US 20250265842A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2025/0265842 A1

Ghose et al. (43) Pub. Date: Aug. 21, 2025

(54) **MULTI-MODAL METADATA EXTRACTION SYSTEM**

(71) Applicant: **ANOKI, INC.**, San Carlos, CA (US)

(72) Inventors: **Susmita Ghose**, Mountain View, CA (US); **Ashutosh Chaubey**, Chhattisgarh (IN); **Sartaki Sinha Roy**, Uttarpara (IN); **Ashish Baldua**, San Jose, CA (US)

(73) Assignee: **ANOKI, INC.**, San Carlos, CA (US)

(21) Appl. No.: **18/581,328**

(22) Filed: **Feb. 19, 2024**

Publication Classification

(51) **Int. Cl.**

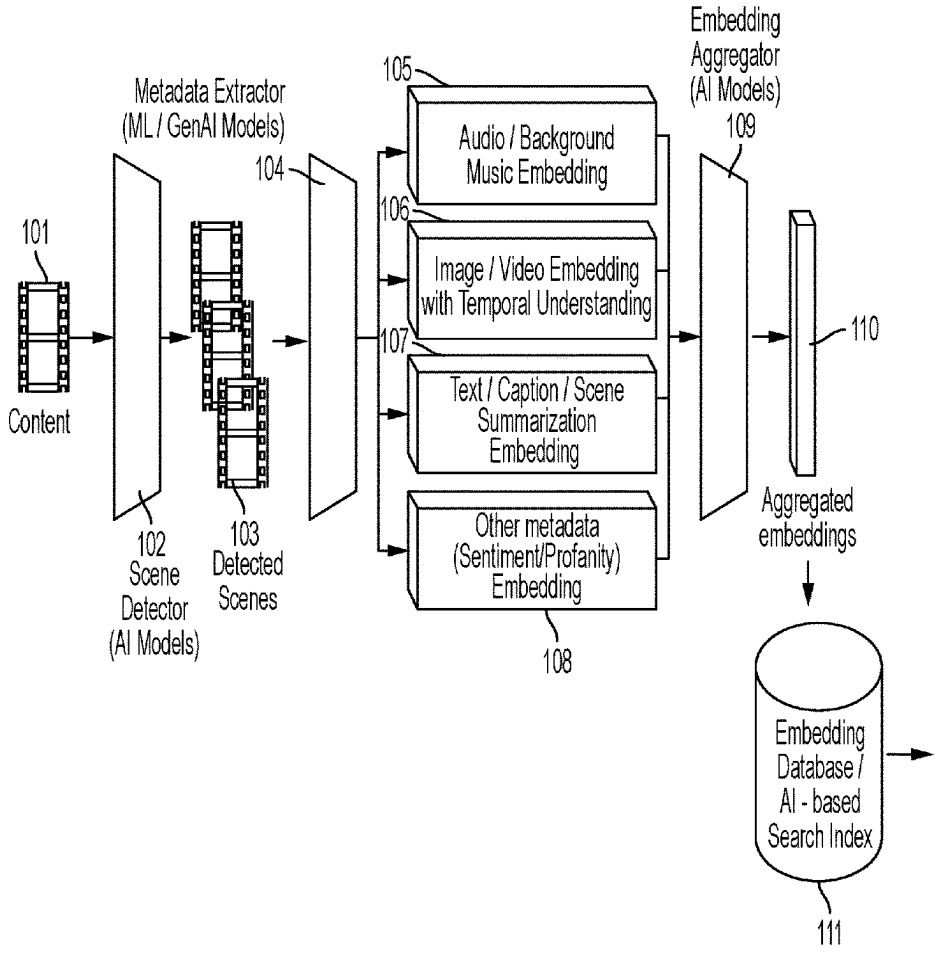
G06V 20/40	(2022.01)
G06V 10/44	(2022.01)
G06V 10/74	(2022.01)
G06V 10/762	(2022.01)

