



(51) International Patent Classification:

H04N 21/234 (2011.01) *G06V 10/25* (2022.01)
G06F 3/04845 (2022.01) *G06V 10/82* (2022.01)
G06T 3/40 (2006.01) *G06V 20/40* (2022.01)
G06T 11/60 (2006.01) *H04N 21/44* (2011.01)

(21) International Application Number:

PCT/US2023/025971

(22) International Filing Date:

22 June 2023 (22.06.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: **GOOGLE LLC** [US/US]; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US).

(72) Inventors: **MENA, Juan Luis Flores**; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US). **JIN, Ying**; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US). **ROSSTON, Sarah Elizabeth**; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US). **VOTEL, Ronald**; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US). **HUBLI, Priyanka Vijay**; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US).

(74) Agent: **TAM, Steven K.** et al.; LOWENSTEIN SANDLER LLP, One Lowenstein Drive, Roseland, New Jersey 07068 (US).

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: AUTOMATICALLY MODIFYING FRAME PRESENTATION CHARACTERISTICS OF A MEDIA ITEM

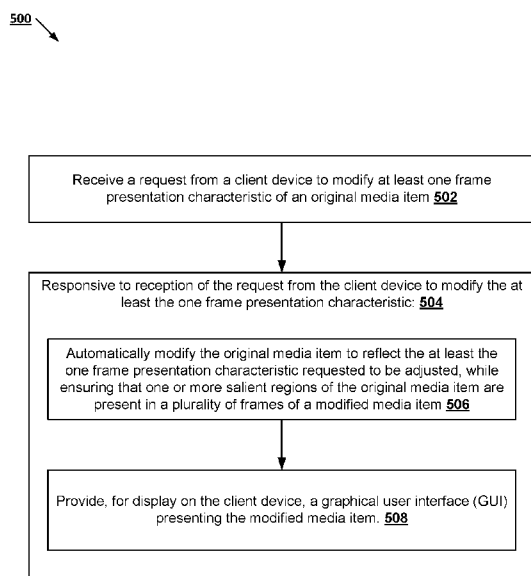


FIG. 5

(57) Abstract: Systems and methods for automatically modifying frame presentation characteristics of a media item are provided herein. a request is received from a client device to modify at least one frame presentation characteristic of an original media item. Responsive to reception of the request from the client device, the original media item is modified to reflect the at least one frame presentation characteristic requested to be adjusted, while ensuring that one or more salient regions of the original media item are present in frames of a modified media item. A user interface (UI) is provided, for display on the client device, to present the modified media item.

Published:

— *with international search report (Art. 21(3))*

AUTOMATICALLY MODIFYING FRAME PRESENTATION CHARACTERISTICS OF A MEDIA ITEM

TECHNICAL FIELD

[0001] Aspects and implementations of the present disclosure relate to automatically modifying frame presentation characteristics of a media item.

BACKGROUND

[0002] A platform (e.g., a content sharing platform) can allow users to upload, view, and share digital content such as media items. Media items can include audio clips, movie clips, music video, images and other multimedia content. For example, a user can generate a video (e.g., using a client device) and can provide the video to the platform (e.g., via the client device) to be accessible by other users of the platform. User can use computing devices (such as smart phones, cellular phones, laptop computers, desktop computers, tablet computers, televisions, and the like) to use, play, and/or otherwise consume the video and other media items (e.g., watch digital videos, and/or listen to digital music).

SUMMARY

[0003] The below summary is a simplified summary of the disclosure in order to provide a basic understanding of some aspects of the disclosure. This summary is not an extensive overview of the disclosure. It is intended neither to identify key or critical elements of the disclosure, nor to delineate any scope of the particular implementations of the disclosure or any scope of the claims. Its sole purpose is to present some concepts of the disclosure in a simplified form as a prelude to the more detailed description that is presented later.

[0004] In some implementations, a system and method are disclosed for automatically modifying frame presentation characteristics of a media item. In an implementation, a method includes receiving a request from a client device to modify at least one frame presentation characteristic of an original media item. Responsive to receiving the request from the client device to modify the at least one frame presentation characteristic of the original media item, the method further includes automatically modifying the original media item to reflect the at least one frame presentation characteristic requested to be adjusted, while ensuring that one or more salient regions of the original media item are present in frames of a modified media item. the method further includes providing, for display on the client device, a user interface (UI) presenting the modified media item.

[0005] In some embodiments, automatically modifying the original media item includes modifying an aspect ratio of the media item to facilitate creation of a short-form version of the media item. In some embodiments, automatically modifying the original media item includes providing at least a subset of frames of the original media item as input to one or more machine learning models. The one or more machine learning models are trained to predict, based on a given frame, bounding boxes for the given frame that each represent a salient region of the given frame. The method further includes obtaining outputs from the one or more machine learning models. The outputs include bounding boxes each indicating a salient region of the original media item. In some embodiments, each of the one or more machine learning models is trained to predict bounding boxes for visual features of a particular type. In some embodiments, the types of visual features include one or more of facial features, objects, frame boundary regions, or shot boundary regions.

[0006] In some embodiments, automatically modifying the original media item includes determining a distribution of saliency signals for each of the at least the subset of frames of the original media item based on the bounding boxes indicating the one or more salient regions of the media item. The method further includes determining a salient area of each of the at least the subset of the frames of the original media item based on the distribution of saliency signals. The method further includes adjusting a frame boundary of each of the at least the subset of the frames of the original media item to exclude content outside of a respective salient area and to include content within the respective salient area to obtain the frames of the modified media item.

[0007] In some embodiments, determining the salient area for each of the at least the subset of the frames of the original media item based on the distribution of saliency signals involves including, in a salient area of a respective frame of the at least the subset of the frames of the original media item, one or more portions of the respective frame that satisfy one or more threshold criteria for an amount of saliency signals; and excluding, from the salient area of the respective frame, one or more portions of the respective frame that do not satisfy one or more threshold criteria for the amount of saliency signals.

[0008] In some embodiments, the salient area for each of the at least the subset of the frames of the original media item is determined based on a distribution of motion signals between the frames of the media item. In some embodiments, the request to adjust the at least one frame presentation characteristic of the original media item is received responsive to a user interaction with one or more elements of the UI provided for display on the client device.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] Aspects and implementations of the present disclosure will be understood more fully from the detailed description given below and from the accompanying drawings of various aspects and implementations of the disclosure, which, however, should not be taken to limit the disclosure to the specific aspects or implementations, but are for explanation and understanding only.

[0010] **FIG. 1** illustrates an example system architecture, in accordance with aspects and implementations of the present disclosure.

[0011] **FIG. 2** illustrates an example user interface (UI) of an editor of a video item, in accordance with aspects and implementations of the present disclosure.

[0012] **FIG. 3** is a block diagram illustrating an example of multiple machine learning detection models used to automatically modify a frame presentation characteristic of a media item, in accordance with aspects and implementations of the present disclosure.

[0013] **FIG. 4A** is an example of an original frame of a video item to be modified, in accordance with aspects and implementations of the present disclosure.

[0014] **FIG. 4B** is an example of a user interface (UI) displaying a modified frame of a modified video item, in accordance with aspects and implementations of the present disclosure.

[0015] **FIG. 5** illustrates a flow diagram of an example method of automatically modifying frame presentations characteristics of a media item, in accordance with aspects and implementations of the present disclosure.

[0016] **FIG. 6** illustrates an example training engine for training and deployment of a deep neural network, in accordance with aspects and implementations of the present disclosure.

[0017] **FIG. 7** is a block diagram illustrating an exemplary computer system, in accordance with aspects and implementations of the present disclosure.

DETAILED DESCRIPTION OF THE DRAWINGS

[0018] Aspects of the present disclosure generally relate to automatically modifying frame presentation characteristics of a media item. A platform (e.g., a content sharing platform, etc.) can enable a user to access a media item (e.g., a video item, an audio item, etc.) provided by another user of the content sharing platform (e.g., via a client device connected to the content sharing platform). For example, a client device associated with a first user (e.g., a content creator) of the content sharing platform can generate the media item

and transmit the media item to the content sharing platform via a network. A client device associated with a second user of the content sharing platform can transmit a request to access the media item and the content sharing platform can provide the client device associated with the second user with access to the media item (e.g., by transmitting the media item to the client device associated with the second user, etc.) via the network.

[0019] Content sharing platforms typically allow users to upload media items (e.g., videos). When the video is uploaded to a content sharing platform, it can be uploaded in various orientations such as landscape, portrait, square, etc. with corresponding aspect ratios. Content sharing platforms can support media items with different aspect ratios, allowing users to upload media items in a suitable orientation. For example, users can upload videos to a content sharing platform in a landscape orientation as it matches an aspect ratio of most computer screens, televisions, and mobile devices when held horizontally. Landscape videos have a wider width than height and typically have a 16:9 aspect ratio. The aspect ratio refers to the proportional relationship between the width and the heights of a video frame. For example, a 16:9 aspect ratio indicates that the width the video frame is 16 units, and the height is 9 units.

[0020] In many instances, users of the content sharing platform can repurpose content uploaded to the content sharing platform in a landscape orientation (16:9 aspect ratio) to a portrait orientation (e.g., aspect ratio 9:16) to upload to another content sharing platform (e.g., a short-form content sharing platform) or to reupload to the same content sharing platform in the modified orientation. A short-form content sharing platform can refer to a platform that focuses on and/or supports brief, concise, and quickly consumable content, often with a limited duration (e.g., less than 60 seconds in length). Short-form content sharing platforms primarily display videos in a portrait format (e.g., aspect ratio 9:16) and are often designed for content viewing via a mobile device (e.g., a smartphone), where users typically hold their phone vertically while creating and consuming content. Short-form content sharing platforms are becoming increasingly popular as they capture user preferences for quick and easily-consumable digital content. Accordingly, many users repurpose content uploaded in landscape format to a portrait format for transmission and upload to a short-form video conference platform.

[0021] Conventional video editing tools can allow users to modify various frame presentation characteristics of a video item. Frame presentation characteristics can include a video orientation, an aspect ratio, a composition of visual elements, a frame location, and the like. For example, some video editing tools can allow users to modify an aspect of a video

item from a first aspect ratio (e.g., 16:9 aspect ratio) for viewing in a landscape orientation to a second aspect ratio (e.g., 9:16 aspect ratio) for viewing in a portrait orientation. However, many of these tools utilize a manual process to allow users to modify video items. For example, such tools can present a video item within a complex user interface (UI) to a user of the tool. The user can draw one or more adjustable boxes over a portion of the video time and manually modify (e.g., using an input device such as a touch screen, a cursor control device such as a mouse, or the like) the box to select a desired region of the video item with a desired second aspect ratio. The tool can modify a frame presentation of the video item according to the portion of the video item occupied by the adjustable box. Users can modify the adjustable box to ensure salient regions of the video item are maintained in the modified version of the video item. Salient regions can refer to specific elements within the video item that are identified to be included in the modified version of the video item and may be, for example, elements that stand out from their surroundings and are likely to capture a viewer's attention or piques a viewer's interest. Salient regions may, however, be any elements that are identified for inclusion in a modified video item. However, salient regions of the video item can constantly change throughout the duration of the video. Accordingly, users can parse the video on a per-frame basis and modify the adjustable boxes for each frame of the video item to ensure the salient regions are retained in the modified version of the video item. This can burden users with additional tasks and require additional computing resources to support these tasks.

[0022] Some video editing tools can modify frame presentation characteristics of a video item without using user-adjustable boxes to capture changing salient regions. However, such tools fail to dynamically capture salient regions throughout the duration of the video. For example, a user can request an original media item to be modified from a landscape orientation (e.g., 16:9 aspect ratio) to a portrait orientation (e.g., 9:16 aspect ratio). In response to the request, a conventional video editing tool can statically crop a portion of a video captured in landscape orientation (e.g., 16:9 aspect ratio) to produce a portrait orientation (e.g., 9:16 aspect ratio) version of the video. However, such a static crop of the video fails to include regions that fall outside of the static crop. Thus, the user of the tool can nevertheless adjust the modified video in portrait orientation to ensure salient regions are retained within frames of the modified video. This can cause frustration for the user and further burden the user with additional tasks that unnecessarily consume computing resources, thereby decreasing overall efficiency and increasing overall latency of the platform.

[0023] Aspects and implementations of the present disclosure address the above and other deficiencies by automatically modifying frame presentation characteristics of a media item, such as a video item, while ensuring salient regions of the video item are present in the modified version of the video item. A content sharing platform can allow users to upload, consume, share, search for, comment on, and otherwise engage with media items. A platform, such as content sharing platform, can allow users to modify frame presentation characteristics of a video item and upload the modified video item to the platform. In some embodiments, the platform can provide a user interface (UI) for presentation on a client device to allow a user to modify frame presentation characteristics of a video item for uploading and sharing on the platform. Frame presentation characteristics can include a frame rate, a resolution, an orientation (aspect ratio), a color space, a display format, and the like. For example, in response to a user interaction with a UI element of the UI, the platform can automatically modify an aspect ratio of the video item from a 16:9 aspect ratio (e.g., for landscape orientation viewing) to a 9:16 aspect ratio (e.g., for portrait orientation viewing) while ensuring salient regions of the original video item are present within the frames of the modified video item. The platform can present the modified video item within the UI provided for display on the client device for the user to upload, share, view, etc. In some embodiments, the video item is modified to convert it into a short-form version. A short-form version can refer to a brief, concise, and quickly consumable video clip of limited duration (e.g., less than 60 seconds in length). Such a video clip can be intended for content viewing via, for example, a mobile device (e.g., a smartphone) that is frequently held vertically by users, necessitating its conversion into a portrait format (e.g., aspect ratio 9:16) for better use of screen space and improved viewing experience of users.

