A Brief Overview of Interactive Tools for Understanding and Effective Manipulation of Multimedia Data

Hemanchal Joshi and Yasushi Akiyama*†

Department of Mathematics and Computing Science, Saint Mary's University, 923 Robie Street, Halifax, B3H 3C3, Nova Scotia, Canada.

*Corresponding author(s). E-mail(s): Yasushi.Akiyama@smu.ca;
Contributing authors: Hemanchal.Joshi@smu.ca;
†These authors contributed equally to this work.

Abstract

With the rapidly increasing use of devices that can easily create, store, and share multimedia data such as pictures, videos, and sound, we have now accumulated an enormous amount of multimedia data. While data mining approaches may provide ways to filter and categorize collective data effectively, understanding individual multimedia content can still be tedious and time-consuming: common ways to understand individual multimedia content are to listen to the sound, view images, or playback video and animation. What types of interaction techniques are available for which types of multimedia data, and can any of these techniques be applied across different domains? To answer the first part of this question, in this article, we surveyed over hundred publications of past and current research on various types of interactive multimedia tools and approaches that assist user interactions with multimedia data and summarized eighty-one prevalent tools. Naturally, such a list of research cannot, of course, be exhaustive, especially given the limit of pages, thus this paper intends to provide an overview of the status of interactive tools for multimedia tasks in the related study areas. We hope that this paper will serve as a hub for researchers of such multimedia tools for finding related and similar tools so that they can leverage the findings and efforts of other researchers to move our knowledge and techniques forward and across typical boundaries of multimedia research domains.

1 Introduction

With the wide usage of mobile device technologies that enable easy ways to create, store, and share multimedia data such as pictures, videos, and sound, there is now an accumulation of an enormous amount of multimedia data to be handled. While approaches that have been developed in such research areas as data mining play an important role in understanding collective data, especially for narrowing down the search space [1], in the case of multimedia data, more assistance is needed to understand and manipulate the content of individual multimedia data files.

Common ways to understand individual multimedia content are to listen to the sound, view images, or play back video and animation, but this process can be tedious and time-consuming. One encounters this issue in different types of tasks, for example, when one tries to select several pictures from a few hundred candidate pictures to be used for certain multimedia tasks. Typically, one would manually go through thumbnails of these pictures while capitalizing on certain heuristic filtering criteria such as dates and locations. A similar task with video data (e.g., a videographer is trying to find particular video segments to be used for a documentary film) can exponentially increase the time spent searching through the collection, as the thumbnail view for videos typically only shows a single video frame that may or may not be the most relevant frame to represent the video content. In these cases, one would need to open each candidate video to play it back and search its content to find relevant sections of the video. Similar and related issues can occur when navigating through a music library or video streaming services.

Further, in multimedia tasks, after identifying the target media from a large collection, one would typically need to adjust a set of parameter values of the selected data in order to make appropriate modifications to the data to be used in the multimedia projects. That is, users do not only need the proper information to be displayed, but they also need to be able to manipulate such information in order to understand and then modify the underlying data. The types of information and interactions that users need to perform these data manipulations can vary. What is the current status of interactive multimedia tools that help users to perform these tasks?

As an initial step toward answering this question, we surveyed over hundred publications of past and current research on various types of interactive multimedia tools and approaches that assist user interactions with multimedia data and summarized eighty-one prevalent tools. Naturally, such a list of research cannot, of course, be exhaustive, thus this paper intends to provide an overview of the current status of interactive tools for multimedia tasks in the related study areas. While there are myriad ways to categorize these tools, we organize the rest of the paper based on the types of general task goals/purposes of these interactive multimedia tools; information/data property retrieval (Section 2), media content search (Section 3), data annotation and labelling (Section 4), and multimedia data visualization (Section 5). Note

that there are understandably some overlaps (e.g., a tool may support interactions for both information retrieval and annotation) as well as certain types of interdependencies between these categories (e.g., information retrieval before visualizing), but the organization was done based on the tools' preeminent focuses of their target multimedia tasks.

2 Interactive Information and Data Property Retrieval

One of the fundamental tasks in dealing with multimedia data is to extract relevant information from raw target media. The retrieved information can then be used in other tasks such as searching in data collections (Section 3), annotating and describing the data (Section 4), and visualizing them in a variety of ways (Section 5). While there are many different approaches to multimedia information retrieval, as in the underlying theme of this paper, this section focuses on approaches and tools that allow for *interactive* multimedia information retrieval. That is, rather than relying on the fully automated approach to information retrieval, the approaches discussed in this section are examples of those that involve human interactions in the retrieval processing loop so as to enhance retrieval performance in certain ways.

Image and video property information such as basic colour correction properties (e.g., colour channels, brightness, contrast, exposure, hue, saturation, opacity) is often trivial and retrieved automatically (i.e., without much user interaction) and available for viewing and manipulation in most popular commercial image and video editing software.In addition to these basic types of image information, advances have been made especially in the study area of computer vision that focuses on image feature extraction for such purposes as content-based image search and object recognition/detection [2, 3]. Image feature extraction has several application domains, one of which is to create semantically meaningful segments in images and videos. Some earlier approaches to object boundary detection used methods with low computational complexity such as minimum spanning tree searching (e.g., [4]) and clustering of similar frames (e.g., [5]). The clustering itself is then often complemented by user interactions to connect the extracted information and the multimedia data (or correct the automatically retrieved information). Digital Fishtank [6] was one of such earlier tools, which allowed for storing the segmented image to create a multimedia database. After the extraction of the object from the video frame, the system allowed the user to manipulate the object and its attributes to edit the existing content.

One of the most common interaction techniques for video segmentation is to specify target objects and/or regions in an image by drawing or scribbling on one of the video frames. For example, the system developed by Giró-i-Nieto and Martos [7]) allows users to draw a boundary around an object. The bounding box will then be expanded to generate an object mask, which triggers an object-tracking algorithm. This process is complemented by an annotating tool

to manipulate the image from an object frame. In a geodesic framework [8], the users scribble on a target object with one colour and around it in another colour. The system then performs the optimal, linear-time computation of weighted geodesic distances to those scribbles to detect the object boundary and segment it from the rest of the image. LIVEcut [9] is an interactive video segmentation tool that implements a graph-cut optimization framework [10]. A set of visual properties of the object (colour, gradient, adjacent colour relationships, spatiotemporal coherence, and motion) are first gleaned from video frames and these properties are then locally weighted in order to perform the graph-cut optimization. The tool then allows the user to correct errors by adjusting the weights and learns from it to optimize its segmentation performance. Grabcut [11] is another graph-cut-based segmentation tool, which allows its users to draw scribbles on (shown as the red lines in Fig. 1) and around the object (shown as the blue lines), resulting in a rectangular box drawn around the object (shown as the green rectangle) to indicate the estimated foreground area (i.e., the area of user's interest). It then slices the frame by the grab-cut operator after estimating the probable foreground region and rectifies the interested region by making sure the object lies within the green rectangle. There are also approaches to accomplish the segmentation even with fewer strokes; Shankar et al. [12] proposed an approach that only requires a single disjoint scribble on the specified object, and uses a combination of motion from point trajectories and integrates a constraint to enforce colour consistency.

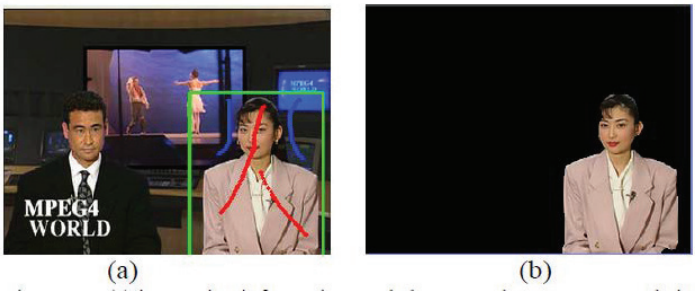

Fig. 1 A screenshot of Grabcut: (a) User Scribbles on the object and boundary (b) segmented object (Image reproduced from [11]).

These interactions to specify objects are often followed by one or more user interactions to optimize the segmentation performance. Fomtrace [13] uses a delineation algorithm derived from Optimum-Path Forest (OPF) [14] to trace a fuzzy object model based on the input mask layer created by tracking the scribbles. Another tool proposed by Levinkov et al. [15] uses the supervoxel-based multi-cut approach, which allows the object with consistent spatiotemporal boundaries to be segmented simultaneously. An interactive image segmentation method by Yang et al. [16] capitalizes on a random walk algorithm [17],

which estimates the constraints around the three types of user inputs (foreground vs. background, specification of region boundaries, and indication of pixels falling on the boundaries) and selects the matching frames.

Some recent approaches have now started incorporating machine-learning algorithms in an attempt to improve the performance of interactive video segmentation and Deep Extreme Cut (DEXTR) [18] is one such approach. Users specify a bounding box with mouse clicks on the extreme endpoints of objects and a 2D Gaussian is centralized around each of the annotated endpoints in order to create a heat-map, which is then concatenated with the standard RGB input values to form a 4-channel input for a convolutional neural network (CNN). The output of the CNN is a probability map of pixels indicating whether each pixel belongs to the object of interest. In Mindcamera [19], the users draw a rough outline of an object and the system uses alpha matting to allocate real-time foreground object extraction to return the scene images consisting of the potential objects. It uses contour extraction to filter the backgrounds and Gradient Field HOG (GF-HOG) [20] to add spatial information to Bag of Visual Word (BoVW) as description, in conjunction with YOLO [21], a deep-learning-based object detection algorithm.