(52) **U.S. Cl.**

CPC ***G06V 20/46*** (2022.01); ***G06V 10/44*** (2022.01); ***G06V 10/761*** (2022.01); ***G06V 10/762*** (2022.01); ***G06V 20/49*** (2022.01); *G06V 2201/10* (2022.01)

(57) **ABSTRACT**

A multimodal metadata extraction system may be provided with a scene detector having a video content input and an output representing scene boundaries. The metadata extractor may be responsive to the content of a scene to extract metadata corresponding to several, plural, or multiple extraction modes. A metadata embedding may be used for each of the modes. An embedding aggregator responsive to the embedding operates to formulate an aggregated embedding for each scene thereby indexing the content of the scene. The scene detector may include a frame analyzer for identifying consecutive frames having similar characteristics. A boundary detector may be provided to identify boundaries of consecutive frames having sufficiently similar characteristics that they likely belong to the same shot. An embedding system may be provided to formulate a composite distance matrix capturing the distance between shot embeddings. A temporal clustering system may be connected to the composite distance matrix. An output of the temporal clustering system identifies the scene boundaries of the content. An embedding database may be connected to the embedding aggregator for storing the aggregated embedding for use as a search index for scenes identified in the content.



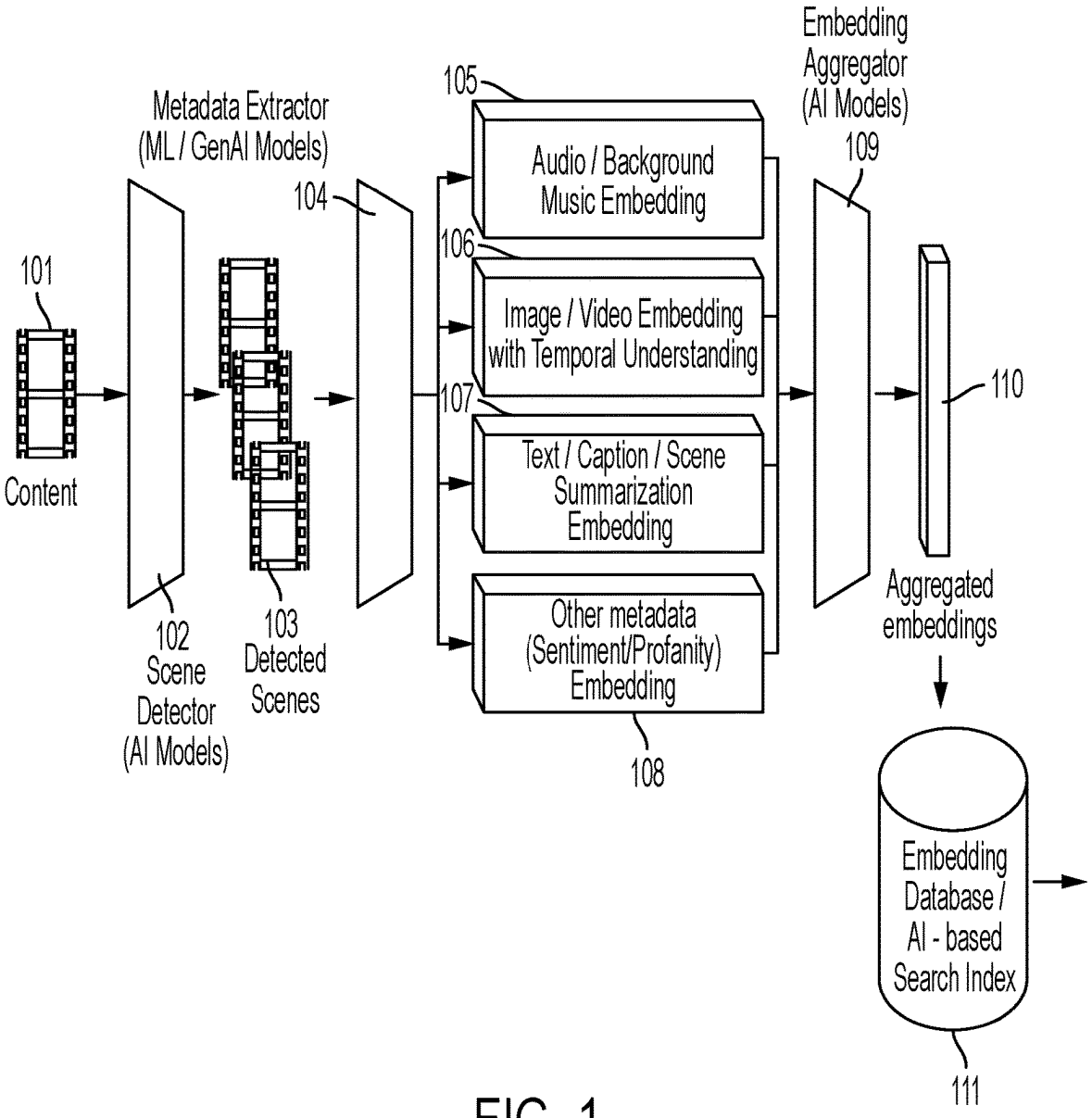

Content
101
Metadata Extractor
(ML / GenAI Models)
102
Scene
Detector
(AI Models)
103
Detected
Scenes
104
105
Audio / Background
Music Embedding
106
Image / Video Embedding
with Temporal Understanding
107
Text / Caption / Scene
Summarization
Embedding
Other metadata
(Sentiment/Profanity)
Embedding
108
Embedding
Aggregator
(AI Models)
109
110
Aggregated
embeddings
Embedding
Database /
AI - based
Search Index
111