[0024] In some embodiments, the platform can utilize one or more machine learning models to identify salient regions of the original media item. The machine learning models can be trained to predict, based on a given frame/image, bounding boxes for the given frame that represent salient regions of the given frame. A bounding box as used herein can refer to a rectangular box that surrounds an object or a salient region of an image or a video frame that defines a spatial location of the object or the salient. In some embodiments, each machine learning model is trained to predict bounding boxes that indicate visual features of a certain types. For example, the platform can use a face detection model to predict bounding boxes that indicate facial features, an object detection model to predict bounding boxes that indicate objects (humans, animal objects, inanimate objects, and the like), a boundary detection model to predict bounding boxes that indicate frame boundary regions, a shot detection model to

predict bounding boxes that indicate shot boundary regions, etc. Each bounding box can identify a salient region of a frame of the video item indicating a visual feature of the respective type.

[0025] The platform can provide frames (e.g., each frame, a subset of the frames, etc.) of the video item as input to the above-described machine learning models and obtain multiple bounding boxes for each of the frames provided as input. In some embodiments, the platform can modify the original media item based on the outputs of the machine learning models in order to ensure the salient regions indicated by the bounding boxes are present in the modified video item. For example, the platform can determine a distribution of saliency signals across frames of the video item based on the bounding boxes. The distribution of saliency signals across a particular frame of the video item can refer to a spatial arrangement of salient regions of the particular frame indicated by the bounding boxes corresponding to the particular frame. In other words, the distribution of saliency signals can illustrate an arrangement and intensity of visually prominent regions across frames of the video item. The platform can determine a salient area for each of the frames based on respective distributions of saliency signals and crop a respective frame to the respective salient area.

[0026] In some embodiments, to determine a salient area of a frame, the platform can include, in the salient area, portions of the frame that satisfy one or more threshold criteria for an amount of saliency signals. For example, the one or more threshold criteria can include a threshold amount of saliency signals. In some embodiments, a developer and/or operator associated with platform 110 can provide (e.g., via a client device) an indication of the one or more threshold criteria for the amount of saliency signals. The platform can further exclude, from the salient area, portions of the frame that do not satisfy the threshold criteria for an amount of saliency signals. In some embodiments, the salient area of a frame can correspond to a region of the frame with the greatest amount of saliency signals according to the distribution of saliency signals. For example, the salient area of the frame can correspond to a 9:16 aspect ratio region of the original media item with the greatest amount of saliency signals.

[0027] In some embodiments, a motion detection model can be used to determine salient areas of frames of a video item. For example, each portion of a particular frame can fail to satisfy one or more threshold criteria for an amount of saliency signals. Accordingly, the platform can rely on a motion detection model to determine the salient area of the particular frame in lieu of the above-described saliency detection models. The motion detection model can be a machine learning model trained to predict a presence of motion, an absence of

motion, and/or an amount of motion present in a frame of a video item. The platform can determine that a salient area of the particular frame is a region of the frame with the greatest amount of motion based on an output of the motion detection model. For example, the salient area of the particular frame can correspond to a 9:16 aspect ratio region of the original media item with the greatest amount of motion present.

[0028] To modify the video item, the platform can modify (crop) a frame boundary of frames of the video item to exclude content outside of the salient area and to include content within the salient area. After salient areas have been determined and frame boundaries have been modified, the platform can provide the modified video item for presentation at a UI of the client device for uploading, sharing, etc.

[0029] Automatically modifying frame presentation characteristics, such as a video orientation or an aspect ratio, of a media item while ensuring that salient regions of the media item are present in the modified version of the media item improves an overall user experience with the content sharing platform as users can easily reupload video items in a different orientation. In addition, aspects and implementations of the present disclosure result in more efficient use of processing resources by automatically modifying frame presentation characteristics upon a user request rather than the users repeatedly modifying frame presentation characteristics manually to capture salient regions on a per-frame basis, thereby avoiding unnecessary consumption of resources to support iterative manual modifications to obtain a suitable modified video item.

[0030] Although the description herein often refers to video as an example type of media item, it is appreciated that aspects and implementations of the present disclosure can apply to other types of media items such as images, audio, and other multi-media without deviating from the scope of the present disclosure.

[0031] **FIG. 1** illustrates an example system architecture 100, in accordance with implementations of the present disclosure. The system architecture 100 (also referred to as “system” herein) includes client devices 102A-N, a data store 110, a platform 120, and/or a server machine 150 each connected to a network 108. In implementations, network 108 can include a public network (e.g., the Internet), a private network (e.g., a local area network (LAN) or wide area network (WAN)), a wired network (e.g., Ethernet network), a wireless network (e.g., an 802.11 network or a Wi-Fi network), a cellular network (e.g., a Long Term Evolution (LTE) network), routers, hubs, switches, server computers, and/or a combination thereof.

[0032] In some embodiments, platform 120 can be a content sharing platform that allows users to consume, upload, share, search for, approve of (“like”), dislike, and/or comment on media items 112. Platform 120 can include a website (e.g., a webpage) or application back-end software used to provide a user with access to media items 112 (e.g., via client devices 102A-N). A media item 112 can be consumed via the Internet or via a mobile device application, such as a content viewer 103A-N of client device 102A-N. In some embodiments, a media item 112 can correspond to a media file (e.g., a video file, and audio file, etc.). In other or similar embodiments, a media item 112 can correspond to a portion of a media file (e.g., a portion or a chunk of a video file, an audio file, etc.). As discussed previously, a media item 112 can be requested for presentation to users of the platform by a user of the platform 120. As used herein, “media,” “media item,” “online media item,” “digital media,” “digital media item,” “content,” and “content item” can include an electronic file that can be executed or loaded using software, firmware or hardware configured to present the digital media item to an entity. In one implementation, the platform 120 can store the media items 112 using the data store 110. In another implementation, the platform 120 can store media item 112 or fingerprints as electronic files in one or more formats using data store 110. Platform 120 can provide media item 112 to a user associated with a client device (e.g., client device 102A) by allowing access to media item 112 (e.g., via a content sharing platform application), transmitting the media item 112 to the client device 102A, and/or presenting or permitting presentation of the media item 112 at a display device 103A of client device 102A.

[0033] In some embodiments, media item 121 can be a video item. A video item refers to a set of sequential video frames (e.g., image frames) representing a scene in motion. For example, a series of sequential video frames can be captured continuously or later reconstructed to produce animation. Video items can be provided in various formats including, but not limited to, analog, digital, two-dimensional and three-dimensional video. Further, video items can include movies, video clips, video streams, or any set of images (e.g., animated images, non-animated images, etc.) to be displayed in sequence. In some embodiments, a video item can be stored (e.g., at data store 110) as a video file that includes a video component and an audio component. The video component can include video data that corresponds to one or more sequential video frames of the video item. The audio component can include audio data that corresponds to the video data.

[0034] In some implementations, data store 110 is a persistent storage that is capable of storing data as well as data structures to tag, organize, and index the data. Data can include audio data and/or video data, in accordance with embodiments described herein. Data store

110 can be hosted by one or more storage devices, such as main memory, magnetic or optical storage based disks, tapes or hard drives, NAS, SAN, and so forth. In some implementations, data store 110 can be a network-attached file server, while in other embodiments data store 110 can be some other type of persistent storage such as an object-oriented database, a relational database, and so forth, that can be hosted by platform 120 or one or more different machines (e.g., server machines 130-160) coupled to the platform 120 via network 108. Data store 110 can include a media cache that stores copies of media items that are received from the platform 120. In one example, media item 112 can be a file that is downloaded from platform 120 and can be stored locally in media cache. In another example, media item 112 can be streamed from platform 120 and can be stored as an ephemeral copy in memory of one or more of server machine 130-160.

[0035] The client devices 102A-N can each include computing devices such as personal computers (PCs), laptops, mobile phones, smart phones, tablet computers, netbook computers, network-connected televisions, etc. In some implementations, client devices 102A-N can also be referred to as “user devices.” Each client device 102A-N can include a content viewer 103A-N. The content viewer 103A-N can include a web browser and/or the client application to present, on a client device 102A-N, a user interface (UI) 124A-N for users to view or upload content, such as images, video items, web pages, documents, etc. For example, the content viewer 103A-N can be a web browser that can access, retrieve, present, and/or navigate content (e.g., web pages such as Hyper Text Markup Language (HTML) pages, digital media items, etc.) served by a web server. The content viewer 103A-N can render, display, and/or present the content to a user. The content viewer 103A-N can also include an embedded media player (e.g., a Flash® player or an HTML5 player) that is embedded in a web page (e.g., a web page that can provide information about a product sold by an online merchant). In another example, the content viewer 103A-N can be a standalone application (e.g., a mobile application or app) that allows users to view digital media items (e.g., digital video items, digital images, electronic books, etc.). According to aspects of the disclosure, the content viewer 103A-N can be a content platform application for users to record, edit, and/or upload content for sharing on platform 120. As such, the content viewers 103A-N and/or the UIs 124A-N associated with the content viewers 103A-N can be provided to client devices 102A-N by platform 120. In one example, the content viewers 103A-N can be embedded media players that are embedded in web pages provided by the platform 120.

[0036] Platform 120 can include multiple channels (e.g., channels A through Z). A channel can include one or more media items 121 available from a common source or media

items 121 having a common topic, theme, or substance. Media item 121 can be digital content chosen by a user, digital content made available by a user, digital content uploaded by a user, digital content chosen by a content provider, digital content chosen by a broadcaster, etc. For example, a channel X can include videos Y and Z. A channel can be associated with an owner, who is a user that can perform actions on the channel. Different activities can be associated with the channel based on the owner's actions, such as the owner making digital content available on the channel, the owner selecting (e.g., liking) digital content associated with another channel, the owner commenting on digital content associated with another channel, etc. The activities associated with the channel can be collected into an activity feed for the channel. Users, other than the owner of the channel, can subscribe to one or more channels in which they are interested. The concept of "subscribing" can also be referred to as "liking," "following," "friending," and so on.

[0037] In some embodiments, system 100 can include one or more third party platforms (not shown). In some embodiments, a third party platform can provide other services associated with media items 121. For example, a third party platform can include an advertisement platform that can provide video and/or audio advertisements. In another example, a third party platform can be a video streaming service provider that provides a media streaming service via a communication application for users to play videos, TV shows, video clips, audio, audio clips, and movies, on client devices 102 via the third party platform.

[0038] In some embodiments, a client device 102 can transmit a request to platform 120 for access to a media item 121. Platform 120 can identify the media item 121 of the request (e.g., at data store 110, etc.) and can provide access to the media item 121 via the UIs 124A-N of the content viewers 103A-N provided by platform 120. In some embodiments, the requested media item 121 can have been generated by another client device 102A-N connected to platform 120. For example, client device 102A can generate a video item (e.g., via an audiovisual component, such as a camera, of client device 102A) and provide the generated video item to platform 120 (e.g., via network 108) to be accessible by other users of the platform. In other or similar embodiments, the requested media item 121 can have been generated using another device (e.g., that is separate or distinct from client device 102A) and transmitted to client device 102A (e.g., via a network, via a bus, etc.). Client device 102A can provide the video item to platform 120 (e.g., via network 108) to be accessible by other users of the platform, as described above. Another client device, such as client device 102B, can transmit the request to platform 120 (e.g., via network 108) to access the video item provided by client device 102A, in accordance with the previously provided examples.

[0039] In some embodiments, platform 120 can include a user interface (UI) manager 161. The UI manager 161 can generate and provide a UI for display at a client device 102. The UI manager can provide a UI to allow users of the platform 120 to edit media items 112 using one or more UI elements included within the UI. For example, the UI manager 161 can provide a UI for display at a client device 102A that includes a UI element to automatically adjust a composition of the media item 112. Responsive to receiving an indication of a user interaction with the UI element of the displayed UI, UI manager 161 can transmit a command causing media item modifier 151 to automatically modify frame presentation characteristics (e.g., an aspect ratio) of the media item 112, as described in greater detail below. Server machine 160 can additionally or alternatively include UI manager 161.

[0040] Training data generator 131 (e.g., residing at server machine 130) can generate training data to be used to train machine learning models 160A-N. Models 160A-N can include machine learning models used or otherwise accessible to media item modifier 151 (e.g., to generate saliency signals). In some embodiments, training data generator 131 can generate the training data based on video frames of training media items and/or training images (e.g., stored at data store 110 or another data store connected to system 100 via network 108) and/or data associated with one or more client devices that accessed the training media items.