Another interactive tool [22] uses a template matching algorithm in order to detect relevant video segments and create summary videos of youth and amateur soccer games. Many amateur-level games and practices are typically recorded with a single camera, and it is often difficult to employ existing approaches (e.g., [23–25]) that capitalize on multiple camera shots or the spectators' audio stream to detect video scenes. The user initially selects a set of reference areas by using bounding boxes in a video frame and the tool uses the reference areas as search keys for relevant frames. The search results are shown as a graph indicating the likelihood of each frame potentially including those reference areas (shown in Figure 2). The user will then use these results as a guide to finding relevant video scenes.

This section discussed multimedia tools that incorporated user interactions that are provided to extract relevant information from target media, in particular from images and videos. These interactions typically allow for initial feature retrievals, followed by other interactions (e.g., feedback to correct results, annotation) and for certain tools, enhanced by the use of machine learning techniques. Table 1 summarizes the tools discussed in this section and Figure 3 shows the corresponding publications in chronological order.

3 Interactive Media Content Search

With myriad ways to produce large amounts of multimedia data, especially with the advancement of, and the easy access to, different types of mobile devices that allow us to create high-resolution multimedia content, we are often faced with issues of finding relevant data, especially in large data collections. In this section, we discuss various approaches to interactive search for multimedia content.

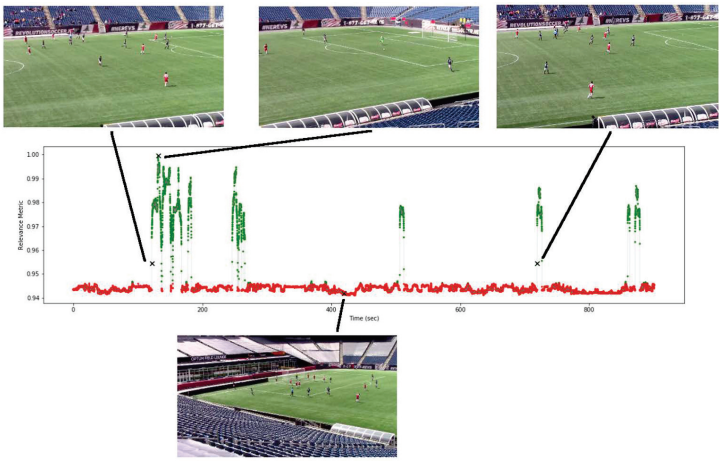

Fig. 2 The tool displays the resulting likelihood of each frame containing a reference image, a soccer goal in this example. [22]

Table 1 Summary of the tools for interactive information retrieval (discussed in Section 2).

Domains/Tasks	Algorithms/Interactions: Tools
Image feature extraction	Minimum spanning tree search, clustering of similar frames, user interactions to connect the extracted information and the multimedia data: [6]
	Scribble/drawing: [7–9, 11, 12]
	Optimization of extraction: [13] (OPF), [15] (supervoxel-based multi-cut), [16] (a random walk), [18] (CNN), [19] (BoVW, YOLO)
	Template Matching: [22]

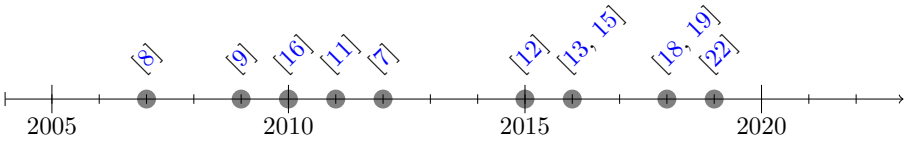

Fig. 3 Timeline of the tools for interactive information retrieval (discussed in Section 2).

3.1 Images and Video Information Retrieval for Interactive Search

There are two fundamental approaches that are often mentioned when discussing multimedia search algorithms. One of them is content-based image retrieval (CBIR), which is a set of techniques that analyze the *contents* of the data (e.g., colours, shapes, textures, moods/emotions) and uses the retrieved content information for search algorithms [26]. In contrast, the other approach,

perhaps a more traditional one, is *concept*-based image retrieval, which capitalizes on common metadata such as data creation dates, keywords, or general descriptions of the medium for their search mechanisms. One of the earlier examples of interactive multimedia search systems based on the CBIR used a relevance feedback approach for image retrieval [27], which allowed the users to interactively give feedback on the search results and the system then modifies its internal search algorithms based on this feedback in such a way as to improve future search results. In terms of retrieval of contents, the signature-based search algorithm [28] is a common technique often employed for the CBIR search, in which, a set of feature signatures such as those based on the colours are extracted from select key frames [29]. Tools based on this technique provide a set of interactions such as allowing users to sketch with simple coloured circles specifying the region of semantically similar contents [30, 31]. For example, NII-UIT [32–34] is a tool that uses the sketch-based search key entry for a known scene in a video database. This interactive search system employs a multi-modal search by describing a known scene by its both visual and audio cues. Users specify the input search key by sketching non-rigid shapes with colours in addition to the audio search capability based on types of sound or audio genres.

Emotional semantics can also be used to execute a query and navigate through the video contents both within a single video or in a collection. For example, Yoo and Cho introduced an interactive video scene retrieval tool [35] (shown in Fig. 4) that uses the genetic algorithm (GA) [36]. This tool implements a video scene retrieval algorithm based on human emotions annotated by the users. After detecting a video frame, the video generates a set of chromosomes which represent video image features such as average colour histogram, brightness, shot duration, and change rate. The extracted chromosomes are then mapped through GA which analyses the emotional space that a user thinks of. Once trained, the system can filter videos according to the mappings between the video features and the emotion criteria.

Most CBIR approaches are often used to complement the traditional concept-based image retrieval in order to enhance the search performance by adding extra dimensions to search algorithms and results filtering. A web-based interactive video contents navigation tool, proposed by Joly and Tjondronegoro (e.g., [37]), is one such example tool to take advantage of both the approaches; videos are first indexed by an MPEG-7 specification-based method [38], and the users can interactively narrow down the search space by providing two types of queries (domain and media queries). The search results are displayed as an interactive page that organizes the information in a hierarchical structure, which allows the users to investigate the results at different abstraction levels. A common interaction technique often used with hierarchical structures in the interactive data search is to allow changes in the data granularity (e.g., zooming in and out). Zoom Slider Interface [39], for example, is a tool that allows users to view the data at various granularity levels. The

Fig. 4 The interface of a video scene retrieval tool that uses the genetic algorithm to detect the emotions associated with the video content (image reproduced from [35].)

users select a medium file as an anchor point and move the slider to zoom into all the related files in the collection by narrowing down the search space.

As in the case of image object detection (2), various image processing techniques such as those developed in computer visions have been proposed and advanced for content retrieval and search of multimedia data. Schoeffmann et al. [40] developed a video browsing tool using motion visualization, which utilizes motion vector information contained in H.264/AVC bit streams and visualizes a set of video motion properties (amount, direction, and speed of motions of a video scene). Based on these properties, the tool creates an interactive navigation index, which, in turn, helps users better understand the content semantics of video scenes. This type of interactive tool can allow users to quickly audition possible media data in the collection as their semantically tagged contents along with their abstract portrayal allows the user to click and quickly preview the target media, and then navigate through the video content. Zhang et al [41] proposed a system that employs the concept filters and faceted navigation approach to allow users to find a video of their interests in a large collection; video attributes such as semantic concepts (e.g., indoor/outdoor, landscape) and object labels (e.g., vehicle, building, faces) are automatically extracted from the video to filter the search results. The faceted navigation is accomplished by first building a set of facets from different object labels or scene categories that are extracted from the video content, and the associated facets are then selected based on the initial text search query so that the users can explore related search spaces adjacent or overlapped with the main query results. A faceted navigation approach was also used in LifeSeeker, Interactive Lifelog Search Engine [42], in addition to the query expansion for solving

certain lexical gaps between the users and the tool. Since the platform is developed for lifelog search purposes, it is easy to see how this approach can be used for more generic multimedia data search purposes as a lifelog is indeed a collection of multimedia data.

The recent trend to employ machine learning (ML) techniques has also been observed in their applications for multimedia content search. A multimodal search tool, VIRET [43, 44], uses deep neural networks for known-item search and video retrieval. Shot transaction and frame selection methods [45] determine the structure of the input videos. In this tool, a deep 3D convolutional neural network (CNN) called TransNet is trained for shot transition detection in down-scaled videos. Once shots are detected, frames will be selected based on a clustering-based method considering the similarity of frames and temporal ordering, and the selected frames' thumbnails and descriptors will be generated. An input query from the user is used as a keyword for the relevance score function in order to rank the top results among which the user can interactively select a threshold to retrieve prominent results.

Verge [46–49] is another video content search tool that integrates multimodal indexing and retrieval modules using Instance Search (INS) [50] and Ad-hoc Video Search (AVS) tasks [51]. As the tool evolved, new approaches to video clustering and categorization for effective browsing are added to the newest version of Verge [48] (shown in Fig. 5).

Fig. 5 A screenshot of Verge (image reproduced from [48].