FIG. 1

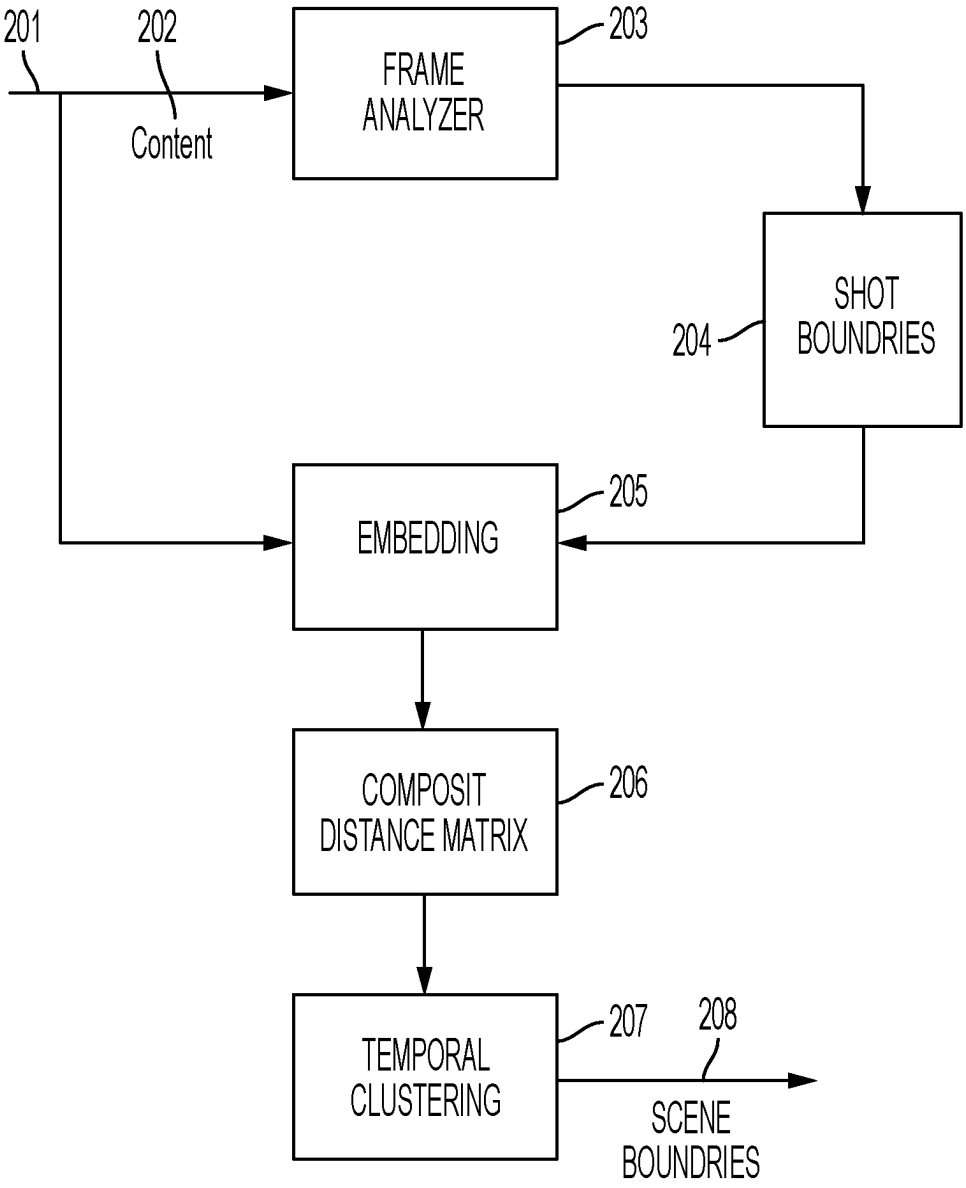

201
202
Content
FRAME ANALYZER
203
SHOT BOUNDRIES
204
EMBEDDING
205
COMPOSIT DISTANCE MATRIX
206
TEMPORAL CLUSTERING
207
208
SCENE BOUNDRIES

FIG. 2

MULTI-MODAL METADATA EXTRACTION SYSTEM

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is related to U.S. application Ser. No. _____ filed on _____, 2023, attorney docket no. 169004; U.S. application Ser. No. _____ filed on _____, 2023, attorney docket no. 169005; U.S. application Ser. No. _____ filed on, _____, 2023, attorney docket no. 169006; U.S. application Ser. No. _____ filed on _____, 2023, attorney docket no. 169007; U.S. application Ser. No. _____ filed on _____, 2023, attorney docket no. 169008; U.S. application Ser. No. _____ filed on _____, 2023, attorney docket no. 169009; and U.S. application Ser. No. _____ filed on _____, 2023, attorney docket no. 169010; the disclosures of all of which are incorporated by reference herein.

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0002] The invention relates to content processing and more particularly multi-modal metadata extraction from content.

2. Description of the Related Technology

[0003] Online advertising is a form of marketing and advertising that uses the Internet to promote products and services to audiences and platform users. Advertisements are increasingly being delivered via automated software systems operating across multiple websites, media services, and platforms, known as programmatic advertising.

[0004] Online advertising often involves a publisher, who integrates advertisements into its online content, and an advertiser, who provides the advertisements to be displayed on the publisher's content. Other potential participants include advertising agencies that help generate and place the ad copy, and an ad server that delivers and tracks the advertising activity.

[0005] Many common online advertising practices are controversial and, as a result, have become increasingly subject to regulation. Many internet users also find online advertising disruptive and have increasingly turned to ad blocking for a variety of reasons. Online ad revenues also may not adequately replace other publishers' revenue streams.

[0006] Display advertising conveys its advertising message visually using text, logos, animations, videos, photographs, or other graphics. The goal of display advertising is to obtain more traffic, clicks, or popularity for the advertising brand or organization. Display advertisers frequently target users to increase the ads' effect.

[0007] Web banners or banner ads typically are graphical ads displayed within a web page. Many banner ads are delivered by a central ad server.

[0008] The online advertising process may involve many parties. In the simplest case, the website publisher selects and serves the ads. Publishers which operate their own advertising departments may use this method. Alternatively, ads may be outsourced to an advertising agency, and served from the advertising agency's servers or ad space may be offered for sale in a bidding market using an ad exchange and real-time bidding, known as programmatic advertising.