[0041] Server machine 140 can include a training engine 141. Training engine 141 can train machine learning models 160A-N using the training data from training data generator 131. In some embodiments, the machine learning models 160A-N can refer to model artifacts created by the training engine 141 using the training data that includes training inputs and corresponding target outputs (correct answers for respective training inputs). The training engine 141 can find patterns in the training data that map the training input to the target output (the answer to be predicted), and provide the machine learning models 160A-N that captures these patterns. The machine learning models 160A-N can be composed of, e.g., a single level of linear or non-linear operations (e.g., a Convolutional Neural Network (CNN) or other deep network, e.g., a machine learning model that is composed of multiple levels of non-linear operations). An example of a deep network is a neural network with one or more hidden layers, and such a machine learning model can be trained by, for example, adjusting weights of a neural network in accordance with a backpropagation learning algorithm or the like. In other or similar embodiments, the machine learning models 160A-N can refer to model artifacts that are created by training engine 141 using training data that includes training inputs. Training engine 141 can find patterns in the training data, identify clusters of

data that correspond to the identified patterns, and provide the machine learning models 160A-N that captures these patterns. Machine learning models 160A-N can use one or more of clustering, supervised machine learning, semi-supervised machine learning, unsupervised machine learning, k-nearest neighbor algorithm (k-NN), linear regression, random forest, neural network (e.g., artificial neural network), a boosted decision forest, etc.

[0042] In some embodiments, machine learning models 160A-N can be trained to predict, based on a given image or frame, such as a frame of a media item 121, bounding boxes for the given frame that each indicate a salient region of the given frame. In some embodiments, each of the machine learning models 160A-N can be trained to predict bounding boxes of a certain type. For example, machine learning model 160A can be an object detection model to predict bounding boxes that indicate objects. Machine learning model 160B can be a face detection model that predicts bounding boxes that indicate facial features. Machine learning model 160C can be a boundary detection model to predict bounding boxes that indicate frame boundary regions. Machine learning model 160D can be a shot detection model to predict bounding boxes that indicate shot boundary regions. Each bounding box output from the machine learning models 160A-D can identify a salient region of a frame of the media item 121 indicating a visual feature of the respective type.

[0043] Server machine 150 can include media item modifier 151. Media item modifier 151 can dynamically (e.g., immediately or not more than a threshold number of seconds after a modification request) modify a frame presentation characteristic of a media item 121 to produce a modified media item while ensuring salient regions the media item 121 are retained within the modified media item. Media item modifier 151 can provide a subset of frames of media item 112 as input to one or more trained machine learning models 106A-N. The media item modifier 151 can modify the media item 121 based on the outputs of the machine learning models in order to ensure the salient regions indicated by the bounding boxes are present in a modified video item. In some embodiments, to analyze the salient regions indicated by the bounding box, the media item modifier 151 can generate a distribution of saliency signals across frames of the video item based on the bounding boxes. The distribution of saliency signals across a particular frame of the video item can refer to a spatial arrangement of salient regions of the particular frame indicated by the bounding boxes corresponding to the particular frame. The media item modifier 151 can modify an aspect ratio of the media item and crop frames of the media item 121 to the modify aspects ration and ensure salient regions remain present within the cropped frames based on the distribution of saliency signals, as described in detail below with respect to **FIG. 4A-B**.

[0044] It should be noted that although **FIG. 1** illustrates media item modifier 151 and user interface (UI) controller 161 as part of platform 120, in additional or alternative embodiments, media item modifier 151 and UI manager 161 can reside on one or more server machines that are remote from platform 120 (e.g., server machine 150, server machine 160). It should be noted that in some other implementations, the functions of server machines 130, 140, 150, 160 and/or platform 120 can be provided by a fewer number of machines. For example, in some implementations, components and/or modules of any of server machines 130, 140, 150, 160 can be integrated into a single machine, while in other implementations components and/or modules of any of server machines 130, 140, 150, 160 can be integrated into multiple machines. In addition, in some implementations, components and/or modules of any of server machines 130, 140, 150, 160 can be integrated into platform 120.

[0045] In general, functions described in implementations as being performed by platform 120 and/or any of server machines 130, 140, 150, 160 can also be performed on the client devices 102A-N in other implementations. In addition, the functionality attributed to a particular component can be performed by different or multiple components operating together. Platform 120 can also be accessed as a service provided to other systems or devices through appropriate application programming interfaces, and thus is not limited to use in websites.

[0046] Although implementations of the disclosure are discussed in terms of platform 120 and users of platform 120 accessing a video item, implementations can also be generally applied to media items generally. Further, implementations of the disclosure are not limited to content sharing platforms that allow user to generate, share, view, and otherwise consume media items such as video items.

[0047] In implementations of the disclosure, a “user” can be represented as a single individual. However, other implementations of the disclosure encompass a “user” being an entity controlled by a set of users and/or an automated source. For example, a set of individual users federated as a community in a social network can be considered a “user.” In another example, an automated consumer can be an automated ingestion pipeline of platform 120.

[0048] Further to the descriptions above, a user can be provided with controls allowing the user to make an election as to both if and when systems, programs, or features described herein can enable collection of user information (e.g., information about a user’s social network, social actions, or activities, profession, a user’s preferences, or a user’s current location), and if the user is sent content or communications from a server. In addition, certain

data can be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity can be treated so that no personally identifiable information can be determined for the user, or a user's geographic location can be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user can have control over what information is collected about the user, how that information is used, and what information is provided to the user.

[0049] FIG. 2 illustrates an example user interface (UI) 200 of an editor of a video item, in accordance with some embodiments of the present disclosure. Users of a content sharing platform (e.g., platform 120 of FIG. 1) can interact with the UI 200 to modify frame presentation characteristics, recompose, crop, or otherwise edit media items (e.g., media items 121 of FIG. 1). The UI 200 can be generated by one or more processing devices of server machine 130, server machine 140, server machine 150, and/or server machine 160. In some embodiments, the UI 200 can be generated by user interface (UI) manager 161 of FIG. 1, for presentation at a client device (e.g., client devices 102A-102N.) The UI manager 161 can provide the UI 200 to enable users of the content sharing platform to cause a presentation characteristic (e.g., a composition, an aspect ratio, etc.) of a media item of the platform 120 to be modified based on a user interaction with the UI 200. In some embodiments, the UI 200 can be generated by a client device (e.g., content viewer application 103 on client device 102).

[0050] UI 200 can include multiple regions, including a region 210 to display one or more media items (e.g., video items) corresponding to video data captured and/or streamed by client devices, such as client devices 102A-102N of FIG. 1. The UI 200 can further include a region 220 for providing selectable layouts and options to adjust a composition of the video item. The region 220 includes a UI element 222, a UI element 224, a UI element 226, a UI element 228, and a UI element 230. Responsive to a user interaction with one of the UI elements displayed with the region 220, one or more components (e.g., media item modifier 151 of FIG. 1) of the platform can modify presentation characteristics of the video item to obtain a modified video item. A UI manager, such as UI manager 161 of FIG. 1, can provide the modified video item for display at region 210 of the UI 200.

[0051] In an illustrative example, in response to a user interaction with UI element 222 of the region 220, a processing device (e.g., media item modifier 151 of FIG. 1) can automatically adjust a composition of the video item to obtain a modified video item. To automatically adjust the composition of the video item, the processing device can analyze one

or more frames of video item to determine salient regions of the video item. Salient regions can refer to meaningful or important regions for the video item or a particular frame of the video item. For example, a region of interest can include specific elements within a frame of the video item that capture a viewer's attention. The processing device can automatically modify a composition of the video item while maintaining focus on one or more salient regions. In some embodiments, the processing device can modify an aspect ratio of the video item to a target aspect ratio as indicated by the user. For example, the processing device can adjust the aspect ratio of the media item from a 16:9 aspect ratio (landscape) to a 9:16 aspect ratio (portrait). In some embodiments, to determine salient regions, the processing device can provide one or more frames of the video item as input to one or more computer vision detectors or machine learning models, such as models 160A-N of **FIG. 1**, that are trained to identify specific objects, patterns, or features within a given video frame or image. The outputs of the one or more machine learning models can be indicative of the salient regions of the video item. In some embodiments, by way of example and not by way of limitation, the one or more machine learning models can be a shot boundary detector, a saliency detector, a motion detector, a face detector, and a border detector. The processing device can analyze outputs of the one or more machine learning models to determine salient regions for individual frames of the video item, as described in detail below with respect to **FIG. 3**. The processing device can automatically modify the composition (e.g., the aspect ratio) of the video item while ensuring one or more of the salient regions of the video item are present in the modified video item. The UI 200 can be updated to display the modified video item at the region 210.

[0052] In some embodiments, the region 210 can be split to display multiple regions of the video item. For example, the region 210 can be split to display a first portion of the video item at an upper portion of the region 210 and a second portion of the video item at a lower portion of the region 210. In an illustrative example, a first area of a video item can be dedicated to display a view of a content creator captured via a webcam, a camera, or the like. A second area of the video item can be dedicated to display other visual content shared by the content creator such as a gameplay footage, content produced by another content creator, or the like. The processing device can identify a first set of salient regions associated with the first area of the video item and a second set of salient regions associated with the second area of the video item. The processing device can modify a composition of the first area and the second area according to the first set of regions of interest and second set of regions of interest respectively to obtain a first modified area of the video item and second modified

area of the video item. The first modified area of the video can be provided for display at the lower portion of the region 210 and the second modified area of the video can be provided for display at the upper portion of the region 210.

[0053] In some embodiments, UI element 224, UI element 226, UI element 228, and UI element 230 can be interactable UI elements that, in response to a user interaction, provide one or more pre-defined layouts of region 210 for a user to manually adjust and display a modified video item. For example, in response to a user interaction with UI element 224, the processing device can modify an aspect ratio of the video item from a first aspect ratio (e.g., 16:9) to a second aspect ratio (e.g., 9:16). To fit the narrower second aspect ratio, the processing device can remove content from the sides of the video, thereby reducing the width of the video to fit within the second aspect ratio. In some embodiments, the processing device can provide a window over an area of the video where content within the window is retained and content outside of the window is removed. In some embodiments, the window is a static window. For example, the processing device can provide a static window that remains at a center portion video item for each frame of the video. In some embodiments, the window can be presented within the region 210 of the video UI 200 and a user can manually modify (e.g., using an input device such as a touch screen, a cursor control device such as a mouse, or the like) the window to capture a desired region of the video item to displaying within region 210 on a per-frame basis in order to ensure regions of interest are maintained in the modified composition of the video item.

[0054] In another example, in response to a user interaction with UI element 226, the processing device can cause an upper region of the region 210 to display a first window of a first area of the video item and lower region of the region 210 to display a second window of the second area of the video item. In some embodiments, a user can manually adjust (e.g., using an input device such as a touch screen, a cursor control device such as a mouse, or the like) the first window and the second window to capture a desired first area and second area of the video item. The UI element 228 and the UI element 230 can provide similar functionality using different predefined layouts. For example, in response to a user interaction with UI element 226, the processing device can cause an upper region of the region 210 to display the first window of the first area of the video item and lower region of the region 210 to display the second window of the second area of the video item. In response to a user interaction with UI element 230, the processing device can cause an upper region of the region 210 to display a first window of a first area of the video item and lower region of the

region 210 to display a second window of the second area of the video item, where the upper region is larger than the lower region.

[0055] FIG. 3 is a block diagram illustrating use of multiple machine learning detection models (also referred to as “machine learning detectors” herein) to automatically modify a presentation characteristic of a media item, in accordance with implementations of the present disclosure. Specially, FIG. 3 illustrates an example trained shot boundary detection model 310, object detection model 312, motion detection model 314, face detection model 316, and border detection model 318. In some embodiments, the detection models can provide bounding boxes as output that represent a salient region of an image/video frame provided as input. In some embodiments, the object detection models can be trained to predict bounding boxes for visual features of a particular type. For example, the object detection model 312 can be trained to predict bounding boxes for objects, the face detection model 316 can be trained to predict bounding boxes for facial features, and the border detection model 318 can be trained to predict bounding boxes for frame boundary regions (borders). Each of the detection models can correspond to a respective model 160A-N of FIG. 1. For example, the shot boundary detection model 310 can correspond to model 160A, the object detection model 312 can correspond to model 160B, and so forth. As illustrated, a video item 302 can be provided as input to the shot boundary detection model 310, the object detection model 312, the motion detection model 314, the face detection model 316, and the border detection model 318. In some embodiments, the video item 302 can correspond to media item 121 of FIG. 1. The detection models can be composed of a convolutional neural network (CNN), a recurrent neural network (RNN), or other deep learning network.

[0056] The shot boundary detection model 310 can be coupled to receive the video item 302 as input. The shot boundary detection model 310 can be trained to predict shot boundaries. Shot boundary detection is a computer vision technique to identify transitions or boundaries between consecutive shots in a video sequence such as video item 302. A shot in the video item 302 refers to a continuous sequence of frames captured by a single camera without interruption. By accurately detecting shot boundaries, the video item 302 can be segmented into individual shots, allowing for further analysis of the individual shots. The shot boundary detection model 310 can be trained on historical data such as videos and labeled shot boundaries. Visual and/or audio features can be extracted from each video frames of the historical videos that represent the content of each of the frames. The features (e.g., video and/or audio features) and corresponding labeled shot boundaries can be used to train the shot boundary detection model 310. After training and deployment of the shot

boundary detection model 310, the shot boundary detection model 310 can process each frame of the video item 302 and predict shot boundaries of the video item 302 based on learned patterns from the training data. The shot boundary detection model 310 can identify one or more bounding boxes indicative of one or more shot boundaries of the video item 302 as output to the framer 322.

[0057] It is appreciated that various other shot boundary detection computer vision techniques and signal processing algorithms can be employed to determine shot boundaries without deviating from the scope of the present disclosure. For example, a pixel-based approach can calculate pixel metrics such as pixel intensity or pixel color to identify shot boundaries based on threshold differences of pixel metrics. In another example, a histogram-based analysis can compare a color distribution between consecutive frames to determine shot boundaries. In yet another example, a motion-based approach can analyze motion vectors between frames to identify shot boundaries. In still another example, temporal analysis can use various metrics such as mean squared error (MSE) or structural similarity index (SSIM) to quantify a similarity level or a difference level between consecutive frames indicative of a shot boundaries. It can be noted that one or more of the above-reference approaches can be used individually or in combination to identify shot boundaries of the video item 302.