Conceptual categorization of multimedia data helps users to easily understand and identify multimedia content. Primus et al. [52] presented a collaborative video search system as a part of the Video Browser Showdown (VBS) competitions [53]. In this system, the video contents are grouped and categorized based on the similarity in the video semantics and displayed in an interactive video inspector view, which consists of a grid of video frames. When a user runs a query, a pivot table based on the nearest neighbour algorithm

is exploited for the fast search to find the heuristically best-resembling target videos in the database. This type of conceptual categorization approach to multimedia data can also be applied in many different domains. For example, Cai et al. [54] developed a tool for medical decision-making. It is used to assist the users in identifying patterns of similar images. This tool reduces the time to cluster the correlating images together and potentially reduces certain types of mistakes that could potentially lead to an incorrect medical diagnosis, which, in turn, may have (often negative) repercussions.

3.2 Music and Sounds Information Retrieval for Interactive Search

Unlike images and video data that contain inherent visual components, humans typically rely heavily on their auditory system to examine the content of sound data. As we now have enormous amounts of digital audio data such as those provided by music streaming services, it is almost impossible and extremely inefficient for humans to find and understand these audio data just by listening to them. To help understand the audio content, many studies have been conducted to provide users with additional (visual, textual, etc) information that can complement the audio data. Other than the common metadata (e.g., mp3 ID3 tags), various research for browsing through such large audio data collections has resulted in many different approaches for browsing and viewing the audio data. Sonic Browser [55, 56] which maps the sound into a two-dimensional or a two-and-a-half-dimensional representations. Certain audio properties may be assigned to the visual properties of this tool (e.g., file size mapped to the size of visual objects, sampling rates to colour, object coordinate/location to file creation time). The users hover over the visual objects representing audio data, and audition one or more audio files simultaneously, enabled by changing the pointer size to hover over more than one data point. This hover and listen technique is also used in other interactive audio visualization tools such as MUSESCAPE [57] that displays audio files in a graphical plot. The tool provides an automatic configuration mechanism based on computer audition techniques and the use of continuous audio-music feedback.

While the above approaches are typically based on the common metadata and relatively crude audio attributes, there are other tools that focus more on the semantic information of the audio data, and use this information for organizing the data. These approaches often employ more sophisticated audio analysis techniques such as clustering of audio data based on its semantics and a mechanism to integrate interactive user feedback. For example, MusicSim [58] is an interactive UI for a large-scale music data visualization that capitalizes on the user feedback (i.e., manipulation of the automatically created clusters) and audio content analysis techniques to organize a music library. Another system called Interactive Music Archive Access [59] uses the optimized version of the source separation algorithm [60] to extract the audio source address. The source address is the unique position in the stereo file identified by a frequency

bin carrying the local minima in an "azimuth frequency" during a single time frame. Other common music attributes such as meter, time signature, key signature, and tempo are also retrieved in this tool, allowing the users to view and manipulate the audio data in several ways (e.g., isolation of individual instruments in stereo mixes, pitch modification, and time-scale modification, and beat-synchronous looping). Songrium [61, 62] (shown in Fig. 6) stores metadata (e.g., frequency, artist, length) of various songs and visualizes them in a graphical representation based on the *relationship* between the songs. The graph allows the users to understand the various relations between web-based music and the existence of any derivative music such as a cover song of an original song. For extraction of the relationships, it mines and synthesizes data from Niconico which is a Japanese video communication service, by analyzing the semantics that the songs carry and defining the distance between them based on the proximity of their semantics.

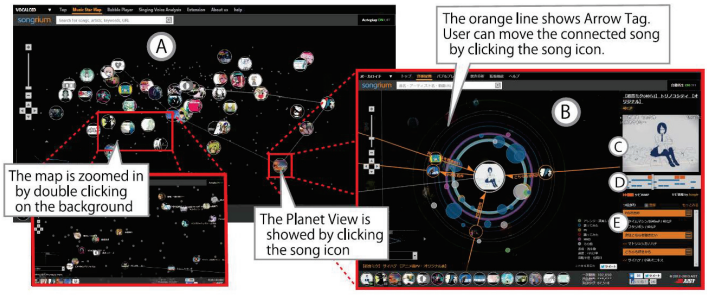

Fig. 6 A screenshot of Songrium (image reproduced from [61]).

Capitalizing on the additional retrieved information such as those described above, some tools provide effective audio data search capabilities. For example, Dunya [63] is a web-based tool that allows users to interact with an audio music collection through the use of musical concepts that are derived from the Carnatic music culture. The tool allows users to execute a query and the search results are obtained based on the relationships among the audio files in the database, or by using culturally relevant similarity measures based on the parameters related to melodic and rhythmic patterns (e.g., raga, tala, sur) as well as common metadata (e.g., artists, concert pieces). Another search tool called MyMoodPlay [64] uses human emotions as a search parameter. Using semantic web technology, the tool presents the *mood* of the audio data based on cognitive/psychological analysis and uses emotions as coordinates in the arousal-valence (AV) space. The emotions then categorize the audio files based on the extracted mood and also use the user's interaction history as feedback to improve the relevance. The users browse the data using the pre-defined query mood coordinates as a cue and by clicking on the mood label to visualize the filtered results. With Moodydb [65], an online MIR system for searching and

browsing music by mood, Hu et al. experimented with a unique visual-physical interaction, which is the eye movement pattern pointing as the Prominent Area of Interest. The semantics related to the specific eye movement were used to clustering of the audio files. Certain eye movement patterns during music retrieval tasks (e.g., search, listening, navigating etc.) are recorded by sensors and are used to analyze the user interactions with the music database.

This section discussed multimedia tools that allow for interactive media content search. Two of the most common approaches were content-based image retrieval (CBIR), which utilizes media content such as colours, shapes, textures, moods/emotions, and concept-based image retrieval, which capitalizes on common metadata such as data creation dates, keywords, or general descriptions of the medium for their search mechanisms. The approaches are also found to be effective for sound/music data. And in both the data types, user feedback is often used to correct and improve the search results, and/or supported by certain machine learning algorithms. Table 2 summarizes the tools discussed in this section and Figure 7 shows the corresponding publications in chronological order.

Table 2 Summary of the tools for interactive media content search (discussed in Section 3).

Domains/Tasks	**Algorithms/Interactions: Tools**
Images and Video (Section 3.1)	Content-based image retrieval (CBIR): [27, 29–35]
	Concept-based image retrieval: [37, 52, 54]
	Computer-vision-based approaches: [40] (motion visualization), [41] (concept filters/-faceted navigation), [42] (faceted navigation), [43, 44] (machine learning)
	Multimodal indexing, INS, AVS: [46–49]
Music and Sounds (Section 3.2)	Based on metadata: [55–57]
	Based on semantics: [58, 59, 61–65]

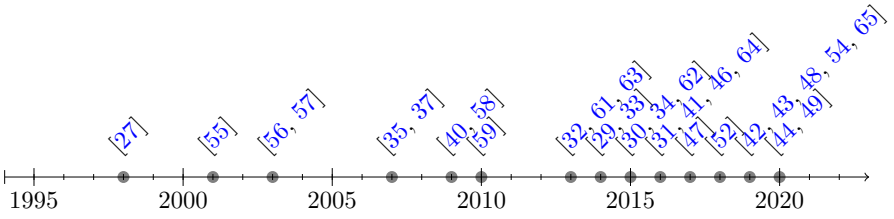

Fig. 7 Timeline of the tools for interactive media content search (discussed in Section 3).

4 Interactive Data Annotation and Labelling

Integrating additional information that is not readily available in raw multimedia data can enhance users' understanding of the data. While some information may automatically be retrieved and estimated as discussed in Section 2, the accuracy and usefulness of such automated estimators still vary, especially when used in more generic domains [66, 67]. Approaches to allowing users to intervene in the automated processes and interactively annotate or label data have thus been investigated [68], and in this section, we discuss interactive annotation and labelling approaches for multimedia data.

Some of the earlier work on interactive data annotation and labelling tools date back to the late 80s and early 90s [69, 70] when researchers started deviating toward constructive approaches for increased interactivity from the older limited interaction techniques. An interactive watermarking tool [71], for example, provides an environment that retrieves various label information such as copyright information, customer information or any additional metadata embedded in the original video files, then allows the watermark insertion of these labels. The watermark is usually invisible in the image or video stream itself but can be made visible when the copyright owners need to access the relevant information. For the retrieval to be possible, the video initially should be annotated with specially instructed copyright information and metadata that supports the retrieval. The interaction occurs with a mouse click when the labels are being embedded using an amplitude-modulation (AM)-based and a discrete cosine transform-based algorithm [72, 73]. An authorized user can easily retrieve the image and compare the images through these visible watermarks. Similarly, I^2A [74] is an interactive image annotation tool derived based on statistical modelling. The annotation of the data is done in 2 steps. First, unlabelled images are clustered based on the low-level image feature extraction and their semantic relationships are hierarchically structured benefiting from the use of WordNet, a lexical database of English [75]. These clusters are then annotated from the training data using the semantic expressions of the nearest cluster. The system will next allow for user interaction by accepting the query from the users either by examples or keywords, and the returned results are validated by the user and refined based on the user feedback.