[0009] Programmatic advertising involves automating the sale and delivery of digital advertising on websites and platforms via software rather than direct human decision-making. Advertisements are selected and targeted to audiences via ad servers which often use cookies, which are unique identifiers of specific computers, to decide which ads to serve to a particular consumer. Cookies can track whether a user left a page without buying anything, so the advertiser can later retarget the user with ads from the site the user visited.

[0010] As advertisers collect data across multiple external websites about a user's online activity, they can create a detailed profile of the user's interests to deliver even more targeted advertising. This aggregation of data is called behavioral targeting. Advertisers also target their audience by using contextual cues to deliver ads related to the content of the web page where the ads appear. Retargeting, behavioral targeting, and contextual advertising all are designed to increase an advertiser's return on investment over untargeted ads.

[0011] Customer information is combined and returned to the supply-side platform creates and provides ad offers to an ad exchange. The ad exchange puts the offer out for bid to demand-side platforms. Demand-side platforms act on behalf of ad agencies that sell ads. Demand-side platforms have ads ready to display and are searching for users to view them. Bidders get the information about the user ready to view the ad and decide, based on that information, how much to offer to buy the ad space. An ad exchange picks the winning bid and informs both parties. The ad exchange then passes the link to the ad back through the supply side platform and the publisher's ad server to the user's browser, which then requests the ad content from the agency's ad server.

[0012] Interstitial ads: An interstitial ad displays before a user can access requested content, sometimes while the user is waiting for the content to load. Interstitial ads are a form of interruption marketing.

[0013] Content marketing is any marketing that involves the creation and sharing of media and publishing content in order to acquire and retain customers. This information can be presented in a variety of formats, including blogs, news, videos, white papers, e-books, infographics, case studies, how-to guides, and more.

[0014] Ad blocking, or ad filtering is a technology that may be used to block advertising.

[0015] An online advertising network or ad network is a company that connects advertisers to websites that want to host advertisements. The key function of an ad network is an aggregation of ad supply from publishers and matching it with the advertiser's demand. The phrase "ad network" by itself is media-neutral in the sense that there can be a "Television Ad Network" or a "Print Ad Network" but is increasingly used to mean "online ad network" as the effect of aggregation of publisher ad space and sale to advertisers is most commonly seen in the online space. The fundamental difference between traditional media ad networks and online ad networks is that online ad networks use a central ad server to deliver advertisements to consumers (ad serving), which enables targeting, tracking, and reporting of impressions in ways not possible with analog media alternatives.

[0016] Targeted networks focus on specific targeting technologies such as behavioral or contextual, that have been built into an ad server. Targeted networks specialize in using consumer clickstream data to enhance the value of the inventory they purchase. Further specialized targeted networks include social graph technologies which attempt to enhance the value of inventory using connections in social networks. Significant targeting methods include behavioral targeting, contextual targeting, and creative optimization by using experimental or predictive methods to explore the optimum creative for a given ad placement and exploiting that determination in further impressions.

[0017] Artificial intelligence (AI) is the intelligence of machines or software, as opposed to the intelligence of human beings or animals.

[0018] Machine learning is the study of programs that can improve their performance on a given task automatically. It has been a part of AI from the beginning.

[0019] There are several kinds of machine learning. Unsupervised learning analyzes a stream of data, finds patterns, and makes predictions without any other guidance. Supervised learning requires a human to label the input data first and comes in two main varieties: classification (where the program must learn to predict what category the input belongs in) and regression (where the program must deduce a numeric function based on numeric input). In reinforcement learning the agent is rewarded for good responses and punished for bad ones. The agent learns to choose responses that are classified as "good". Transfer learning is when the knowledge gained from one problem is applied to a new problem. Deep learning uses artificial neural networks for these types of learning.