[0058] The object detection model 312 can be coupled to receive the video item 302 as input. The object detection model 312 can be a machine learning model trained to predict visually salient regions or objects in a frame of a video item 302. A salient region or object can be a region or object that stands out from its surroundings in the frame and is likely to capture a viewer's attention. The object detection model 312 can be trained on historical data such as videos and labeled data indicating salient regions and/or objects. Training data (historical data) can be labeled using ground truth data indicating salient regions or objects of the historical videos. The ground truth data can be established by human observers or by eye-tracking data where eye movements of human observers are recorded while viewing the historical videos. After training and deployment of the object detection model 312, the object detection model 312 can process one or more frames of the video item 302 and predict salient regions and/or objects of the video item 302 based on learned patterns from the training data. The object detection model 312 can identify one or more bounding boxes indicative of one or more salient regions/objects for each processed frame of the video item 302 as output to the framer 322.

[0059] The motion detection model 314 can be coupled to receive the video item 302 as input. The motion detection model 314 can be a machine learning model trained to predict a presence of motion, an absence of motion, and/or an amount of motion in given video frame. The motion detection model 314 can be trained using historical data composed of frames of video items where the frames are labeled with a quantitative measure of a magnitude of motion present in the respective frames. Once the motion detection model 314 is trained and deployed, during an inference process, a set of frames of the video item 302 can be provided as input to the motion detection model 314. The motion detection model 314 can predict a magnitude of motion present in each frame of the set of frames of the video item 302 based on learned patterns from the training data. In some embodiments, the motion detection model 314 can output data to create a motion magnitude graph (also referred to as a “distribution of motion signals” herein) for each processed frame of the video item 302. For example, the output of the motion detection model 314 can be synthesized to create a line graph corresponding to a frame of the video item 302 where the x-axis represents a column of pixels of the frame and the y-axis represents a magnitude of motion, as illustrated below with respect to **FIG. 4A**. It is appreciated that other methods of computer vision motion detection (e.g., frame differencing, pixel thresholding, etc.) can be leveraged to determine motion within frames of the video item 302 without deviating from the scope of the present disclosure.

[0060] The face detection model 316 can be coupled to receive the video item 302 as input. The face detection model 316 can be a machine learning model trained to predict (e.g., identify and locate) faces within a frame of a video item or an image. The face detection model 316 can be trained with a historical dataset of images and/or video frames. The historical dataset can be labeled with bounding boxes around facial features contained within the images/video frames of the historical dataset. After training and deployment of the face detection model 316, the face detection model 316 can process unlabeled data such as frames of the video item 302 and predict presence and location of facial features within frames of the video item 302. The face detection model 316 can identify one or more bounding boxes indicative of a location of one or more facial features within a respective frame of the video item 302.

[0061] The border detection model 318 is coupled to receive the video item 302 as input. The border detection model 318 can be a machine learning model trained to predict borders of edges of video frames/images. The border detection model 318 can be trained with a historical dataset of images and/or video frames labeled with ground truth border or edge

information. For example, border pixels of the historical training images/video frames can be marked as positive examples and non-border pixels can be marked as negative examples. After training and deployment, the border detection model 318 can process unlabeled data such as frames of the video item 302 and predict a location of a border around frames of the video item 302. The border detection model 318 can identify a bounding box indicative of a location of a border within a respective frame of the video item 302. It is appreciated that various other methods and approaches can be used, such as a gradient-based approach, to detect borders of the video item 302 without deviating from the scope of the present disclosure.

[0062] Each of the shot boundary detection model 310, the object detection model 312, the motion detection model 314, the face detection model 316, and the border detection model 318 can process and analyze each frame of the video item 302 or a subset of frames of the video item 302. For example, the video item 302 can be a series of sequential video frames captured at a rate of 24 individual frames per second (FPS). However, the detection models can process and analyze the video item 302 at a rate of 5 FPS. In another example, the detection models can process the video item 302 at a rate of 24 FPS (e.g., each frame of the video item 302).

[0063] The framer 322 can be coupled to receive one or more outputs of the shot boundary detection model 310, the object detection model 312, the motion detection model 314, the face detection model 316, and the border detection model 318. The framer can determine a salient area for frames of the video item 302 based on outputs (e.g., bounding boxes, motion data, etc.) obtained from the above-described detection models. For example, the framer 322 can ensure that a salient area of a frame of the video item 302 does not include a predicted border of the frame as indicated by a bounding box output from the border detection model.

[0064] In some embodiments, the framer 322 can combine outputs from one or more of the detection models to generate a distribution of saliency signals across frames of the video item 302. The saliency signals can indicate visual saliency or importance of specific regions within video frames based on bounding boxes provided by the detection models. In some embodiments, the detection models can be trained to predict bounding boxes for visual features of a particular type. For example, the object detection model 312 can be trained to predict bounding boxes for objects, the face detection model 316 can be trained to predict bounding boxes for facial features, and the border detection model 318 can be trained to predict frame boundary regions (e.g., borders of frames). The framer 322 can combine

bounding boxes output from detection models trained to predict bounding boxes for visual features of a particular type. For example, the framer 322 can combine bounding boxes output from the object detection model 312, and the face detection model 316, and the border detection model 318 to generate a distribution of combined saliency signals across frames of the video item 302.

[0065] In some embodiments, saliency signals from one or more of the detection models can be combined across an axis of a video frame. For example, the saliency signals can be combined across a horizontal axis of the video frame, as illustrated with respect to **FIG. 4A** in which combined saliency signals are represented. The framer 322 can analyze the distribution of the combined saliency signals across the video frame to determine a salient area of the video frame and crop the video frame to the salient area. In some embodiments, the framer 322 can make such a determination according to threshold criteria for an amount of saliency (based on the distribution of the combined saliency signals) pertaining to a particular region of the video frame. If the threshold criteria are satisfied, the framer 322 can include the particular region in the salient area. If the threshold criteria are not satisfied, the framer 322 can exclude the particular region from the salient area. The framer 322 can perform similar analysis across regions of the video frame. For example, the video frame can be divided into uniform columns of pixels where each column of pixels corresponds to a portion of the distribution of the combined saliency signals. The framer 322 can analyze each column of pixels and corresponding saliency signals and include, in the salient area of the video frame, each column of pixels that satisfies the threshold criteria. For example, a saliency score for each column having a width of a predetermined number of pixels may be determined by combining the saliency score associated with each pixel width of the column. The framer 322 can exclude, from the salient area of the video frame, each column of pixels that does not satisfy the threshold criteria. In some embodiments, the salient area can be centered around a point of the frame with the greatest amount of saliency signals according to the distribution of saliency signals and expand outwards. Further details regarding determining a salient area of a video frame and a corresponding illustration are provided below with respect to **FIG. 4A**.

[0066] It can be noted that salient areas of frames of the video item 302 can differ from frame to frame as distributions of saliency signals can be different for each analyzed frame. Accordingly, smoother 324 can make additional modifications to frame boundaries of frames of the video item 302 to ensure smooth transitions between frames of the video item 302. For example, the smoother 324 can reduce abrupt fluctuations between frames of the video item

302 to create a more visually appealing final modified video output. In some embodiments, the smoother 324 can use one or more filters to analyze motion within salient areas of the video frames and compensate for motion within the salient areas based on the analysis. Many filters and techniques can be employed individually and in combination to smooth transitions between the modified frames of the video item 302 based on pixel characteristics of modified frame, motion between the modified frames, and the like.

[0067] At block 326, frames of the video item 302 can be modified (cropped) to a respective salient area by modifying a frame boundary of the respective video frame to exclude content outside of the respective salient area and to include content within the respective salient area to generate a video output 328. In some embodiments, the video item 302 can be padded with additional space around the modified video frame to ensure the video item fits within a display format of a client device, as illustrated below with respect to **FIG. 4B**. In some embodiments, the video output 328 can be provided for display within a user interface (UI) such as UI 200 of **FIG. 2**.

[0068] It is appreciated that the operations described above with respect to **FIG. 3** can be performed on one or more client devices, one or more servers, or any combination thereof. For example, one or more of shot boundary detection model 310, object detection model 312, motion detection model 314, face detection model 316, and border detection model 318 can be deployed and operate on one or more servers, such as server machines 130-160 of **FIG. 1**. A client device, such as client device 102A, can obtain outputs of the machine learning models from the one or more servers and perform operations associated with framer 322, smoother 324, and crop & pad 326.

[0069] **FIG. 4A** is an example of an original frame 400 of a video item to be modified, in accordance with aspects and implementations of the present disclosure. As described above, a request is received from a client device (e.g., client device 102A) to modify a frame presentation characteristic of an original media item (e.g., video item 302). For example, the request can include a request to modify an aspect ratio of the original video item from a 16:9 aspect ratio (for landscape viewing) to a 9:16 aspect ratio (for portrait viewing). The original video item can be automatically modified to reflect the frame presentation characteristic (e.g., the aspect ratio) requested to be modified while ensuring one or more salient regions of the original video item are present in the frames of the modified media item. That is, a modified frame may be generated such that the modified frame includes the one or more salient regions and has the frame presentation characteristic.

[0070] In an illustrative example, an original frame 400 of a video item illustrates multiple bounding boxes that represent salient regions of the original frame 400. In at least one embodiment, the bounding boxes are determined based on outputs from corresponding machine learning models (e.g., saliency detectors), as described above with respect to **FIG. 3**. For example, the bounding box 402 corresponds to an output from object detection model 312 which is trained to predict bounding boxes for objects and/or face detection model 316 which is trained to predict bounding boxes for facial feature. The bounding box 406 corresponds to an output from border detection model 318 which is trained to predict bounding boxes for frame boundary regions, and bounding box 408 corresponds to an output from shot boundary detection model 310 trained to predict shot boundary regions.

[0071] The original frame 400 additionally includes a salient area 404 which represents the area the original frame 400 will be cropped to produce the modified frame illustrated with respect to **FIG. 4**. In some embodiments, the salient area 404 is determined based on a distribution of saliency signals 410. In some embodiments, portions of the original frame 400 that satisfy one or more threshold criteria for an amount of saliency signals are included in the salient area 404 and portions of the original frame 400 that do not satisfy the threshold criteria for an amount of saliency signals are excluded from the salient area 404 according to the distribution of saliency signals 410. In some embodiments, the salient area 404 can be centered around a point of the original frame 400 with the greatest amount of saliency signals according to the distribution of saliency signals 410 and expand outwards to include other salient regions of the original frame 400, as illustrated. In some embodiments, the distribution of saliency signals 410 can be determined based on bounding boxes representing salient regions of the original frame 400. In some embodiments, bounding boxes can be given a lesser or greater weight in determining the distribution of saliency signals based on a location of the bounding boxes within the original frame 400. For example, a center portion of bounding boxes 402 can result in the greatest amount of saliency signals for the original frame 400 that decreases (linearly, exponentially, etc.) as the bounding box 402 expands outward. In another example, a bounding box (not illustrated) near (e.g., within 5 pixels of) the border of the original frame 400 can result in the least amount of saliency signals when compared against bounding boxes closer to center of the original frame 400.

[0072] In some embodiments, the distribution of saliency signals 410 can be a discrete distribution, as illustrated, in which discrete values of the distribution indicate (e.g., point to) salient regions and a magnitude of the discrete values indicate one or more of a confidence level or a size of a bounding box indicating the salient region. In some embodiments, the

original frame 400 can include many (e.g., 20 or more) bounding boxes representing salient regions of the original frame 400 such that the distribution of saliency signals 410 can be a continuous distribution (not illustrated), where the continuous distribution of saliency signals indicates a density of bounding boxes representing salient regions at a given point across the horizontal axis of the original frame 400. In some embodiments, the magnitude of a given point of the continuous distribution of saliency signals can represent one or more of a density of bounding boxes at the given point, a confidence level of the bounding boxes associated with the given point, or a size of the bounding box associated with the given point. In some embodiments, the distribution of saliency signals 410 can be semi-continuous (not illustrated) that exhibits characteristics of a discrete distribution and a continuous distribution. For example, the original frame 400 can include several (e.g., 5) bounding boxes representing salient regions of the original frame 400. Four of the five bounding boxes can be located near each other (e.g., within 10 pixels) or overlap to create a continuous distribution across first portion of the vertical axis of the original frame 400. The fifth bounding box can be located farther (e.g., 100 pixels or more) from the other bounding boxes such that the fifth bounding box results in a discrete signal within the distribution of saliency signals 410 representing a salient area at a second portion of the vertical axis of the original frame 400.

[0073] In some embodiments, processing logic can ensure the salient area 404 is enclosed within the bounding box 406 representing the frame boundary region of the original frame 400 and within the bounding box 408 representing the shot boundary region of the original frame 400. That is, a modified frame is generated such that the salient area 404 is included in the modified frame. For example, the salient area 404 can be an area within the frame boundary region (represented by bounding box 406) with the greatest amount of saliency signals according to the distribution of saliency signals 410.

