789, 2005.

[29] T. M. Green, W. Ribarsky, and B. Fisher. Building and applying a human cognition model for visual analytics. *SAGE InfoVis*, 8(1):1–13, 2009.

[30] T. L. Griffiths, M. Steyvers, and A. Firl. Google and the mind: Predicting fluency with PageRank. *Psychological Science*, 18(12):1069–1076, 2007.

[31] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[32] A. G. Hauptmann and M. A. Smith. Text, speech and vision for video segmentation: The Informedia project. In *AAAI Symp Compu Mod for Int Lang and Vis*, 1995.

[33] A. G. Hauptmann, J. J. Wang, W. H. Lin, J. Yang, and M. Christel. Efficient search: The Informedia video retrieval system. In *ACM CIVR*, pages 543–544, 2008.

[34] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[35] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. Active learning for interactive multimedia retrieval. *Proc IEEE*, 96(4):648–667, 2008.

[36] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *ACM MIR*, pages 89–98, 2006.

[37] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010.

[38] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos. Scalable active learning for multiclass image classification. *TPAMI*, 34(11):2259–2273, 2012.

[39] S. Kandel, E. Abelson, H. Garcia-Molina, A. Paepcke, and M. Theobald. PhotoSpread: A spreadsheet for managing photos. In *ACM CHI*, 2008.

[40] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010.

[41] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, pages 76–90. Springer, 2008.

[42] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image search with relative attribute feedback. In *CVPR*, pages 2973–2980, 2012.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[44] C. H. Lampert, H. Nickish, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[45] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[46] Y. LeCun and Y. Bengio. Convolutional networks for images, speech and time series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. The MIT Press, 1995.

[47] M. Li and I. K. Sethi. Confidence-based active learning. *TPAMI*, 28(8):1251–1261, 2006.

[48] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *TMM*, 11(7):1–14, 2009.

[49] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Fusing concept detection and geo context for visual search. In *ACM ICMR*, 2012.

[50] Z. Li, E. Gavves, K. E. A. van de Sande, C. G. M. Snoek, and A. W. M. Smeulders. Codemaps: Segment, classify and search objects locally. In *IEEE International Conference on Computer Vision*, pages 2136–2143, 2010.

[51] H. Liu, X. Xie, X. Tang, Z. W. Li, and W. Y. Ma. Effective browsing of web image search results. In *ACM MIR*, pages 84–90, 2004.

[52] S. P. Lloyd. Least squares quantization in PCM. *TIT*, 28(2):129–137, 1982.

[53] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[54] H. Luo, J. Fan, J. Yang, W. Ribarsky, and S. Satoh. Analyzing large-scale news video databases to support knowledge visualization and intuitive retrieval. In *VAST*, pages 107–114, 2007.

[55] G. Marchionini. Exploratory search: From finding to understanding. *Comm ACM*, 49(6):41–46, 2006.

[56] A. H. Meghdadi and P. Irani. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *TVCG*, 19(12):2119–2128, 2013.

[57] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *to appear in ACM Computing Surveys*, 2014.

[58] J. L. Mey. *Pragmatics: An Introduction*. Oxford: Blackwell, 2nd edition, 2001.

[59] P. Mitra, C. A. Murthy, and S. K. Pal. A probabilistic active support vector learning algorithm. *TPAMI*, 26(3):413–418, 2004.

[60] G. P. Nguyen and M. Worring. Interactive access to large image collections using similarity-based visualization. *J Vis Lang and Comp*, 19(2):203–224, 2008.

[61] G. P. Nguyen, M. Worring, and A. W. M. Smeulders. Interactive search by direct manipulation of dissimilarity space. *TMM*, 9(7):1404–1415, 2007.

[62] C. North. Towards measuring visualization insight. *TCGA*, 26(3):6–9, 2006.

[63] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int J Comp Vis*, 42(3):145–175, 2001.

[64] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In *CVPR*, pages 3384–3391, 2010.

[65] W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The science of interaction. *SAGE InfoVis*, 8(4):263–274, 2009.