[0020] Natural language processing (NLP) allows programs to read, write, and communicate in human languages such as English. Specific problems include speech recognition, speech synthesis, machine translation, information extraction, information retrieval, and question answering.

[0021] Modern deep learning techniques for NLP include word embedding (how often one word appears near another), transformers (which finds patterns in text), and others. Feature detection helps AI compose informative abstract structures out of raw data.

[0022] Machine perception is the ability to use input from sensors (such as cameras, microphones, wireless signals, active lidar, sonar, radar, and tactile sensors) to deduce aspects of the world. Computer vision is the ability to analyze visual input. The field includes speech recognition, image classification, facial recognition, object recognition, and robotic perception.

[0023] Deep learning uses several layers of neurons between the network's inputs and outputs. The multiple layers can progressively extract higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits, letters, or faces.

[0024] Generative artificial intelligence (AI) is artificial intelligence capable of generating text, images, or other media, using generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics. A generative AI system is constructed by applying unsupervised or self-supervised machine learning to a data set.

[0025] The capabilities of a generative AI system depend on the modality or type of the data set used.

[0026] A foundation model (also called base model) is a large machine learning (ML) model trained on a vast quantity of data at scale (often by self-supervised learning or semi-supervised learning) such that it can be adapted to a wide range of downstream tasks. Foundation models can in turn be used for task and/or domain-specific models using targeted datasets of various kinds. Beyond text, several visual and multimodal foundation models have been produced—including DALL-E, Flamingo, Florence, and NOOR. Visual foundation models (VFMs) have been combined with text-based LLMs to develop sophisticated task-specific models. There is also Segment Anything by Meta AI for general image segmentation. For reinforcement learning agents, there is GATO by Google DeepMind.

[0027] Foundation models may be further developed through additional training. A foundation model is a "paradigm for building AI systems" in which a model trained on a large amount of unlabeled data can be adapted to many applications. Foundation models are "designed to be adapted (e.g., finetuned) to various downstream cognitive tasks by pre-training on broad data at scale".

[0028] Key characteristics of foundation models are emergence and homogenization. Because training data is not labeled by humans, the model emerges rather than being explicitly encoded. Properties that were not anticipated can appear. For example, a model trained on a large language dataset might learn to generate stories of its own or to do arithmetic, without being explicitly programmed to do so. Furthermore, these properties can sometimes be hard to predict beforehand due to breaks in downstream scaling laws. Homogenization means that the same method is used in many domains, which allows for powerful advances but also the possibility of "single points of failure".

SUMMARY OF THE INVENTION

[0029] It is an object to provide a system that is computationally efficient yet allows for a deep understanding of content. It is an object to provide a system that is computationally efficient yet allows for a deep understanding of video content.

[0030] It is an object to provide a versatile system that may be used in different applications where machine understanding of content and video content is useful or required.

[0031] It is an object to provide a system capable of indexing content, including video content, according to multiple domains. It is a further object to provide a system where the indexing of video content may be on a scene-by-scene basis and/or a frame-by-frame basis.

[0032] It is an object to utilize artificial intelligence (AI) techniques to generate the index of video content. It is a further object to provide a system that extracts scene and frame-level detail from video content, audio content, and/or other content as the content is streamed. It is a further object to provide a system that extracts scene and frame-level detail from video content and/or other content for storage in an embedding database in the form of AI indexing vectors. According to an advantageous feature, a multimodal metadata extraction system may be provided with a scene detector having a video content input and an output representing scene boundaries. The metadata extractor may be responsive to the content of a scene as identified by the scene bound-

aries to extract metadata corresponding to several, plural, or multiple extraction modes. A metadata embedding may be used for each of the modes.

[0033] An embedding aggregator response to the embedding operates to formulate an aggregated embedding for each scene thereby indexing the content of the scene. The output representing the identified scenes may be a set of video clips of each scene or an index to the video content corresponding to the identified scenes.