LabelMe [76], a web-based image annotation tool, went through a few evolutions of its features. Initially, the tool was built to read images from publicly available databases of visual objects and allow users to manually add annotations for objects (i.e., draw boundaries then add labels) in the image, as shown in Fig. 8. The users are also allowed to correct any errors in the existing labels (e.g., redraw object boundaries or edit attached labels). These annotations can then be used to train a machine learning model to automatically detect and label objects in unseen images. The authors then extended to include the capability of video annotation [77], and further added the feature of 3-D reconstruction of scenes from 2-D images by capitalizing on certain information extracted from the annotations of the visual objects [78].

Fig. 8 User Interface of LabelMe (image reproduced from [76])

There are several areas in which the image and video annotation approaches are applied. For example, video surveillance is one of those popular application areas of video annotation approaches; ViSOR is a framework for collecting, annotating, retrieving, and sharing surveillance videos [79]. The system allows the uploading and downloading of surveillance videos and annotations. Its annotation tool allows users to draw points, bounding boxes and oriented rectangles, ellipses, polygons and circles to indicate areas of interest and attach annotations to them. The annotations may include the context of the video (e.g., indoor vs outdoor, traffic surveillance, etc) and its content (e.g., building, person, car), which can be either physical objects or actions/events. Another interactive scene annotation tool for video surveillance [80] employs a three-stage procedure for camera calibration to annotate scenes from videos captured by pin-hole cameras. In the first stage, the users specify vertical lines, which will be used to estimate the vanishing point and its corresponding horizon line. The users will then interactively refine the estimation in the second stage, by allowing them to add more vertical and horizontal lines. In the final stage, the users can label major surfaces such as a ground plane, walls, and other surfaces constraining objects' activity (e.g., stairs).

In the context of learning activities, Kalboussi et al. [81] proposed a web-browser-based interactive annotation tool. It allows the users (i.e., the learners) to make annotations on a web page and the annotations are stored based on their ontological implications. The goal of the annotation is then interpreted and used to invoke a search query of web services. The results of this query are filtered and the subset of the results that are only relevant to the learner's annotation is returned to the user. The annotation itself is done by a popular annotation plug-in called Annozila. Temporal Summary Images (TSIs) [82] is an approach to exploring and annotating complex, multidimensional, and time-varying data. It employs comic strip-style data snapshots and textual

annotations that are superimposed on the data visualization. The purpose of this tool is to provide ways to allow for narrative visualization and data storytelling. Once the user chooses the base visualization with the previously saved annotations, they can initiate the semi-automatic annotation process by first selecting the data point of interest and its attributes, which will generate the data-driven annotations. These automatically generated annotations can then be edited by direct interactions with them like dragging, deleting, and filtering, as well as modifying the underlying algorithm's annotation suggestions. ECAT (Endoscopic Concept Annotation Tool) [83] is a web-based tool for surgical video frame annotation. It allows the user to cluster video frame images by directing them into folders tagged with a similar concept. The tool consists of two parts, predefined taxonomical labels on the left-hand side of the UI in a hierarchical structure, and a set of images on a grid view on the right-hand side of the UI. The user either selects a set of correlated images and chooses a label that defines all selected images in unison, or clicks to select a label and view all the images related semantically.

While the annotated data such as these discussed above are typically used to generate labelled datasets for AI and machine learning model training, there are also attempts to enable semi-supervised/semi-automatic annotations that integrate computer vision techniques and machine learning algorithms for labelling images and videos. IGAnn's [84] approach employs semi-supervised clustering that constructs a hierarchical classifier of images based on the user's relevance feedback. This approach allows unlabelled/unannotated images to be interactively annotated by using only a few labelled images at the initial iterations. iVAT [85]) also provides both semi- and fully-automatic annotation modes in addition to the more traditional manual annotation mode (i.e., by specifying object boundaries and attaching a label). The semi-automatic annotation will be triggered once an object in a video frame is completed, by linear tracking of that object or spatially-constrained template matching throughout the frames. Automatic mode is initiated by users choosing a label from the list and selecting a preferred supervised object-detection model (a cascade of boosted classifiers working with Haar-like features, Histograms of Oriented Gradients (HOG) features, and Local Binary Patterns (LBP) features). Another automatic interactive video authoring tool proposed by Yoon et al. [86] uses Faster R-CNN based object recognition [87] and natural language processing (NLP) based keyword extraction techniques for the data annotation. The tool first detects shots/video segments and objects in the video, it then extracts the metadata from the detected objects. The extracted metadata is linked to the object-related classes (i.e., objects and annotations) and the video-related class (the shot that includes the objects). The tool then performs the image search on Google to related textual information attached to the images found based on the detected objects as an image search key. VIAN [88] is a visual annotation tool for a film analysis that supports deep learning-driven segmentation for spatially aware colour analysis. It refines the

segmentation by further categorizing the manually incorporated key semantics of the video frame (shown in Fig. 9).

This section discussed multimedia tools for interactive data annotation. Over the past decades since the earlier implementations of interactive annotation tools were introduced, approaches of data annotation matured to include the interactive feedback process to correct or complement automatic object recognition algorithms and supplemented by certain machine learning algorithms to enhance the annotation processes. This trend of integration of machine learning is seen in all these multimedia domains, and it is most likely to continue. The annotated data will then be used as training sets for machine learning modelling processes, even furthering the improvements of these approaches. Table 3 summarizes the tools discussed in this section and Figure 10 shows the corresponding publications in chronological order.

Table 3 Summary of the tools for interactive data annotation (discussed in Section 4).

Domains/Tasks	Algorithms/Interactions: Tools
Generic image annotation	Embedded with AM-based and a discrete cosine transform-based algorithm: [71]
	Statistical models to associate image features and semantics, then refined by relevance feedback: [74]
	ML model to automatically detect and label objects: [76–78, 84–86, 88]
Video surveillance	Indicate areas (shapes, lines) and attach annotations: [79, 80]
Learning activities	Web page annotations based on ontological implications: [81]
Data storytelling	Comic strip-style data snapshots and textual annotations for complex, multidimensional, and time-varying data: [82]
Surgical video	Cluster video frames with a similar concept and label the cluster: [83]

5 Multimedia Data Visualization

After extracting relevant information from multimedia data and attaching relevant annotations and labels, we need to address in what ways such information can be displayed and what types of interactions with it are to be provided so that humans can utilize the information to gain certain new knowledge that is otherwise not easily acquirable. There are typically multiple ways to visualize the same information given the data and the corresponding information, and some approaches may often be more suitable for certain tasks, purposes, and

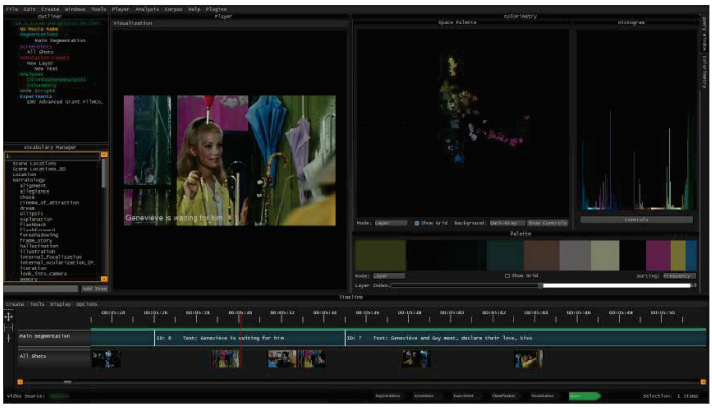

Fig. 9 The user interface of VIAN, an interactive visual annotation tool for film analysis (image reproduced from [88]).

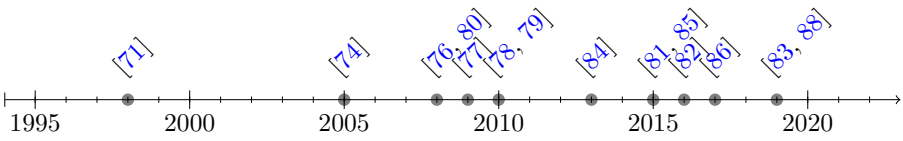

Fig. 10 Timeline of the tools for interactive data annotation (discussed in Section 4).

user characteristics than others. This section discusses examples of interactive information visualization tools developed for different task domains.

5.1 Music Analysis through Visualization

In the case of audio data, music, for example, has its own notation to describe lines of notes on staffs with tempo and dynamics markings, and sound property visualization techniques such as waveforms and spectral frequency displays are common in most sound production software tools. Music visualization tools that are often found in popular media players provide users with abstract and artistic imagery based on the music. However, all these audio data representations have clear limitations as to when to be used–only a very small population of musicians can recreate the music in their head from the written score, specialized sound property visualizations can only be understood by expert sound engineers, and animated artistic imagery of music visualization is often too abstract to mentally recreate the sound accurately from the visualization itself. In this section, we discuss example tools that are focused on visualization and interaction approaches that are more informative for general users.

Hiraga et. al [89] developed a prototype music visualization tool that utilizes both the traditional music score notation to visualize the music to be played and the Chernoff faces [90] to visualize the musician's performance. A Chernoff face uses a set of facial expression parameters to indicate certain

aspects of the music performance. In their implementation, the musical parameters such as tempo, articulation, and sound level are visualized by mapping them to the position of the eyes, the contour of the face and the shape of the mouth, and the shape of the nose, respectively. While the user study results suggest the Chernoff face expressions had many aspects to be improved, it was an interesting sound visualization approach to explore. SeeGroove [91, 92] is another music visualization system that employs a module-oriented multithreaded architecture representing the *grooves* of music, generalizing certain musical aspects such as surges, propulsive rhythmic feel, dynamics, and togetherness of sound to reflect overall rhythmic pleasure. Grooves are the visualization scheme based on an orbit metaphor where an audio file and its details are broken down and represented in the form of an orbiting object.