[0074] The original frame 400 further illustrates a distribution of motion signals 420 of the original frame 400. In some embodiments, the distribution of motion signals 420 can be determined based on a motion detection model (e.g., motion detection model 314) that is trained to predict a presence of motion, an absence of motion, and/or an amount of motion in given video frame. In some embodiments, the salient area 404 of the original frame 400 can be determined based on the distribution of motion signals 420. In some embodiments, the salient area 404 of the original frame 400 can be determined based on the distribution of motion signals 420 in response to a determination that saliency signals are not present in the original frame 400. In some embodiments, the salient area 404 of the original frame 400 can be determined based on the distribution of motion signals 420 in response to determining the

distribution of saliency signals 410 does not satisfy a threshold criteria. For example, bounding boxes predicted by detectors described with respect to **FIG. 3** can fail to result in a distribution of saliency signals 410 that satisfies the threshold criteria for an amount of saliency signals. Accordingly, the distribution of motion signals 420 can be utilized to produce the salient area 404. In some embodiments, portions of the original frame 400 that satisfy one or more threshold criteria for an amount of motion signals are included in the salient area 404 and portions of the original frame 400 that do not satisfy the threshold criteria for an amount of motion signals are excluded from the salient area 404 according to the distribution of motion signals 420. In some embodiments, the salient area 404 can be centered around a point of the original frame 400 with the greatest amount of motion signals according to the distribution of motion signals 420 and expand outwards.

[0075] In some embodiments, the salient area 404 of the original frame 400 can be determined based on both the distribution of saliency signals 410 and the distribution of motion signals 420. For example, a weight can be applied to distribution of saliency signals 410 according to a relative amount saliency present in the original frame 400 and another weight can be applied to the distribution of motion signals 420 according to a relative amount of motion present in the original frame 400. The weighted distribution of saliency signals and weighted distribution of motions signals can indicate an important level of the distribution of saliency signals 410 and distribution of motion signal 420 respectively in determining the salient area 404. In some embodiments, the highest weighted distribution can indicate a distribution to utilize (e.g., the distribution of saliency signals 410 or the distribution of motion signals 420) in determining the salient area 404.

[0076] **FIG. 4B** is an example of a user interface (UI) 411 displaying a modified frame 412 of a modified video item, in accordance with aspects and implementations of the present disclosure. In response to modifying the video item, as described above, the UI 411 can be provided, for display on a client device (e.g., client device 102A), to present the modified video item. As illustrated, modified frame 412 of the modified video item corresponds to the salient area 404 of the original frame 400 (illustrated with respect to **FIG. 4A**). In some embodiments, the modified video item can be padded with additional space around the modified frame 412 to ensure the video item fits within a desired display format (e.g., within a certain aspect ratio). For example, a first region 416A can be included within the UI 411 above the modified frame 412 and a second region 416B can be include within the UI 411 below the modified frame 412 to maintain a display format of the client device without stretching or otherwise distorting the modified frame 412.

[0077] FIG. 5 illustrates a flow diagram of an example method 500 of automatically modifying frame presentations characteristics of a media item, in accordance with aspects and implementations of the present disclosure. Method 500 can be performed by processing logic that can include hardware (circuitry, dedicated logic, etc.), software (e.g., instructions run on a processing device), or a combination thereof. In one implementation, some or all the operations of method 500 can be performed by one or more components of system 100 of FIG. 1. In at least one embodiment, some or all of the operations of method 500 can be performed by a server device, as described above.

[0078] For simplicity of explanation, the methods of this disclosure are depicted and described as a series of acts. However, acts in accordance with this disclosure can occur in various orders and/or concurrently, and with other acts not presented and described herein. Furthermore, not all illustrated acts can be required to implement the methods in accordance with the disclosed subject matter. In addition, those skilled in the art will understand and appreciate that the methods could alternatively be represented as a series of interrelated states via a state diagram or events. Additionally, it should be appreciated that the methods disclosed in this specification are capable of being stored on an article of manufacture to facilitate transporting and transferring such methods to computing devices. The term “article of manufacture,” as used herein, is intended to encompass a computer program accessible from any computer-readable device or storage media.

[0079] At operation 502, processing logic can receive a request from a client device, such as client device 102A, to modify at least one frame presentation characteristic of a media item, such as media item 121.

[0080] At operation 504, responsive to reception of the request from the client device to modify that at least the one frame presentation characteristic, the processing logic can perform operations 506 and 508. In some embodiments, the request to adjust the at least one frame presentation characteristic of the original media item is received responsive to a user interaction with one or more elements of a UI provided for display on the client device.

[0081] At operation 506, the processing logic can automatically modify the original media item to reflect the at least one frame presentation characteristic requested to be adjusted, while ensuring that one or more salient regions of the original media item are present in frames of a modified media item. In some embodiments, to automatically modify the original media item, the processing logic can modify an aspect ratio of the media item.

[0082] In some embodiments, to automatically modify the original media item, the processing logic can provide at least a subset of frames of the original media item as input to

one or more machine learning models. The one or more machine learning models can be trained to predict, based on a given frame, bounding boxes for the given frame that each represent a salient region of the given frame. The processing logic can obtain outputs from the one or more machine learning models, wherein the outputs include bounding boxes that each indicate a salient region of the original media item. In some embodiments, each of the one or more machine learning models is trained to predict bounding boxes for visual features of a particular type. In some embodiments, the types of visual features include one or more of facial features, objects, frame boundary regions, or shot boundary regions.

[0083] In some embodiments, to automatically modify the original media item, the processing logic can determine a distribution of saliency signals, such as the distribution of saliency signals 410, for each of the at least the subset of the frames of the original media item based on the bounding boxes indicating the one or more salient regions of the media item. The processing logic can further determine a salient area, such as salient area 404, of each of the at least the subset of the frames of the original media item based on the distribution of saliency signals. The processing logic can adjust a frame boundary of each of the at least the subset of the frames of the original media item to exclude content outside of a respective salient area and to include content within the respective salient area to obtain frames of the modified media item.

[0084] In some embodiments, to determine the salient area for each of the at least the subset of the frames of the original media item based on the distribution of saliency signals, the processing logic can include, in a salient area of a respective frame of the at least the subset of the frames of the original media item, one or more portions of the respective frame that satisfy one or more threshold criteria for an amount of saliency signals. The processing logic can exclude, from the salient area of the respective frame, one or more portions of the respective frame that do not satisfy one or more threshold criteria for the amount of saliency signals. In some embodiments, the salient area for each of the at least the subset of the frames of the original media item is determined based on a distribution of motion signals between frames of the media item.

[0085] At operation 508, the processing logic can provide, for display on the client device, a user interface (UI) presenting the modified media item. In some embodiments, the video item is modified to convert it into a short-form version. A short-form version can refer to a brief, concise, and quickly consumable video clip of limited duration (e.g., less than 60 seconds in length). Such a video clip can be intended for content viewing via, for example, a mobile device (e.g., a smartphone) that is frequently held vertically by users, necessitating its

conversion into a portrait format (e.g., aspect ratio 9:16) for better use of screen space and improved viewing experience of users.

[0086] FIG. 6 illustrates an example training engine 141 for training and deployment of a deep neural network, in accordance with aspects and implementations of the present disclosure. In some embodiments, untrained neural network 606 is trained using a training dataset 602. In some embodiments, training data generator 131 can be configured to generate training dataset 202. In some embodiments, training framework 604 is a PyTorch framework, whereas in other embodiments, training framework 604 is a TensorFlow, Boost, Caffe, Microsoft Cognitive Toolkit/CNTK, MXNet, Chainer, Keras, DeepLearning4j, or other training framework. In at least one embodiment, training framework 604 trains an untrained neural network 606 and enables it to be trained using processing resources described herein to generate a trained neural network 608. The training framework 604 can be used to generate a trained shot boundary detection model 310, object detection model 312, motion detection model 314, face detection model 316, and/or border detection model 318. In some embodiments, weights can be chosen randomly or by pre-training using a deep belief network. In some embodiments, training can be performed in either a supervised, partially supervised, or unsupervised manner.

[0087] In some embodiments, untrained neural network 606 is trained using supervised learning, wherein training dataset 602 includes an input paired with a desired output for an input, or where training dataset 602 includes input having a known output and an output of neural network 606 is manually graded. For example, an untrained face detection model 316 can be trained using supervised learning, where training dataset 602 includes an input of historical images containing facial features such as human faces paired with a desired output of the historical images with corresponding bounding boxes that indicate a location of each face in the respective image. The dataset can include diverse images capturing variations in location, orientation, and presence of facial features. In at least one embodiment, untrained neural network 606 is trained in a supervised manner and processes inputs from training dataset 602 and compares resulting outputs against a set of expected or desired outputs. In at least one embodiment, errors are then propagated back through untrained neural network 606. In at least one embodiment, training framework 604 adjusts weights that control untrained neural network 606. In at least one embodiment, training framework 604 includes tools to monitor how well untrained neural network 606 is converging towards a model, such as trained neural network 608, suitable to generating correct answers, such as in result 614, based on input data such as a new dataset 612. In at least one embodiment, training

framework 604 trains untrained neural network 606 repeatedly while adjust weights to refine an output of untrained neural network 606 using a loss function and adjustment algorithm, such as stochastic gradient descent. In at least one embodiment, training framework 604 trains untrained neural network 606 until untrained neural network 606 achieves a desired accuracy. In at least one embodiment, trained neural network 608 can then be deployed to implement any number of machine learning operations.

[0088] In at least one embodiment, untrained neural network 606 is trained using unsupervised learning, wherein untrained neural network 606 attempts to train itself using unlabeled data. In at least one embodiment, unsupervised learning training dataset 602 will include input data without any associated output data or “ground truth” data. In at least one embodiment, untrained neural network 606 can learn groupings within training dataset 602 and can determine how individual inputs are related to training dataset 602. In at least one embodiment, unsupervised training can be used to generate a self-organizing map in trained neural network 608 capable of performing operations useful in reducing dimensionality of new dataset 612. In at least one embodiment, unsupervised training can also be used to perform anomaly detection, which allows identification of data points in new dataset 612 that deviate from normal patterns of new dataset 612.

[0089] In at least one embodiment, semi-supervised learning can be used, which is a technique in which in training dataset 602 includes a mix of labeled and unlabeled data. In at least one embodiment, training framework 604 can be used to perform incremental learning, such as through transferred learning techniques. In at least one embodiment, incremental learning enables trained neural network 608 to adapt to new dataset 612 without forgetting knowledge instilled within trained neural network 608 during initial training.

[0090] **FIG. 7** is a block diagram illustrating an exemplary computer system 700, in accordance with implementations of the present disclosure. The computer system 700 can correspond to platform 120 and/or client devices 102A-N, described with respect to **FIG. 1**. Computer system 700 can operate in the capacity of a server or an endpoint machine in endpoint-server network environment, or as a peer machine in a peer-to-peer (or distributed) network environment. The machine can be a television, a personal computer (PC), a tablet PC, a set-top box (STB), a Personal Digital Assistant (PDA), a cellular telephone, a web appliance, a server, a network router, switch or bridge, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine. Further, while only a single machine is illustrated, the term “machine” shall also be taken to

include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

[0091] The example computer system 700 includes a processing device (processor) 702, a main memory 704 (e.g., read-only memory (ROM), flash memory, dynamic random access memory (DRAM) such as synchronous DRAM (SDRAM), double data rate (DDR SDRAM), or DRAM (RDRAM), etc.), a static memory 706 (e.g., flash memory, static random access memory (SRAM), etc.), and a data storage device 718, which communicate with each other via a bus 740.

[0092] Processor (processing device) 702 represents one or more general-purpose processing devices such as a microprocessor, central processing unit, or the like. More particularly, the processor 702 can be a complex instruction set computing (CISC) microprocessor, reduced instruction set computing (RISC) microprocessor, very long instruction word (VLIW) microprocessor, or a processor implementing other instruction sets or processors implementing a combination of instruction sets. The processor 702 can also be one or more special-purpose processing devices such as an application specific integrated circuit (ASIC), a field programmable gate array (FPGA), a digital signal processor (DSP), network processor, or the like. The processor 702 is configured to execute instructions 705 for performing the operations discussed herein.

[0093] The computer system 700 can further include a network interface device 708. The computer system 700 also can include a video display unit 710 (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)), an input device 712 (e.g., a keyboard, and alphanumeric keyboard, a motion sensing input device, touch screen), a cursor control device 714 (e.g., a mouse), and a signal generation device 720 (e.g., a speaker).

[0094] The data storage device 718 can include a non-transitory machine-readable storage medium 724 (also non-transitory computer-readable storage medium) on which is stored one or more sets of instructions 705 embodying any one or more of the methodologies or functions described herein. The instructions can also reside, completely or at least partially, within the main memory 704 and/or within the processor 702 during execution thereof by the computer system 700, the main memory 704 and the processor 702 also constituting machine-readable storage media. The instructions can further be transmitted or received over a network 730 via the network interface device 708.

[0095] In one implementation, the instructions 705 include instructions for automatically modifying frame presentation characteristics of a media item. While the computer-readable storage medium 724 (machine-readable storage medium) is shown in an exemplary

implementation to be a single medium, the terms “computer-readable storage medium” and “machine-readable storage medium” should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) that store the one or more sets of instructions. The terms “computer-readable storage medium” and “machine-readable storage medium” shall also be taken to include any medium that is capable of storing, encoding or carrying a set of instructions for execution by the machine and that cause the machine to perform any one or more of the methodologies of the present disclosure. The terms “computer-readable storage medium” and “machine-readable storage medium” shall accordingly be taken to include, but not be limited to, solid-state memories, optical media, and magnetic media.