[0034] The scene detector may include a frame analyzer for identifying consecutive frames having similar characteristics. A boundary detector may be provided to identify boundaries of consecutive frames having sufficiently similar characteristics that they likely belong to the same shot. An embedding system may be provided to formulate a composite distance matrix capturing the distance between shot embeddings. A temporal clustering system may be connected to the composite distance matrix. An output of the temporal clustering system identifies the scene boundaries of the content.

[0035] An embedding database may be connected to the embedding aggregator for storing the aggregated embedding for use as a search index for scenes identified in the content.

[0036] The multimodal metadata extraction system may be provided with extraction modes to adequately characterize the content. The extraction modes and number of extraction modes may be in accordance with the application for which the metadata will be used. Extraction modes include at least one of Audio (speech recognition, music recognition); Image recognition (feature recognition with temporal understanding); Text (caption, scene summarization, text recognition); and scene interpretation (sentiment, profanity, action level). Many other extraction modes may be implemented.

[0037] Various other objects, features, aspects, and advantages of the disclosed system will become more apparent from the following detailed description of preferred embodiments of the invention, along with the accompanying drawings in which the same numerals are repeated for similar components.

[0038] Moreover, the above objects and advantages are illustrative, and not exhaustive, of those that can be achieved by the or with the system. Thus, these and other objects and advantages will be apparent from the description herein, both as embodied herein and as modified in view of any variations that will be apparent to those skilled in the art.

BRIEF DESCRIPTION OF THE DRAWINGS

[0039] FIG. **1** shows a multimedia metadata extraction system.

[0040] FIG. **2** shows the operation of a scene detector **102**.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0041] Before the present invention is described in further detail, it is to be understood that the invention is not limited to the particular embodiments described, as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting, since the scope of the present invention will be limited only by the appended claims.

[0042] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention.

[0043] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can also be used in the practice or testing of the present invention, a limited number of the exemplary methods and materials are described herein.

[0044] It must be noted that as used herein and in the appended claims, the singular forms "a", "an", and "the" include plural referents unless the context clearly dictates otherwise.

[0045] All publications mentioned herein are incorporated herein by reference to disclose and describe the methods and/or materials in connection with which the publications are cited. The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates, which may need to be independently confirmed.

[0046] A system is provided for processing video content to gain a rich understanding of the video content. In order to effectively process the content and achieve sufficient computational efficiency, even using artificial intelligence (AI) techniques, a content stream may be divided into scenes made up of one or more segments of the content. Each segment is likely to correspond to a shot and is made up of one or more sequential frames having a high level of commonality. Two or more segments having a high level of commonality may be grouped together and processed as a single scene.

[0047] In content production, a "shot" is typically considered to be a continuous view captured by a single camera without interruption. A processor can identify continuous frames that are likely to be in the same groups of frames in a shot by examining local color distribution. These shots are identified as a segment of content. Similar shots (or segments) may be grouped into scenes. Similar shots are taken to be part of a scene. Shots having sufficient similarity in a scene are assumed to convey a homogeneous storyline or concept.

[0048] FIG. **1** shows a multimodal metadata extraction system with a video content input **101**. The content input **101** is provided to a scene detector **102**. The scene detector **102** operates to break a video content stream to smaller (or shorter) scenes. A video stream is made up of a series of frames. Frames of content can be grouped into segments based on commonality. Segments can also be grouped into scenes based on commonality.

[0049] FIG. **2** shows the operation of a scene detector **102**. A content stream is provided to a stream analyzer for performing a frame-by-frame analysis to identify boundary frames for a series of consecutive frames having a high level of similarity. The frame-by-frame analysis may be performed by using significant average color distribution differences between consecutive frames. The shot boundaries may be stored in a boundary table **204** and used to access the frames of a shot. The video content may be in content storage **206**. Alternatively, the frames of a shot may be processed in a stream.

[0050] The frames of the shots are provided to the embedding system **205**. The embedding system may be implemented using a convolutional neural network or a Vision Transformer based on a Deep Learning image featurizer. The embedding system may generate a composite distance matrix **206** by capturing the distance between shot embedding based on a distance metric and potentially the temporal distance between shots.