Several music visualization systems have been proposed, capitalizing on the recent development of the MIR approaches such as those based on machine learning techniques. For example, SmartDJ [93] provides an interface with a reduced feature space using the Principle Component Analysis (PCA), on which the users click and select songs from various graphical regions. The system first extracts 28 low-level features that include brightness, centroid, rolloff, and Mel-frequency Cepstral Coefficient (MFCC), and the dimensionality is reduced by the PCA. Its UI has multiple views of similarity maps of songs, which allow the users (DJs) to select the next songs. In a web-based Interactive system for Multi-modal Music Analysis (IMMA)) [94], tonal tension and timbre tension are calculated from the music score and the music performance, respectively, an audio-to-score alignment algorithm based on dynamic time warping was used to visualize the automatically synchronized score and performance for music analysis. It also provides the visualization of similar music segments based on semantic tonality and clustering. MixMash [95, 96] allows the interactive visualization of multidimensional musical attributes that are extracted from a collection of audio files such as hierarchical harmonic compatibility, onset density, spectral region, and timbral similarities. Harmonically similar audio tunes are categorized and visualized as tree nodes whose distances and edge connections indicate their harmonic compatibility as a result of a force-directed graph. Users interact with the UI to explore the data-space changing how they are organized (e.g., changing values of forces of attraction and repulsion between nodes) and zooming and panning to view more details.

Several tools have incorporated extra dimensions of music performance by detecting and analyzing some human body movements as part of the interaction with the visualization tools. MIMOSE [97] uses multimodal user interactions to enter music data utilizing physical gestures and speech in such a way as to mimic the orchestra conductor's movements. Initially, the user's gestures and voice instructions are detected by using body movement and voice sensors, and they are then translated to simulate mouse clicks and key presses that are used to enter music data. As part of i-Maestro [98], a project that focuses on the education of music theory and performance, Ng et al. developed an interface that integrates the sound and gesture analysis and provides the

visualization of the analysis results for educators and learners to understand their performance. Although their use of gestures is not for the explicit musical data, there are systems that utilize unorthodox interactions with music visualizations and are worth mentioning here: Vuzik [99, 100] visualizes music by a painting metaphor (shown in Fig. 11), by interpreting simple painting gestures into musical parameters (e.g., thickness defines loudness of music, colours define instruments, note lengths define pitches). Users draw on the provided interface and the body motion is converted to audio data based on the corresponding musical parameters. Shimon [101] also uses the hand movement of the user as input and deals with robots learning audio synchronization with the human in a jam session using various body movement sensors and modules namely nodding and saliency-based gazing. The robot shifts its gaze to perceive the music and maintains synchronization with the human.

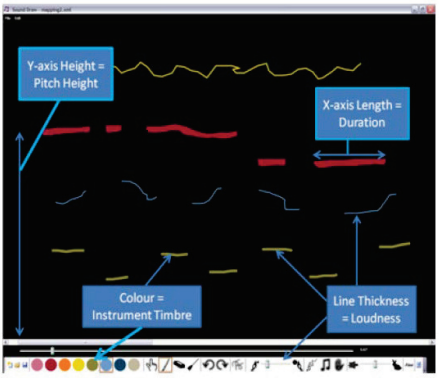

Fig. 11 Vuzik mapping of sound to visuals (image reproduced from [99]).

5.2 Interactive Visualization in Other Domains

In addition to supporting multimedia content authoring tasks, many different multimedia data visualization approaches and techniques have also been developed and adapted in other domains. Two such domains are Data Analytics (Section 5.2.1) and Teaching and Learning (Section 5.2.2). This section discusses examples of interactive visualization tools used in these areas.

5.2.1 Visualization for Data Analytics

Interactions with visualized data often introduce some useful insights that are not otherwise easily attainable from static visualizations. Interactive data analysis is one of the research areas that were developed, capitalizing on the recent advancement of the data analytics approaches and the powerful computational resources [102]. Various tools have been introduced for this purpose. An interactive visualization, called a table of video contents (TOVC) [103], extracts video frames and rearranges them on a 2D plane, similar to the shape of a

rosary. This visualization, created based on the results of the feature extraction and clustering of their similarities, serves as a handy analysis tool for understanding video content. Matchpad [104] is a glyph-based visualization, designed for real-time sports performance analysis. It visualizes some of the sports-specific statistics and provides interactions for the users to rapidly seek information in real time while watching sports games. An interactive visual analytics tool, eVADE [105], was designed for investigating big earth observation data content, with its main application areas being emergency services and disaster management. The tool allows its users to view the geospatial visualization of the earth data and interactively investigate the data as well as annotating/labelling the visual contents.

Instead of focusing on a specific application domain, there are generic interactive data analytics tools. For example, Heer and Bostock [106] developed declarative, domain-specific languages for constructing interactive visualizations. A declarative language, called Protovis, was designed to visualize the numerical data within the semantics of the grammatical rules. It uses a multi-stage pipeline, (1) bind, (2) build, (3) evaluate, (4) interpolate, (5) render, and (6) event, to instantiate visualization specifications. Vega [107] is one of the earlier attempts to construct the declarative visualization grammar that enables sharing and reuse of the data and presents the data in an interactive context. Lyra [108] is a tool that is based on Vega in order to visualize data and allow for user interactions with visual objects, such as drag and drop data points for expressive designs, click and select filters, and rotate handles. Extending Vega, Reactive Vega [109] draws on Event-Driven Functional Reactive Programming (E-FRP) integrating streaming database techniques. Using the streaming techniques allows the user to visualize the real-time data. Furthermore, this tool is extended as Vega-Lite [110] (shown in Fig. 12), which enables rapid specification of interactive data visualizations allowing the user to transform, zoom and select the real-time data for flexibility.

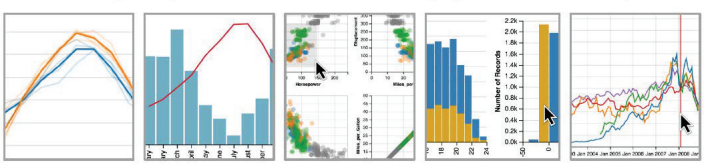

Fig. 12 Visualization by Vega-Lite (image reproduced from [110]).

5.2.2 Interactive Multimedia Visualization in Education

With the increasing use of teaching and learning technologies such as Learning Management Systems (LMS) and the most recent shift to the online teaching platforms largely due to the worldwide pandemic situations, multimedia data play a vital role in education, from grabbing learner's attention [111], to enhancing their understanding of topics by providing multiple communication

channels (e.g., visual, auditory, motor). Providing appropriate interactions for the users (i.e., both instructors and learners) to create, view/listen to, and manipulate multimedia data for educational purposes has thus become almost inevitable in teaching and learning.

Among the domains of interactive multimedia for education, scientific simulations are perhaps one of the most popular application areas: 3DNormalModes [112] simulates a molecule model with its properties to study it in a real 3D environment (shown in Fig. 13). *World in Motion* [113] is a multimedia teaching tool that allows students to practice various physics experiments in a simulated environment. Both these tools allow learners to manipulate visual objects for interactive investigations of these objects.

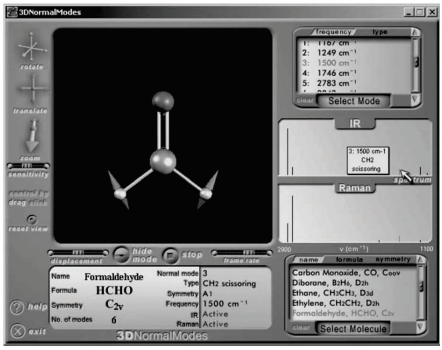

Fig. 13 A screenshot of 3DNormalModes (image reproduced from [112]).

Another popular educational interactive multimedia visualization is for early childhood education: an RFID-Bluetooth-based tool [114] allows children to learn about new objects/entities by tapping on physical objects and produces the semantic description of the corresponding objects. e-Pumapunku [115], a mobile app, is used at children's homes or at a museum to study about the Cañari and Inca indigenous cultures, by identifying QR codes to show multimedia material to children, integrating 3D objects in the augmented reality (AR) environment, and capitalizing on a data mining algorithm for the analysis of museum visitors' data. Alongside such tools used in a specific domain, tools like Powerchalk [116] and Power Electronics Library [117] are used to provide a robust, interactive learning surface to any intelligent environment focused on blended learning. The users can be of any age ground using the surface and navigating through it.

This section discussed tools for interactive multimedia data visualization. In these applications, the users interact with visualized data objects to modify how they appear on displays. They often change which information/features of the data to be displayed through various types of interaction. This interactivity allows users to gain certain new knowledge that is otherwise not easily

acquirable through static visualizations. The approaches can be used in different domains such as data analytics and education. Table 4 summarizes the tools discussed in this section and Figure 14 shows the corresponding publications in chronological order.

Table 4 Summary of the tools for interactive data annotation (discussed in Section 5).