[66] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Int Conf Intel Analysis*, 2005.

[67] Z. Pousman, J. T. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *TVCG*, 13(6):1145–1152, 2007.

[68] N. Quadrianto, K. Kersting, T. Tuytelaars, and W. L. Buntine. Beyond 2D-grids: A dependence maximization view on image browsing. In *ACM MIR*, pages 339–348, 2010.

[69] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Does organisation by similarity assist image browsing? In *ACM CHI*, pages 190–197, 2001.

[70] S. Santini and R. Jain. Similarity measures. *TPAMI*, 21(9):871–883, 1999.

[71] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[72] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.

[73] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380, 2000.

[74] C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *Computer*, 43(6):76–78, 2010.

[75] C. G. M. Snoek and M. Worring. Concept-based video retrieval. In *Foundations and Trends in Information Retrieval*, 2009.

[76] R. Szeliski, N. Snavely, and S. M. Seitz. Navigating the worldwide community of photos. *ACM TOMCCAP*, 9(1s):47:1–47:4, 2013.

[77] Y. Tang, R. Salakhutdinov, and G. E. Hinton. Robust Boltzmann machines for recognition and denoising. In *CVPR*, pages 2264–2271, 2012.

[78] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Computer Society, 2005.

[79] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM MM*, pages 107–118, 2001.

[80] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010.

[81] M. L. Viaud, J. Thièvre, H. Goëau, A. Saulnier, and O. Buisson. Interactive components for visual exploration of multimedia archives. In *ACM CIVR*, pages 609–616, 2008.

[82] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*, pages 2262–2269, 2009.

[83] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, pages 3035–3042, 2010.

[84] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, pages 1903–1910, 2009.

[85] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-SI: Scalable architecture for analyzing latent topical-level information from social media data. *Comp Graph For*, 31(3):1275–1284, 2012.

[86] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma. AnnoSearch: Image auto-annotation by search. In *CVPR*, pages 1483–1490, 2006.

[87] R. C. F. Wong and C. H. C. Leung. Automatic semantic annotation of real-world web images. *TPAMI*, 30(11):1933–1944, 2008.

[88] M. Worring. Easy categorization of large image collections by automatic analysis and information visualization. In *Class & Vis Int UDC Sem*, pages 235–242, 2013.

[89] M. Worring, A. Engl, and C. Smeria. A multimedia analytics framework for browsing image collections in digital forensics. In *ACM MM*, pages 289–298, 2012.

[90] M. Worring and D. C. Koelma. Multimedia pivot tables. In *VAST*, 2013.

[91] Y. Wu, E. Y. Chang, and B. L. Tseng. Multimodal metadata fusion using causal strength. In *ACM MM*, pages 872–881, 2005.

[92] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, and W. Ribarsky. Semantic image browser: Bridging information visualization with automated intelligent image analysis. In *VAST*, pages 191–198, 2006.

[93] J. Yang, D. Luo, and Y. Liu. Newdle: Interactive visual exploration of large online news collections. *TCGA*, 30(5):32–41, 2010.

[94] J. S. Yi, Y. Kang, J. T. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *TVCG*, 13(6):1224–1231, 2007.

[95] J. S. Yi, Y. Kang, J. T. Stasko, and J. A. Jacko. Understanding and characterizing insights: How do people gain insights using information visualization? In *ACM BELIV 2008*, 2008.

[96] J. Zahálka. Multimedia analytics article library. `http://staff.fnwi.uva.nl/j.zahalka/maal.html`.

[97] E. Zavesky, S. F. Chang, and C. C. Yang. Visual islands: Intuitive browsing of visual search results. In *ACM CIVR*, pages 617–626, 2008.

[98] L. Zhang and Y. Rui. Image search — from thousands to billions in 20 years. *ACM TOMCCAP*, 9(1s), 2013.

[99] X. S. Zhou and T. S. Huang. Small sample learning during multimedia retrieval using BiasMap. In *CVPR*, pages I–11–I–17, 2001.

[100] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.