[0051] Temporal clustering **207** based on dynamic programming is applied on the composite distance matrix **206** to group similar, shots together to obtain scene boundaries **208**.

[0052] The scene boundaries **208** define the detected scenes **103**. The detected scenes **103** are provided to a metadata extractor **104**. The metadata extractor **104** considers the content of the scenes individually according to selected aspects anticipated to be potentially present in the content. FIG. **1** illustrates four aspects for processing and embedding. The aspects illustrated in FIG. **1** are examples, Audio/Background Music embedding **105**, Image/Video embedding with temporal understanding **106**, Text/Caption/Scene Summation embedding **107**, and other metadata (sentiment profanity) **108**. In practice, many more modes are contemplated. For example, location, time of day, weather, genre, etc.

[0053] The extraction frame level detail may include objects, logos, locations, sentiment, action detection, scene summarizer, etc. All of the information is then encoded using an embedding model for every scene and a vector search index for each scene is then built. This allows for free-form, contextual, and detailed video indexing/searches for example the metadata for "a romantic scene with a glass of wine by a lake" can be easily identified.

[0054] The embeddings are provided to an embedding aggregator **109** to generate aggregated embeddings **110**. The aggregated embeddings may be stored in an embedding database **111**.

[0055] The techniques, processes and apparatus described may be utilized to control operation of any device and conserve use of resources based on conditions detected or applicable to the device or otherwise made available for further processing.

[0056] The system is described in detail with respect to preferred embodiments, and it will now be apparent from the foregoing to those skilled in the art that changes and modifications may be made without departing from the invention in its broader aspects, and the invention, therefore, as defined in the claims, is intended to cover all such changes and modifications that fall within the true spirit of the invention.

[0057] Thus, specific apparatus for and methods of metadata extraction have been disclosed. It should be apparent, however, to those skilled in the art that many more modifications besides those already described are possible without departing from the inventive concepts herein. The inventive subject matter, therefore, is not to be restricted except in the spirit of the disclosure. Moreover, in interpreting the disclosure, all terms should be interpreted in the broadest possible manner consistent with the context. In particular, the terms "comprises" and "comprising" should be interpreted as referring to elements, components, or steps in a non-exclusive manner, indicating that the referenced elements, components, or steps may be present, or utilized, or combined with other elements, components, or steps that are not expressly referenced.

1. A multimodal metadata extraction system comprising:
- a scene detector having a video content input and an output representing scene boundaries;
- a metadata extractor responsive to content of a scene as identified by said scene boundaries to extract metadata corresponding to several extraction modes;
- a metadata embedding for each extraction mode; and
- an embedding aggregator response to said metadata embedding is to formulate aggregated embedding for each scene indexing said content.

2. The multimodal metadata extraction system according to claim **1** wherein said output representing identified scenes is a set of video clips in each scene.

3. The multimodal metadata extraction system according to claim **1** wherein said output representing identified scenes is an index to said video content corresponding to said identified scenes.

4. The multimodal metadata extraction system according to claim **1** wherein said scene detector further comprising:
- a frame analyzer for identifying consecutive frames having similar characteristics,
- a boundary detector identifying boundaries of consecutive frames having such similar characteristics responsive to said frame analyzer;
- an embedding system formulating a composite distance matrix capturing distance between shot embedding; and
- a temporal clustering system connected to said composite distance matrix and having an output identifying scene boundaries of said content.

5. The multimodal metadata extraction system according to claim **1** further comprising an embedding database connected to said embedding aggregator for storing said aggregated embedding for use as a search index for scenes of said content.

6. The multimodal metadata extraction system according to claim **1** wherein said extraction modes include at least one of Audio (speech recognition, music recognition); Image recognition (feature recognition with temporal understanding); Text (caption, scene summarization, text recognition; and scene interpretation (sentiment, profanity, acting level).

* * * * *