Domains/Tasks	**Algorithms/Interactions: Tools**
Music Analysis through Visualization (Section 5.1)	Visualize common musical parameters (tempo, articulation, surges, propulsive rhythmic feel, dynamics): [89, 91, 92]
	ML model to visualize similarities in audio features (brightness, tonal/timbre tension, MFCC, harmony, etc.) :[93–96]
	Integrate human body movements: [97–101]
Visualization for Data Analytics (Section 5.2.1)	Feature extraction and clustering of similarities of video content: [103]
	Glyph-based visualization, designed for real-time sports performance analysis: [104]
	Geospatial interactive visualization of the earth data: [105]
	Declarative visualization grammar: [106–110]
Interactive Multimedia Visualization in Education (Section 5.2.2)	Scientific simulations: [112, 113]
	RFID-Bluetooth based tool early childhood education: [114]
	QR codes to show multimedia to children in an AR environment: [115]
	Robust, interactive learning surface focused on blended learning: [116, 117]

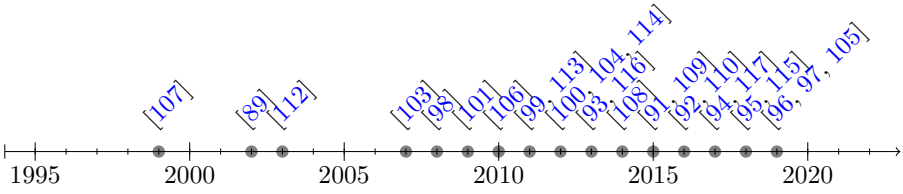

Fig. 14 Timeline of the tools for interactive data annotation (discussed in Section 5).

6 Conclusion

In this paper, we presented a summary of past and current research on various types of tools and approaches that assist user interactions with multimedia data. The list of research discussed in this paper is not exhaustive one in any means, as such a list can scale exponentially and easily go beyond the scope of a single article. What this article provides is, however, an overview of the current status of interactive tools for multimedia tasks in several research domains. Figure 15 shows the timeline of all the tools we discussed in this article. In order to convey the updated statuses of the related fields, we focused on more recent publications than older ones, and thus, the actual temporal distribution of all existing interactive multiple tools is not likely reflected in the distribution depicted in this figure.

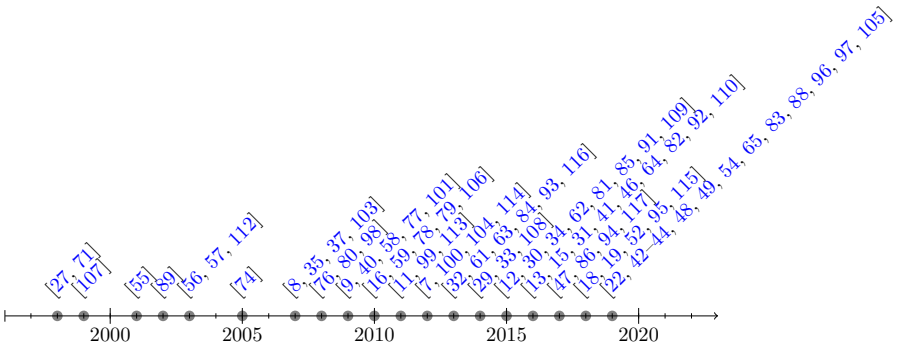

Fig. 15 Timeline of the tools discussed in this article (combined).

This work can be expanded in two main directions. One is a rather traditional approach to look at those tools within a single domain and run deeper analyses of how each tool was influenced by past techniques and how it influenced tools that followed it. This type of analysis may help come up with ideas to provide new interaction techniques for multimedia data as extensions or modifications of past techniques. The other is to look further from multi- and inter-disciplinary perspectives and analyze these tools more in detail and see if and how their approaches may be applied to different types of multimedia tasks. We hope that this article will serve as a starting point for researchers in the related domains and will inspire them to look beyond the boundaries of traditional study areas/topics so as to learn and understand new and potentially old techniques and interactions that have been experimented and implemented for different multimedia data for different purposes.

With the increasing use of digital devices, especially mobile devices that can easily produce and share different types of multimedia content such as pictures, videos, and sound, we are now in more need of tools to understand and manipulate multimedia data in efficient and effective ways than ever. This paper will serve as a hub for researchers of such multimedia tools for finding

related and similar tools so that they can leverage the findings and efforts of other researchers to move our knowledge and techniques forward.

Declarations

Authors can confirm that all relevant data are included in the article and/or its supplementary information files.

References

[1] Li, T., Ogihara, M., Tzanetakis, G.: Music Data Mining. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, Boca Raton (2011)

[2] Goodrum, A.: Image information retrieval: An overview of current research. Informing Science **3**, 2000 (2000)

[3] Zhou, W., Li, H., Tian, Q.: Recent advance in content-based image retrieval: A literature survey. CoRR **abs/1706.06064** (2017) https://arxiv.org/abs/1706.06064

[4] Neumann, F., Wegener, I.: Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. Theoretical Computer Science **378**(1), 32–40 (2007)

[5] Liu, T., Kender, J.R., Hjelsvold, R., Pizano, A.: A fast image segmentation algorithm for interactive video hotspot retrieval. In: Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL 2001), pp. 3–8 (2001). IEEE

[6] Masaki, T., Yamaguchi, T., Kitamura, Y.: An interactive digital fishtank based on live video images. In: International Conference on Advanced Multimedia Content Processing, pp. 386–396 (1998). Springer

[7] Giró-i-Nieto, X., Martos, M.: Interactive segmentation and tracking of video objects. In: 2012 13th International Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4 (2012). IEEE

[8] Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007). IEEE

[9] Price, B.L., Morse, B.S., Cohen, S.: Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 779–786 (2009). IEEE

[10] Greig, D.M., Porteous, B.T., Seheult, A.: Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society, Series B **51**, 271–279 (1989). https://doi.org/10.1111/j.2517-6161.1989.tb01764.x

[11] Yang, L., Wu, X., Guo, Y., Li, S.: An interactive video segmentation approach based on grabcut algorithm. In: 2011 4th International Congress on Image and Signal Processing, vol. 1, pp. 367–370 (2011). IEEE

[12] Shankar Nagaraja, N., Schmidt, F.R., Brox, T.: Video segmentation with just a few strokes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3235–3243 (2015)

[13] Spina, T.V., Falcao, A.X.: Fomtrace: Interactive video segmentation by image graphs and fuzzy object models. arXiv preprint arXiv:1606.03369 (2016)

[14] Rauber, P.E., Falcão, A., Spina, T., Rezende, P.J.: Interactive segmentation by image foresting transform on superpixel graphs. 2013 XXVI Conference on Graphics, Patterns and Images, 131–138 (2013)

[15] Levinkov, E., Tompkin, J., Bonneel, N., Kirchhoff, S., Andres, B., Pfister, H.: Interactive Multicut Video Segmentation. In: Grinspun, E., Bickel, B., Dobashi, Y. (eds.) Pacific Graphics Short Papers, pp. 33–38. The Eurographics Association, Geneva, Switzerland (2016)

[16] Yang, W., Cai, J., Zheng, J., Luo, J.: User-friendly interactive image segmentation through unified combinatorial user inputs. IEEE Transactions on Image Processing **19**(9), 2470–2479 (2010)

[17] Shenvi, N., Kempe, J., Whaley, K.B.: Quantum random-walk search algorithm. Physical Review A **67**(5), 052307 (2003)

[18] Maninis, K.-K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 616–625 (2018)

[19] Wang, J., Zhao, Y., Qi, Q., Huo, Q., Zou, J., Ge, C., Liao, J.: Mindcamera: Interactive sketch-based image retrieval and synthesis. IEEE Access **6**, 3765–3773 (2018)

[20] Bui, T., Collomosse, J.: Scalable sketch-based image retrieval using color gradient features. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1–8 (2015)

[21] Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015) https://arxiv.org/abs/1506.02640

[22] Akiyama, Y., Barrantes, R.G., Hynes, T.: Video scene extraction tool for soccer goalkeeper performance data analysis. In: IUI Workshops (2019)

[23] Oyama, T., Nakao, D.: Automatic extraction of specific scene from sports video. In: 2015 10th Asian Control Conference (ASCC), pp. 1–4 (2015)

[24] Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing **12**(7), 796–807 (2003)

[25] Baillie, M., Jose, J.M.: An audio-based sports video segmentation and event detection algorithm. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 110–110 (2004)

[26] Kato, T.: Database architecture for content-based image retrieval. In: Jamberdino, A.A., Niblack, C.W. (eds.) Image Storage and Retrieval Systems, vol. 1662, pp. 112–123. SPIE, Bellingham WA (1992). International Society for Optics and Photonics

[27] Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Transactions on circuits and systems for video technology **8**(5), 644–655 (1998)

[28] Khakoo, S.A.: Signature-based search algorithm. In: International Conference on Acoustics, Speech, and Signal Processing,, pp. 1874–18773 (1989)

[29] Lokoč, J., Blažek, A., Skopal, T.: Signature-based video browser. In: International Conference on Multimedia Modeling, pp. 415–418 (2014). Springer

[30] Blažek, A., Lokoč, J., Matzner, F., Skopal, T.: Enhanced signature-based video browser. In: International Conference on Multimedia Modeling, pp. 243–248 (2015). Springer