[0096] Reference throughout this specification to “one implementation,” “one embodiment,” “an implementation,” or “an embodiment,” means that a particular feature, structure, or characteristic described in connection with the implementation and/or embodiment is included in at least one implementation and/or embodiment. Thus, the appearances of the phrase “in one implementation,” or “in an implementation,” in various places throughout this specification can, but are not necessarily, referring to the same implementation, depending on the circumstances. Furthermore, the particular features, structures, or characteristics can be combined in any suitable manner in one or more implementations.

[0097] To the extent that the terms “includes,” “including,” “has,” “contains,” variants thereof, and other similar words are used in either the detailed description or the claims, these terms are intended to be inclusive in a manner similar to the term “comprising” as an open transition word without precluding any additional or other elements.

[0098] As used in this application, the terms “component,” “module,” “system,” or the like are generally intended to refer to a computer-related entity, either hardware (e.g., a circuit), software, a combination of hardware and software, or an entity related to an operational machine with one or more specific functionalities. For example, a component can be, but is not limited to being, a process running on a processor (e.g., digital signal processor), a processor, an object, an executable, a thread of execution, a program, and/or a computer. By way of illustration, both an application running on a controller and the controller can be a component. One or more components can reside within a process and/or thread of execution and a component can be localized on one computer and/or distributed between two or more computers. Further, a “device” can come in the form of specially designed hardware; generalized hardware made specialized by the execution of software

thereon that enables hardware to perform specific functions (e.g., generating interest points and/or descriptors); software on a computer readable medium; or a combination thereof.

[0099] The aforementioned systems, circuits, modules, and so on have been described with respect to interact between several components and/or blocks. It can be appreciated that such systems, circuits, components, blocks, and so forth can include those components or specified sub-components, some of the specified components or sub-components, and/or additional components, and according to various permutations and combinations of the foregoing. Sub-components can also be implemented as components communicatively coupled to other components rather than included within parent components (hierarchical). Additionally, it should be noted that one or more components can be combined into a single component providing aggregate functionality or divided into several separate sub-components, and any one or more middle layers, such as a management layer, can be provided to communicatively couple to such sub-components in order to provide integrated functionality. Any components described herein can also interact with one or more other components not specifically described herein but known by those of skill in the art.

[00100] Moreover, the words “example” or “exemplary” are used herein to mean serving as an example, instance, or illustration. Any aspect or design described herein as “exemplary” is not necessarily to be construed as preferred or advantageous over other aspects or designs. Rather, use of the words “example” or “exemplary” is intended to present concepts in a concrete fashion. As used in this application, the term “or” is intended to mean an inclusive “or” rather than an exclusive “or.” That is, unless specified otherwise, or clear from context, “X employs A or B” is intended to mean any of the natural inclusive permutations. That is, if X employs A; X employs B; or X employs both A and B, then “X employs A or B” is satisfied under any of the foregoing instances. In addition, the articles “a” and “an” as used in this application and the appended claims should generally be construed to mean “one or more” unless specified otherwise or clear from context to be directed to a singular form.

[00101] Finally, implementations described herein include collection of data describing a user and/or activities of a user. In one implementation, such data is only collected upon the user providing consent to the collection of this data. In some implementations, a user is prompted to explicitly allow data collection. Further, the user can opt-in or opt-out of participating in such data collection activities. In one implementation, the collect data is anonymized prior to performing any analysis to obtain any statistical patterns so that the identity of the user cannot be determined from the collected data.

CLAIMS

What is claimed is:

1. A method comprising:
receiving a request from a client device to modify at least one frame presentation characteristic of an original media item; and
responsive to receiving the request from the client device to modify the at least one frame presentation characteristic of the original media item:
automatically modifying the original media item to reflect the at least one frame presentation characteristic requested to be adjusted, while ensuring that one or more salient regions of the original media item are present in a plurality of frames of a modified media item; and
providing, for display on the client device, a user interface (UI) presenting the modified media item.
2. The method of claim 1, wherein automatically modifying the original media item comprises modifying an aspect ratio of the media item to facilitate creation of a short-form version of the media item.
3. The method of claim 1, wherein automatically modifying the original media item comprises:
providing at least a subset of a plurality of frames of the original media item as input to one or more machine learning models, wherein the one or more machine learning models are trained to predict, based on a given frame, bounding boxes for the given frame that each represent a salient region of the given frame; and
obtaining a plurality of outputs from the one or more machine learning models, wherein the plurality of outputs comprises a plurality of bounding boxes each indicating a salient region of the original media item.
4. The method of claim 3, wherein each of the one or more machine learning models is trained to predict bounding boxes for visual features of a particular type of a plurality of types.

5. The method of claim 4, wherein the plurality of types of visual features comprises one or more of facial features, objects, frame boundary regions, or shot boundary regions.

6. The method of claim 3, wherein automatically modifying the original media item comprises:

determining a distribution of saliency signals for each of the at least the subset of the plurality of frames of the original media item based on the plurality of bounding boxes indicating the one or more salient regions of the media item;

determining a salient area of each of the at least the subset of the plurality of frames of the original media item based on the distribution of saliency signals; and

adjusting a frame boundary of each of the at least the subset of the plurality of frames of the original media item to exclude content outside of a respective salient area and to include content within the respective salient area to obtain the plurality of frames of the modified media item.

7. The method of claim 6, wherein determining the salient area for each of the at least the subset of the plurality of frames of the original media item based on the distribution of saliency signals comprises:

including, in a salient area of a respective frame of the at least the subset of the plurality of frames of the original media item, one or more portions of the respective frame that satisfy one or more threshold criteria for an amount of saliency signals; and

excluding, from the salient area of the respective frame, one or more portions of the respective frame that do not satisfy the one or more threshold criteria for the amount of saliency signals.

8. The method of claim 7, wherein the salient area for each of the at least the subset of the plurality of frames of the original media item is determined based on a distribution of motion signals between the plurality of frames of the media item.

9. The method of claim 1, wherein the request to adjust the at least one frame presentation characteristic of the original media item is received responsive to a user interaction with one or more elements of the UI provided for display on the client device.

10. A system comprising:
a memory device; and
a processing device coupled to the memory device, the processing device to perform operation comprising:
receiving a request from a client device to modify at least one frame presentation characteristic of an original media item; and
responsive to receiving the request from the client device to modify the at least one frame presentation characteristic of the original media item:
automatically modifying the original media item to reflect the at least one frame presentation characteristic requested to be adjusted, while ensuring that one or more salient regions of the original media item are present in a plurality of frames of a modified media item; and
providing, for display on the client device, a user interface (UI) presenting the modified media item.
11. The system of claim 10, wherein automatically modifying the original media item comprises modifying an aspect ratio of the media item to facilitate creation of a short-form version of the media item.
12. The system of claim 10, wherein automatically modifying the original media item comprises:
providing at least a subset of a plurality of frames of the original media item as input to one or more machine learning models, wherein the one or more machine learning models are trained to predict, based on a given frame, bounding boxes for the given frame that each represent a salient region of the given frame; and
obtaining a plurality of outputs from the one or more machine learning models, wherein the plurality of outputs comprises a plurality of bounding boxes each indicating a salient region of the original media item.
13. The system of claim 12, wherein each of the one or more machine learning models is trained to predict bounding boxes for visual features of a particular type of a plurality of types.

14. The system of claim 13, wherein the plurality of types of visual features comprises one or more of facial features, objects, frame boundary regions, or shot boundary regions.

15. The system of claim 12, wherein automatically modifying the original media item comprises:

determining a distribution of saliency signals for each of the at least the subset of the plurality of frames of the original media item based on the plurality of bounding boxes indicating the one or more salient regions of the media item;

determining a salient area of each of the at least the subset of the plurality of frames of the original media item based on the distribution of saliency signals; and

adjusting a frame boundary of each of the at least the subset of the plurality of frames of the original media item to exclude content outside of a respective salient area and to include content within the respective salient area to obtain the plurality of frames of the modified media item.

16. The system of claim 15, wherein determining the salient area for each of the at least the subset of the plurality of frames of the original media item based on the distribution of saliency signals comprises:

including, in a salient area of a respective frame of the at least the subset of the plurality of frames of the original media item, one or more portions of the respective frame that satisfy one or more threshold criteria for an amount of saliency signals; and

excluding, from the salient area of the respective frame, one or more portions of the respective frame that do not satisfy one or more threshold criteria for the amount of saliency signals.

17. The system of claim 16, wherein the salient area for each of the at least the subset of the plurality of frames of the original media item is determined based on a distribution of motion signals between the plurality of frames of the media item.

18. The system of claim 10, wherein the request to adjust the at least one frame presentation characteristic of the original media item is received responsive to a user interaction with one or more elements of the UI provided for display on the client device.

19. A non-transitory computer-readable storage medium comprising instructions for a server that, when executed by a processing device, cause the processing device to perform operations comprising:

receiving a request from a client device to modify at least one frame presentation characteristic of an original media item; and

responsive to receiving the request from the client device to modify the at least one frame presentation characteristic of the original media item:

automatically modifying the original media item to reflect the at least one frame presentation characteristic requested to be adjusted, while ensuring that one or more salient regions of the original media item are present in a plurality of frames of a modified media item; and

providing, for display on the client device, a user interface (UI) presenting the modified media item.

20. The non-transitory computer-readable storage medium of claim 19, wherein automatically modifying the original media item comprises:

providing at least a subset of a plurality of frames of the original media item as input to one or more machine learning models, wherein the one or more machine learning models are trained to predict, based on a given frame, bounding boxes for the given frame that each represent a salient region of the given frame; and

obtaining a plurality of outputs from the one or more machine learning models, wherein the plurality of outputs comprises a plurality of bounding boxes each indicating a salient region of the original media item.