[31] Kuboň, D., Blažek, A., Lokoč, J., Skopal, T.: Multi-sketch semantic video browser. In: International Conference on Multimedia Modeling, Berlin, Heidelberg, pp. 406–411 (2016). Springer

[32] Le, D.-D., Lam, V., Ngo, T.D., Tran, V.Q., Nguyen, V.H., Duong, D.A., Satoh, S.: NII-UIT-VBS: A video browsing tool for known item search. In: Advances in Multimedia Modeling, pp. 547–549. Springer, New York,

NY (2013)

[33] Ngo, T.D., Nguyen, V.H., Lam, V., Phan, S., Le, D.-D., Duong, D.A., Satoh, S.: NII-UIT: a tool for known item search by sequential pattern filtering. In: International Conference on Multimedia Modeling, pp. 419–422 (2014). Springer

[34] Ngo, T.D., Nguyen, V.-T., Nguyen, V.H., Le, D.-D., Duong, D.A., Satoh, S.: Nii-uit browser: a multimodal video search system. In: International Conference on Multimedia Modeling, pp. 278–281 (2015). Springer

[35] Yoo, H.-W., Cho, S.-B.: Video scene retrieval with interactive genetic algorithm. Multimedia Tools and Applications **34**(3), 317–336 (2007)

[36] Sanchez, E., Shibata, T., Zadeh, L.A.: Genetic Algorithms and Fuzzy Logic Systems: Soft Computing Perspectives vol. 7. World Scientific, Singapore (1997)

[37] Joly, A., Tjondronegoro, D.: An adaptive and extensible web-based interface system for interactive video contents browsing. In: Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp. 1–6. Springer, New York, NY (2007)

[38] Tjondronegoro, D.: Content-based video indexing for sports applications using intergrated multimodal approach. PhD thesis, Deakin University (2005)

[39] Hurst, W., Jarvers, P.: Interactive, dynamic video browsing with the zoomslider interface. In: 2005 IEEE International Conference on Multimedia and Expo, p. 4 (2005). IEEE

[40] Schoeffmann, K., Taschwer, M., Boeszoermenyi, L.: Video browsing using motion visualization. In: 2009 IEEE International Conference on Multimedia and Expo, pp. 1835–1836 (2009). IEEE

[41] Zhang, Z., Li, W., Gurrin, C., Smeaton, A.F.: Faceted navigation for browsing large video collection. In: International Conference on Multimedia Modeling, pp. 412–417 (2016). Springer

[42] Le, T.-K., Ninh, V.-T., Dang-Nguyen, D.-T., Tran, M.-T., Zhou, L., Redondo, P., Smyth, S., Gurrin, C.: Lifeseeker: Interactive lifelog search engine at lsc 2019. In: Proceedings of the ACM Workshop on Lifelog Search Challenge, pp. 37–40 (2019). ACM

[43] Lokoč, J., Kovalčík, G., Souček, T., Moravec, J., Čech, P.: VIRET: A video retrieval tool for interactive known-item search. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp.

177–181 (2019). ACM

[44] Kovalčík, G., Škrhak, V., Souček, T., Lokoč, J.: VIRET: Tool with advanced visual browsing and feedback. In: Proceedings of the Third Annual Workshop on Lifelog Search Challenge. LSC '20, pp. 63–66. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3379172.3391725. https://doi-org.library.smu.ca/10.1145/3379172.3391725

[45] Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: 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)

[46] Moumtzidou, A., Mironidis, T., Apostolidis, E., Markatopoulou, F., Ioannidou, A., Gialampoukidis, I., Avgerinakis, K., Vrochidis, S., Mezaris, V., Kompatsiaris, I., *et al.*: Verge: A multimodal interactive search engine for video browsing and retrieval. In: International Conference on Multimedia Modeling, pp. 394–399 (2016). Springer

[47] Moumtzidou, A., Mironidis, T., Markatopoulou, F., Andreadis, S., Gialampoukidis, I., Galanopoulos, D., Ioannidou, A., Vrochidis, S., Mezaris, V., Kompatsiaris, I., *et al.*: Verge in vbs 2017. In: International Conference on Multimedia Modeling, pp. 486–492 (2017). Springer

[48] Andreadis, S., Moumtzidou, A., Galanopoulos, D., Markatopoulou, F., Apostolidis, K., Mavropoulos, T., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I., *et al.*: Verge in vbs 2019. In: International Conference on Multimedia Modeling, pp. 602–608 (2019). Springer

[49] Andreadis, S., Moumtzidou, A., Apostolidis, K., Gkountakos, K., Galanopoulos, D., Michail, E., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: Verge in vbs 2020. In: Ro, Y.M., Cheng, W.-H., Kim, J., Chu, W.-T., Cui, P., Choi, J.-W., Hu, M.-C., De Neve, W. (eds.) MultiMedia Modeling, pp. 778–783. Springer, Cham (2020)

[50] Meng, J., Yuan, J., Wang, G., Xu, J.: Object instance search in videos. In: 2013 9th International Conference on Information, Communications & Signal Processing, pp. 1–4 (2013). IEEE

[51] Ueki, K., Hirakawa, K., Kikuchi, K., Ogawa, T., Kobayashi, T.: Waseda meisei at trecvid 2017: Ad-hoc video search. In: TRECVID Workshop (2017)

[52] Primus, M.J., Münzer, B., Leibetseder, A., Schoeffmann, K.: The ITEC collaborative video search system at the video browser showdown 2018. In: International Conference on Multimedia Modeling, pp. 438–443

(2018). Springer

[53] Schoeffmann, K.: Video browser showdown 2012-2019: A review. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–4 (2019). https://doi.org/10.1109/CBMI.2019.8877397

[54] Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., *et al.*: Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, p. 4 (2019). ACM

[55] Fernstrom, M., Brazil, E.: Sonic browsing: an auditory tool for multimedia asset management. In: International Conference on Auditory Display, 2001 (2001). Georgia Institute of Technology

[56] Brazil, E., Fernstrom, M.: Audio information browsing with the sonic browser. In: Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization-CMV 2003-, pp. 26–31 (2003). IEEE

[57] Tzanetakis, G.: MUSESCAPE: An interactive content-aware music browser. In: Proc. Conference on Digital Audio Effects (DAFX) (2003)

[58] Chen, Y.-X., Butz, A.: MusicSim: integrating audio analysis and user feedback in an interactive music browsing UI. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, pp. 429–434 (2009). ACM

[59] Gallagher, M., Gainza, M., Fitzgerald, D., Barry, D., Cranitch, M., Coyle, E.: Interactive music archive access system. In: 2010 IEEE International Conference on Multimedia and Expo, pp. 723–724 (2010). IEEE

[60] Rickard, S.: The duet blind source separation algorithm. In: Blind Speech Separation, pp. 217–241. Springer, New York, NY (2007)

[61] Hamasaki, M., Goto, M.: Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community. In: Proceedings of the 9th International Symposium on Open Collaboration, p. 4 (2013). ACM

[62] Hamasaki, M., Goto, M., Nakano, T.: Songrium: Browsing and listening environment for music content creation community. Proc. of SMC 2015, 23–30 (2015)

[63] Porter, A., Sordo, M., Serra, X.: Dunya: A system for browsing audio

music collections exploiting cultural context. In: Britto A, Gouyon F, Dixon S. 14th International Society for Music Information Retrieval Conference (ISMIR); 2013 Nov 4-8; Curitiba, Brazil: ISMIR; 2013. P. 101-6. (2013). International Society for Music Information Retrieval (ISMIR)

[64] Allik, A., Fazekas, G., Barthet, M., Sandler, M.: mymoodplay: An interactive mood-based music discovery app. In: Freeman, J., Lerch, A., Paradis, M. (eds.) Proceedings of the International Web Audio Conference. WAC '16. Georgia Tech, Atlanta, GA, USA (2016)

[65] Hu, X., Que, Y., Kando, N., Lian, W.: Analyzing user interactions with music information retrieval system: An eye-tracking approach. In: Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR) (2019)

[66] Hauptmann, A., Yan, R., Lin, W.-H.: How many high-level concepts will fill the semantic gap in news video retrieval? In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 627–634 (2007). ACM

[67] Snoek, C.G., Huurnink, B., Hollink, L., De Rijke, M., Schreiber, G., Worring, M.: Adding semantics to detectors for video retrieval. IEEE Transactions on multimedia **9**(5), 975–986 (2007)

[68] Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., Kompatsiaris, Y.: A survey of semantic image and video annotation tools. In: Knowledge-driven Multimedia Information Extraction and Ontology Evolution, pp. 196–239. Springer, New York, NY (2011)

[69] Mackay, W.E., Davenport, G.: Virtual video editing in interactive multimedia applications. Communications of the ACM **32**(7), 802–810 (1989)

[70] Vajda, F.: Techniques and trends in digital image processing and computer vision. In: IEE Colloquium on Mathematical Modelling and Simulation of Industrial and Economic Processes, pp. 1–1 (1994). IET

[71] Dittmann, J., Neck, F., Steinmetz, A., Steinmetz, R.: Interactive watermarking environments. In: Proceedings. IEEE International Conference on Multimedia Computing and Systems (Cat. No. 98TB100241), pp. 286–294 (1998). IEEE

[72] Kutter, M., Jordan, F.D., Bossen, F.: Digital signature of color images using amplitude modulation. In: Storage and Retrieval for Image and Video Databases V, vol. 3022, pp. 518–526 (1997). International Society for Optics and Photonics