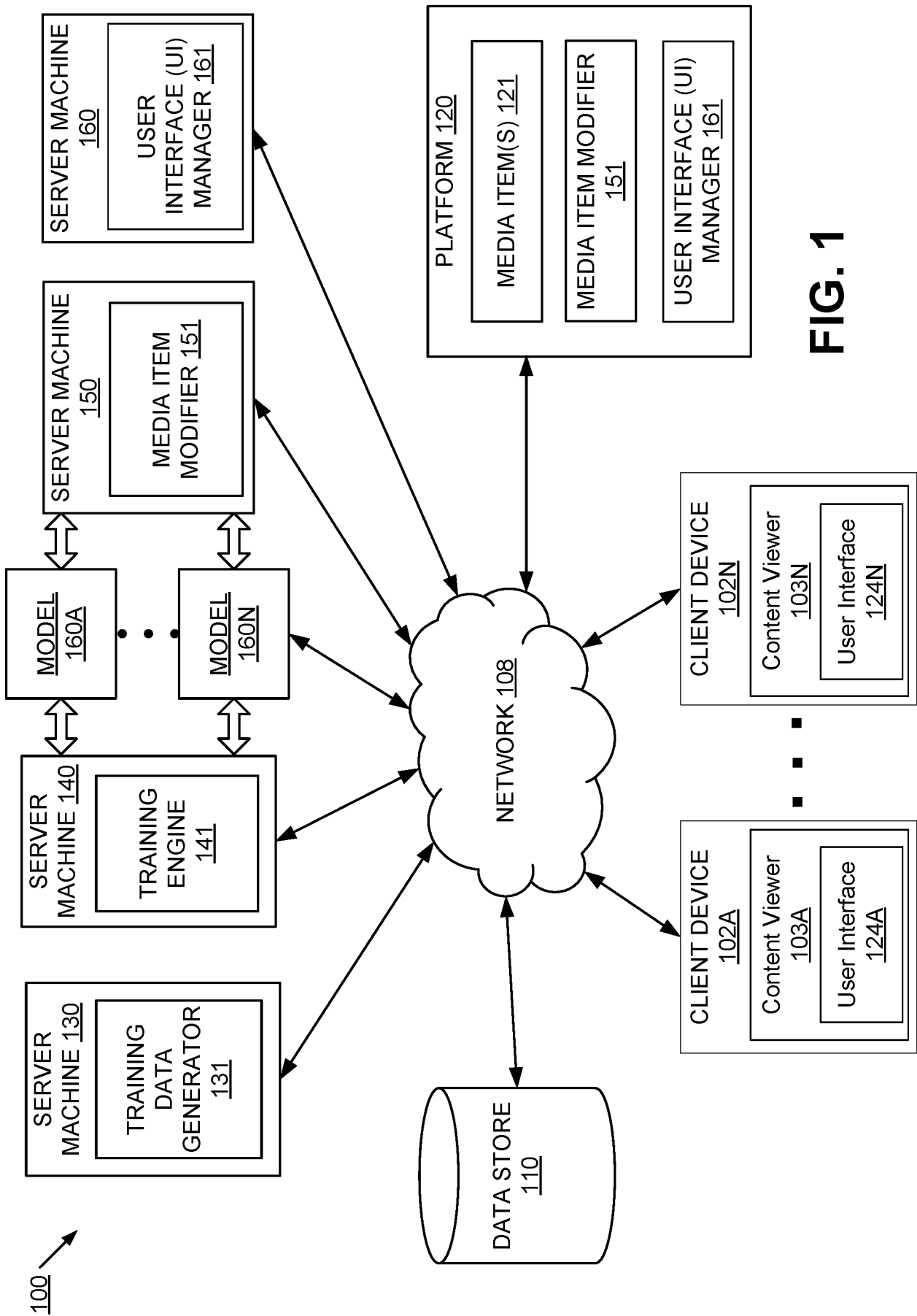


FIG. 1

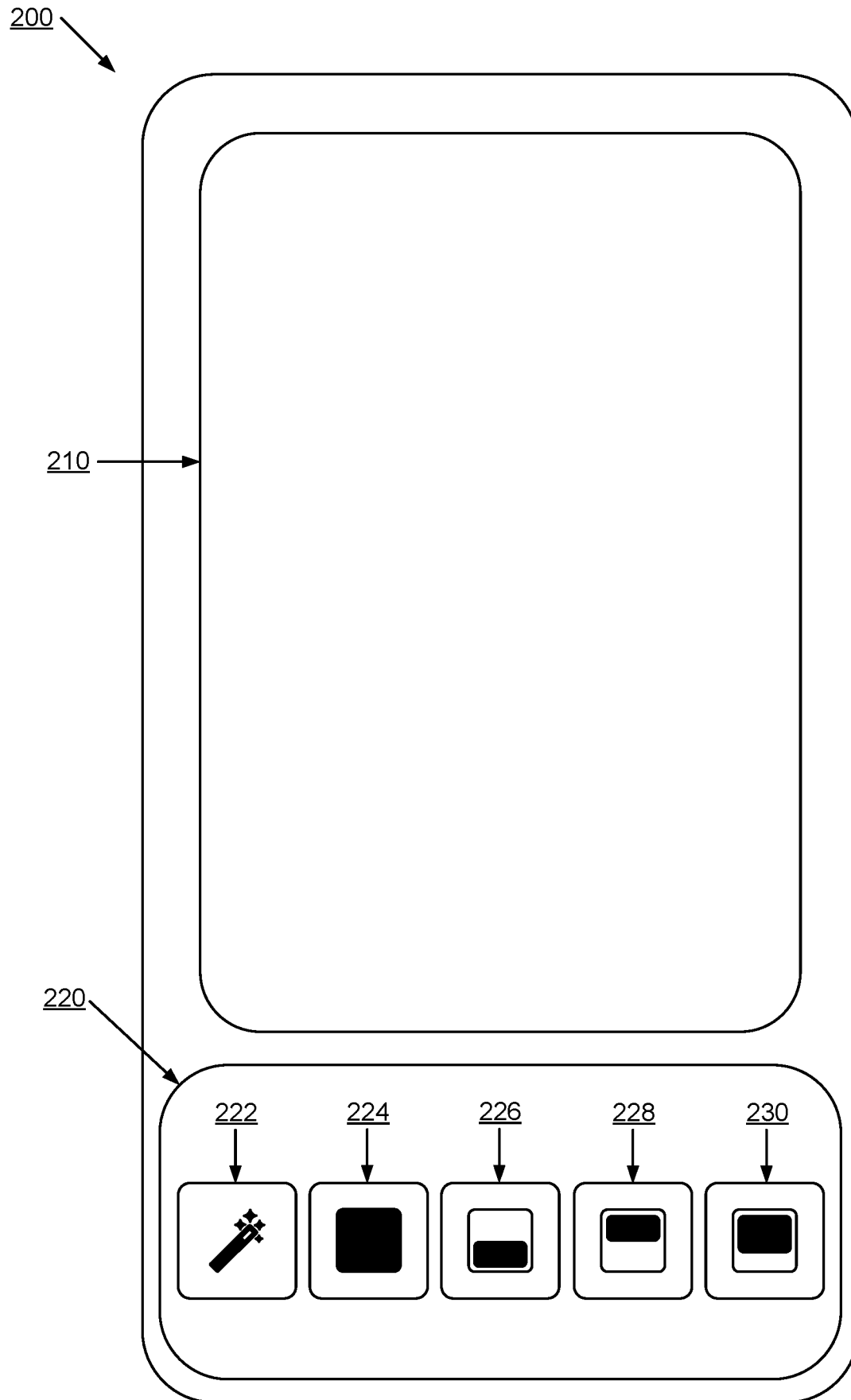
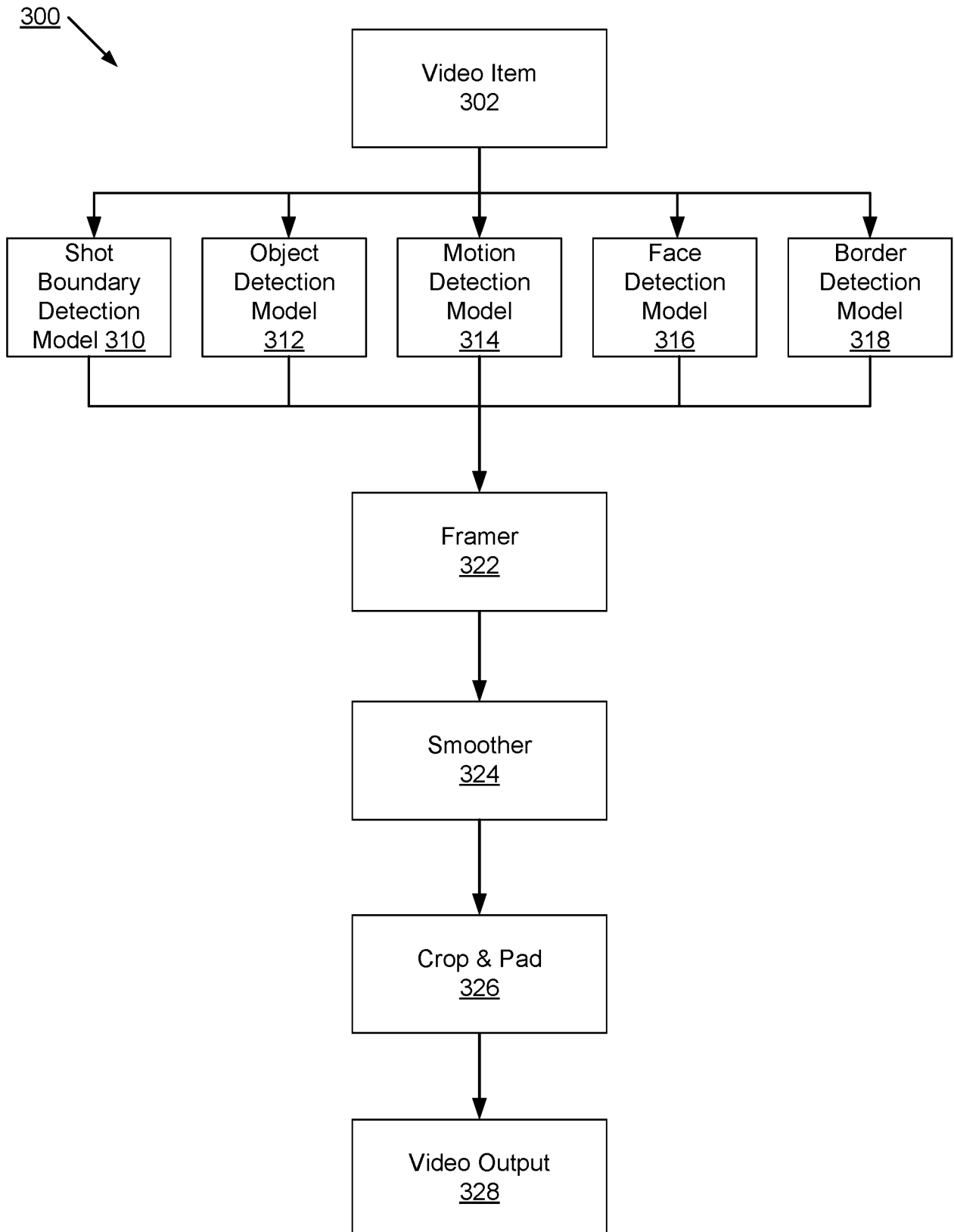