[73] Dittmann, J., Stabenau, M.: Robust mpeg video watermarking technologies. In: Multimedia and Security Workshop at ACM Multimedia, vol. 98, pp. 1998–113 (1998). Citeseer

[74] Yang, C., Dong, M., Fotouhi, F.: I^2A: an interactive image annotation system. In: 2005 IEEE International Conference on Multimedia and Expo, p. 4 (2005). IEEE

[75] Fellbaum, C.: WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA (1998). https://doi.org/10.7551/mitpress/7287.001.0001. https://doi.org/10.7551/mitpress/7287.001.0001

[76] Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International journal of computer vision **77**(1-3), 157–173 (2008)

[77] Yuen, J., Russell, B., Liu, C., Torralba, A.: Labelme video: Building a video database with human annotations. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1451–1458 (2009). IEEE

[78] Torralba, A., Russell, B.C., Yuen, J.: Labelme: Online image annotation and applications. Proceedings of the IEEE **98**(8), 1467–1484 (2010)

[79] Vezzani, R., Cucchiara, R.: Video surveillance online repository (ViSOR): an integrated framework. Multimedia Tools and Applications **50**(2), 359–380 (2010)

[80] Hu, W., Wen, J., Gong, H., Wang, Y.: An interactive scene annotation tool for video surveillance. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4 (2008). IEEE

[81] Kalboussi, A., Omheni, N., Mazhoud, O., Kacem, A.H.: An interactive annotation system to support the learner with web services assistance. In: 2015 IEEE 15th International Conference on Advanced Learning Technologies, pp. 409–410 (2015). IEEE

[82] Bryan, C., Ma, K.-L., Woodring, J.: Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement. IEEE transactions on visualization and computer graphics **23**(1), 511–520 (2016)

[83] Münzer, B., Leibetseder, A., Kletz, S., Schoeffmann, K.: Ecat-endoscopic concept annotation tool. In: International Conference on Multimedia Modeling, pp. 571–576 (2019). Springer

[84] Chiang, C.-C.: Interactive tool for image annotation using a semi-supervised and hierarchical approach. Computer Standards & Interfaces **35**(1), 50–58 (2013)

[85] Bianco, S., Ciocca, G., Napoletano, P., Schettini, R.: An interactive tool for manual, semi-automatic and automatic video annotation. Computer Vision and Image Understanding **131**, 88–99 (2015)

[86] Yoon, U.-N., Hong, M.-D., Jo, G.-S.: Automatic interactive video authoring method via object recognition. In: Asian Conference on Intelligent Information and Database Systems, pp. 589–598 (2017). Springer

[87] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc., Red Hook, NY (2015). https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[88] Halter, G., Ballester-Ripoll, R., Flueckiger, B., Pajarola, R.: Vian: A visual annotation tool for film analysis. Computer Graphics Forum **38**, 119–129 (2019). https://doi.org/10.1111/cgf.13676

[89] Hiraga, R., Watanabe, F., Fujishiro, I.: Music learning through visualization. In: Second International Conference on Web Delivering of Music, 2002. WEDELMUSIC 2002. Proceedings., pp. 101–108 (2002). IEEE

[90] Lee, M.D., Reilly, R.E., Butavicius, M.E.: An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In: Proceedings of the Asia-Pacific Symposium on Information visualisation-Volume 24, pp. 1–10 (2003). Australian Computer Society, Inc.

[91] Fujishiro, I., Haga, N., Nakayama, M.: Seegroove: Supporting groove learning through visualization. In: 2015 International Conference on Cyberworlds (CW), pp. 189–192 (2015). IEEE

[92] Jin, N., Haga, N., Fujishiro, I.: Seegroove2: An orbit metaphor for interactive groove visualization. In: 2016 International Conference on Cyberworlds (CW), pp. 131–134 (2016). IEEE

[93] Aw, M.S., Lim, C.S., Khong, A.W.: Smartdj: An interactive music player for music discovery by similarity comparison. In: 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–5 (2013). IEEE

[94] Herremans, D., Chuan, C.-H.: A multi-modal platform for semantic music analysis: visualizing audio-and score-based tension. In: 2017 IEEE

11th International Conference on Semantic Computing (ICSC), pp. 419–426 (2017). IEEE

[95] Maçãs, C., Rodrigues, A., Bernardes, G., Machado, P.: Mixmash: A visualisation system for musical mashup creation. In: 2018 22nd International Conference Information Visualisation (IV), pp. 471–477 (2018). IEEE

[96] Maçãs, C., Rodrigues, A., Bernardes, G., Machado, P.: Mixmash: An assistive tool for music mashup creation from large music collections. International Journal of Art, Culture and Design Technologies (IJACDT) **8**(2), 20–40 (2019)

[97] Coletta, A., De Marsico, M., Panizzi, E., Prenkaj, B., Silvestri, D.: Mimose: multimodal interaction for music orchestration sheet editors. Multimedia Tools and Applications, 1–28 (2019)

[98] Ng, K., Nesi, P.: i-maestro: Technology-enhanced learning and teaching for music. In: NIME, pp. 225–228 (2008)

[99] Pon, A., Ichino, J., Eagle, D., Sharlin, E., Carpendale, S.: Vuzik: Music visualization and creation on an interactive surface. Technical report, University of Calgary (2011)

[100] Pon, A., Ichino, J., Eagle, D., Sharlin, E., d'Alessandro, N., Carpendale, S.: Vuzik: A painting graphic score interface for composing and control of sound generation. In: ICMC (2012)

[101] Weinberg, G., Raman, A., Mallikarjuna, T.: Interactive jamming with shimon: a social robotic musician. In: 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 233–234 (2009). IEEE

[102] Heer, J., Kandel, S.: Interactive analysis of big data. XRDS **19**(1), 50–54 (2012). https://doi.org/10.1145/2331042.2331058

[103] Goeau, H., Thièvre, J., Viaud, M.-L., Pellerin, D.: Interactive visualization tool with graphic table of video contents. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 807–810 (2007). IEEE

[104] Legg, P., Chung, D., Parry, M., Jones, M., Long, R., Griffiths, I., Chen, M.: Matchpad : Interactive glyph-based visualization for real-time sports performance analysis. Computer Graphics Forum **31**, 1255–1264 (2012). https://doi.org/10.1111/j.1467-8659.2012.03118.x

[105] Daniela, F., GRIPARIS, A., STOICA, A., MOUGNAUD, P., DATCU, M.: An interactive visual analytics tool for big earth observation data

content estimation. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 9518–9521 (2019). IEEE

[106] Heer, J., Bostock, M.: Declarative language design for interactive visualization. IEEE Transactions on Visualization and Computer Graphics **16**(6), 1149–1156 (2010)

[107] Hipke, C.A., Schuierer, S.: Vega-a user-centered approach to the distributed visualization of geometric algorithms. Technical report, Institut für Informatik Universität Freiburg (1999)

[108] Satyanarayan, A., Heer, J.: Lyra: An interactive visualization design environment. Computer Graphics Forum **33**(3), 351–360 (2014)

[109] Satyanarayan, A., Russell, R., Hoffswell, J., Heer, J.: Reactive vega: A streaming dataflow architecture for declarative interactive visualization. IEEE transactions on visualization and computer graphics **22**(1), 659–668 (2015)

[110] Satyanarayan, A., Moritz, D., Wongsuphasawat, K., Heer, J.: Vega-lite: A grammar of interactive graphics. IEEE transactions on visualization and computer graphics **23**(1), 341–350 (2016)

[111] Toomey, R., Ketterer, K.: Using multimedia as a cognitive tool. Journal of Research on Computing in Education **27**(4), 472–482 (1995)

[112] Sigalas, M.P., Charistos, N.D., Teberekidis, V.I., Tsipis, C.A.: 3DNormalModes. ACS Publications (2003)

[113] Žovínová, M., Ožvoldová, M.: New multimedia teaching tool using remote physics experiments. In: 2011 14th International Conference on Interactive Collaborative Learning, pp. 395–399 (2011). IEEE

[114] Karime, A., Hossain, M.A., Rahman, A.M., Gueaieb, W., Alja'am, J.M., El Saddik, A.: Rfid-based interactive multimedia system for the children. Multimedia Tools and Applications **59**(3), 749–774 (2012)

[115] Arias-Espinoza, P., Medina-Carrión, A., Robles-Bykbaev, V., Robles-Bykbaev, Y., Pesántez-Avilés, F., Ortega, J., Matute, D., Roldán-Monsalve, V.: e-pumapunku: An interactive app to teach children the cañari and inca indigenous cultures during guided museum visits. In: 2018 Congreso Internacional de Innovación Y Tendencias en Ingeniería (CONIITI), pp. 1–5 (2018). IEEE

[116] Esponda-Argüero, M., Rojas, R., *et al.*: Powerchalk: An adaptive e-learning application. In: Multimedia and Internet Systems: Theory and Practice, pp. 179–188. Springer, New York, NY (2013)

[117] Trujillo-Aguilera, F., Sotorrío-Ruiz, P.J., Blázqucz-Parra, E.: Improving the power electronics laboratory teaching/learning process: an interactive web tool. In: 2017 7th World Engineering Education Forum (WEEF), pp. 273–278 (2017). IEEE