FIG. 2

**FIG. 3**

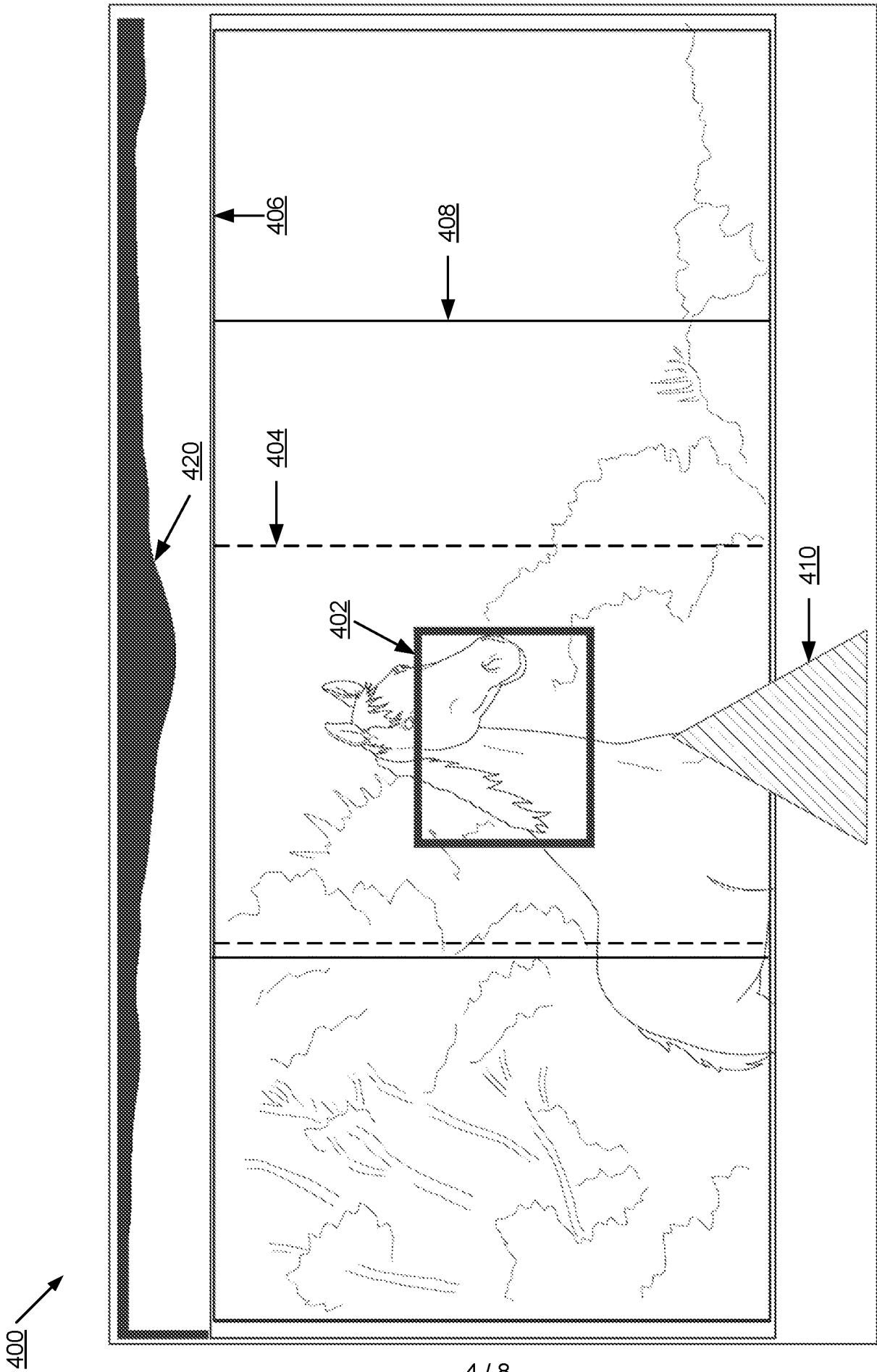


FIG. 4A

411

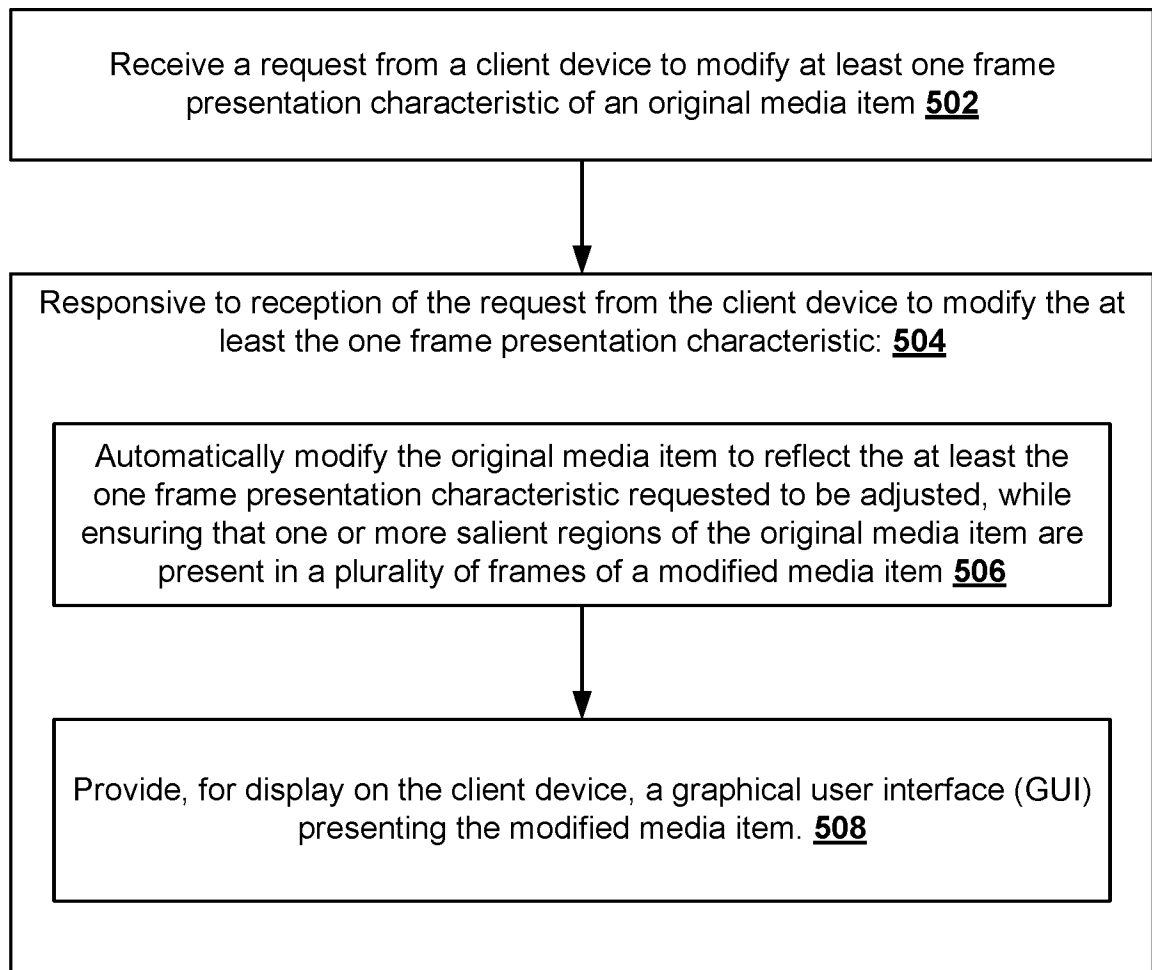


← 414A

← 412

← 414B

FIG. 4B

500**FIG. 5**

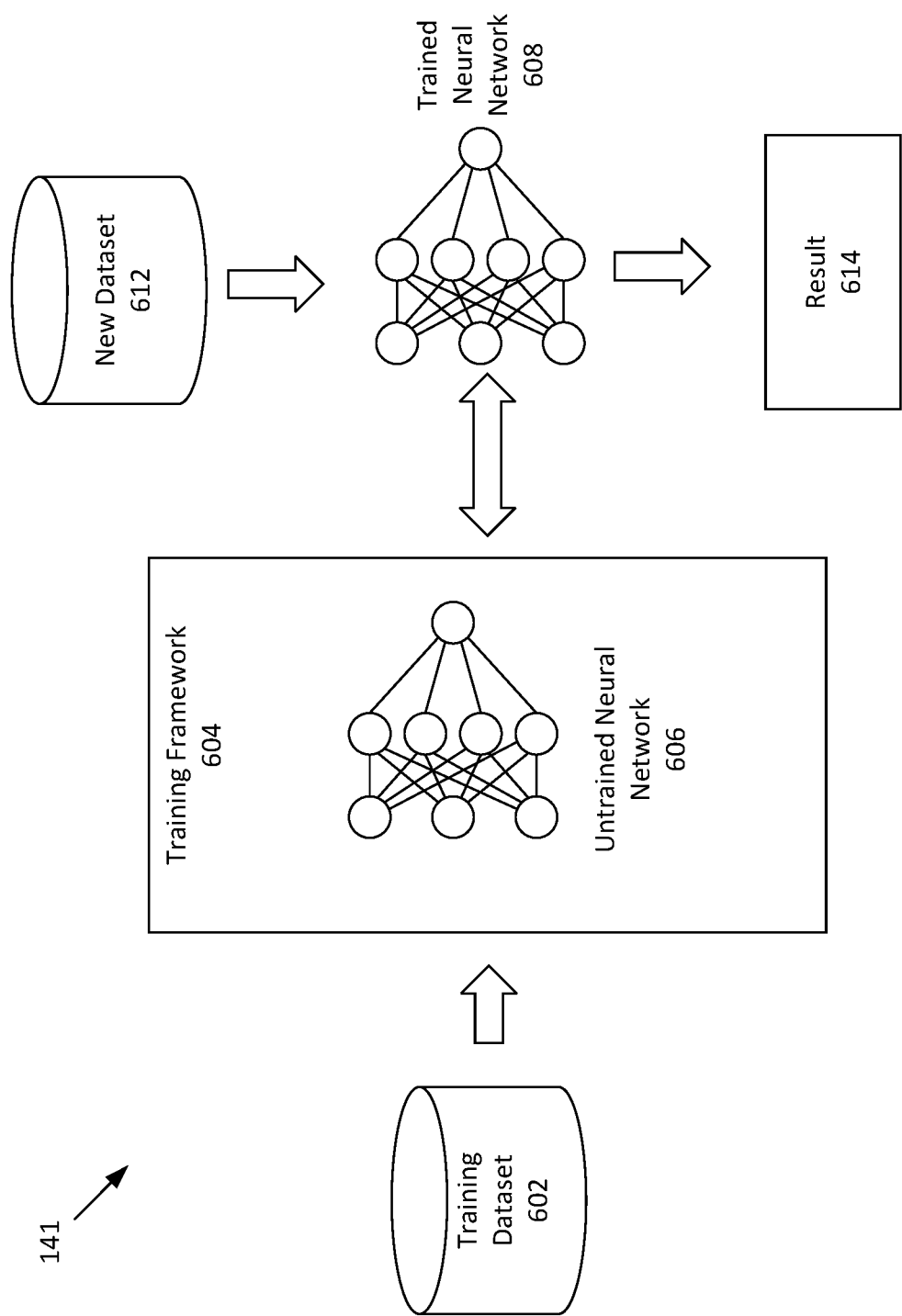
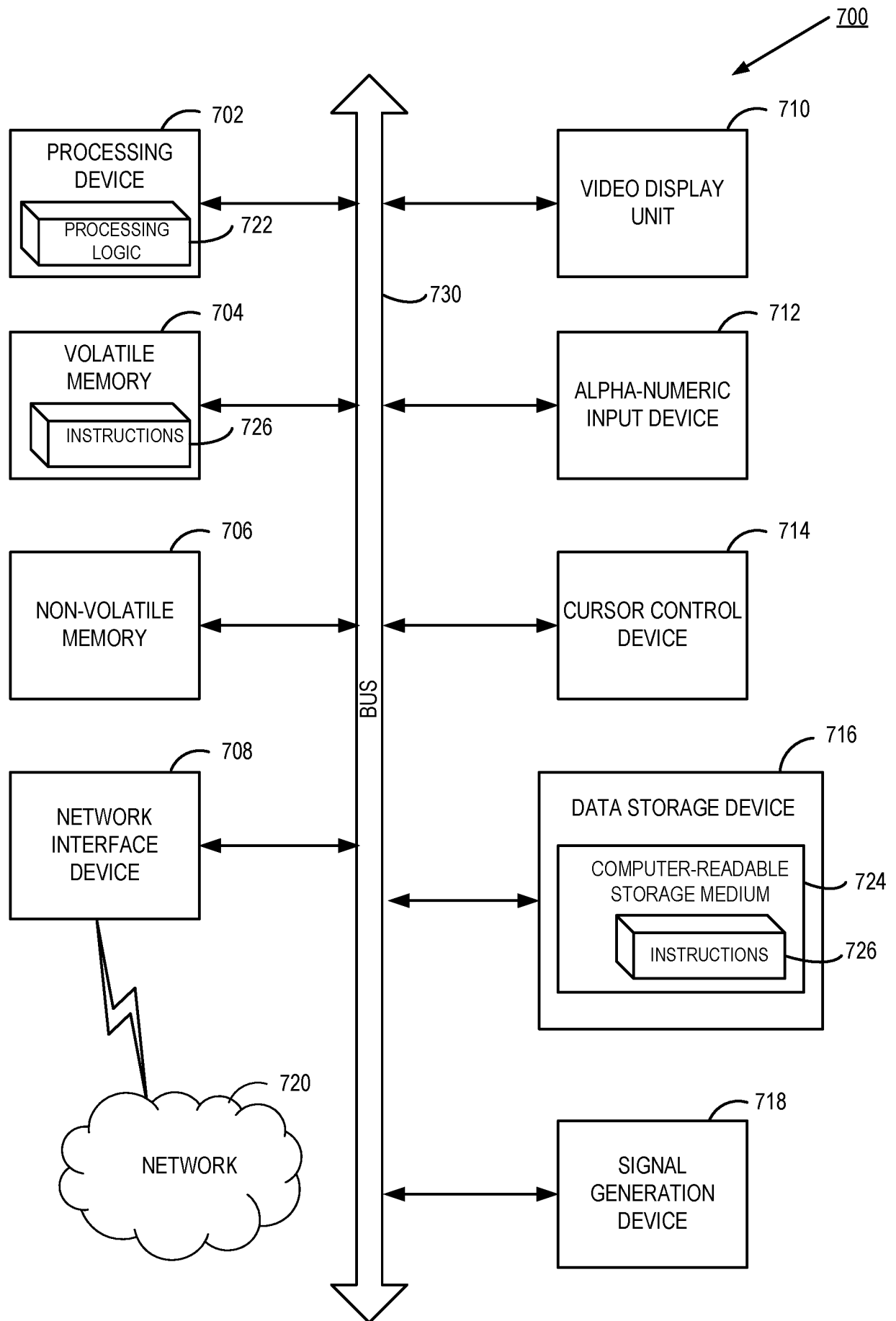


FIG. 6

**FIG. 7**

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2023/025971

A. CLASSIFICATION OF SUBJECT MATTER

INV.	H04N21/234	G06F3/04845	G06T3/40	G06T11/60	G06V10/25
	G06V10/82	G06V20/40	H04N21/44		

ADD .

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04N G06V G06T G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2022/383032 A1 (GARG ANSHUL [IN] ET AL) 1 December 2022 (2022-12-01)	1-7, 9-16, 18-20
Y	paragraphs [0016] - [0019], [0065] - [0068], [0074]; figures 3a, 4b -----	8, 17
X	US 2021/398333 A1 (KALU KALU O [US] ET AL) 23 December 2021 (2021-12-23) paragraph [0007] -----	1, 2, 10, 11, 19, 20
X	US 2023/059805 A1 (KANSARA APURVAKUMAR DILIPKUMAR [US] ET AL) 23 February 2023 (2023-02-23) paragraph [0068]; figures 6-10 ----- -/--	1-20

☒ Further documents are listed in the continuation of Box C.

x See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

13 December 2023

Date of mailing of the international search report

22/12/2023

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer _____

Bertrand, Frédéric

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2023/025971

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	UMEKI YO ET AL: "Salient Object Detection With Importance Degree", IEEE ACCESS, IEEE, USA, vol. 8, 6 August 2020 (2020-08-06), pages 147059-147069, XP011805960, DOI: 10.1109/ACCESS.2020.3014886 [retrieved on 2020-08-17] the whole document -----	1-20
A	WANG BAOYAN ET AL: "Salient object detection based on objectness", 2015 IEEE INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING, COMMUNICATIONS AND COMPUTING (ICSPCC), IEEE, 19 September 2015 (2015-09-19), pages 1-5, XP032817776, DOI: 10.1109/ICSPCC.2015.7338816 ISBN: 978-1-4799-8918-8 [retrieved on 2015-11-25] the whole document -----	1-20
Y	US 2019/005659 A1 (PIEKNIIEWSKI FILIP [US] ET AL) 3 January 2019 (2019-01-03)	8, 17
A	paragraphs [0088], [0158] - [0161] -----	1, 10, 19, 20
A	US 2013/108175 A1 (PTUCHA RAYMOND WILLIAM [US]) 2 May 2013 (2013-05-02) the whole document -----	1-20
A	ZHU TUN ET AL: "Horizontal-to-Vertical Video Conversion", IEEE TRANSACTIONS ON MULTIMEDIA, IEEE, USA, vol. 24, 23 June 2021 (2021-06-23), pages 3036-3048, XP011910957, ISSN: 1520-9210, DOI: 10.1109/TMM.2021.3092202 [retrieved on 2021-06-24] the whole document -----	5, 14

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2023/025971

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2022383032	A1	01-12-2022	NONE
US 2021398333	A1	23-12-2021	CN 113822898 A 21-12-2021
		EP 3934237 A1	05-01-2022
		KR 20210157319 A	28-12-2021
		US 2021398333 A1	23-12-2021
US 2023059805	A1	23-02-2023	AU 2020304345 A1 23-12-2021
		AU 2023202133 A1	04-05-2023
		BR 112021026664 A2	15-02-2022
		CA 3138218 A1	30-12-2020
		EP 3991439 A1	04-05-2022
		US 10834465 B1	10-11-2020
		US 2021021900 A1	21-01-2021
		US 2023059805 A1	23-02-2023
		US 2023300392 A1	21-09-2023
		WO 2020263655 A1	30-12-2020
US 2019005659	A1	03-01-2019	US 10268919 B1 23-04-2019
		US 2016086050 A1	24-03-2016
		US 2016086051 A1	24-03-2016
		US 2016086052 A1	24-03-2016
		US 2018293742 A1	11-10-2018
		US 2019005659 A1	03-01-2019
		US 2019043208 A1	07-02-2019
		US 2019251386 A1	15-08-2019
US 2013108175	A1	02-05-2013	NONE