(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
01 June 2023 (01.06.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/094861 A1

(51) International Patent Classification:
G06T 7/00 (2017.01) *G06V 10/40* (2022.01)

(21) International Application Number:
PCT/IB2021/060976

(22) International Filing Date:
25 November 2021 (25.11.2021)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **L&T TECHNOLOGY SERVICES LIMITED** [IN/IN]; DLF IT SEZ Park, 2nd Floor – Block 3, Mount Poonamallee Road, Ramapuram, Tamil Nadu, Chennai 600089 (IN).

(72) Inventors: **SINGH, Dr. Madhusudan**; B-603, Ajmera Stone Park, 1st Cross, Neeladri Road, Electronic City - 1, Karnataka, Bangalore 560100. (IN). **MALLICK, Triptesh**; N0026, Vill- Fulhari, Post-Pukhuria, West Bengal., Bankura 722160 (IN). **PARIKH, Shyamal**; 29, Kunika Society, Opp. Anand Nagar, Karelibaug, Gujarat, Vadodara, 390022 (IN). **BHADAURIA, Sudhir**; 25, Madhav Park-3, Near Madhav School, Pranami Nagar, Vastral Road, Gujarat, Ahmedabad 382418 (IN). **PATEL, Meet**; A/P-33 Patel Falia, Rahej, Ta.- Gandevi, Dist.-Gujarat, Navsari 396360 (IN).

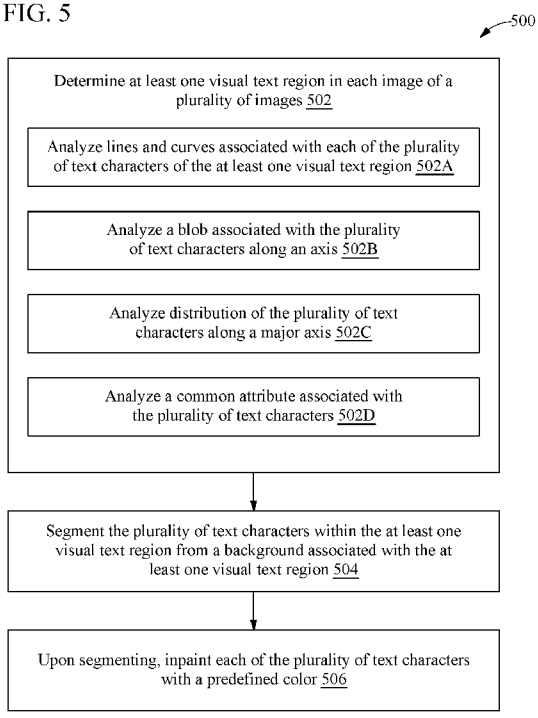
(74) Agent: **KOSHY VARGHESE, Robin**; DLF IT SEZ Park, 2nd Floor – Block 3, Mount Poonamallee Road, Ramapuram, Tamil Nadu, Chennai 600089 (IN).

(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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: A SYSTEM AND METHOD FOR VISUAL TEXT TRANSFORMATION

FIG. 5



(57) Abstract: A method and system for extracting text from a video stream is disclosed. The method may include determining at least one visual text region in each image of a plurality of images. The at least one visual text region may include a plurality of text characters and determining the at least one visual text region is based on analysis of one of lines and curves associated with each of the plurality of text characters, blob associated with the plurality of text characters along an axis, distribution of the plurality of text characters along a major axis, and a common attribute associated with the plurality of text characters. The method may further include segmenting the plurality of text characters from a background, and inpainting each of the plurality of text characters with a predefined color.

WO 2023/094861 A1

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i))*

Published:

— *with international search report (Art. 21(3))*
— *with amended claims (Art. 19(1))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

A SYSTEM AND METHOD FOR VISUAL TEXT TRANSFORMATION

DESCRIPTION

TECHNICAL FIELD

[001] This disclosure relates generally to image or video processing, and more particularly to a method and a system for extracting text from an image or a video stream.

BACKGROUND

[002] Videos like movies, news videos, commercials, and television programs may sometime contain textual information that can be informative and useful to a viewer. However, the viewer may not have options for extracting the text, other than manually writing down the text.

[003] Some text extraction applications can be used for detecting and extracting text present in the video frames. These image or video processing applications may be used in relation to motion pictures for digitization of credit scenes, subtitle verification, media library metadata tagging, and the like. For example, text from natural scenes that are captured in video frames may be identified using any conventional histogram of oriented gradients (HOG) feature extraction-based techniques, and the text may further be segmented using hierarchical clustering algorithms.

[004] However, the hierarchical clustering algorithms may be inefficient because of sensitivity to divergence and high computational cost. Further, the existing techniques for detecting text in videos of natural scenes may focus on characters of the text instead of a background noise removal. As a result, such techniques may be computationally expensive and may iterate multiple times, thereby increasing memory requirements of the associated systems.

[005] Accordingly, there is a need for a computationally efficient and accurate text detection and segmentation solution.

1.0 SUMMARY OF THE INVENTION

[006] In an embodiment, a method of extracting text from a video stream is disclosed. The method may include determining at least one visual text region in each image of a plurality of images. The video stream may include the plurality of images in sequential order. The at least one visual text region may include a plurality of text characters. The method may further include determining the at least one visual text region. The at least one visual text region may be determined based on analysis of one of lines and curves associated with each of the plurality of text characters, a blob associated with the plurality of text characters along an axis, distribution of the plurality of

text characters along a major axis (i.e. Horizontal axis or Vertical axis), and a common attribute associated with the plurality of text characters. The method may further include segmenting the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region. The method may further include inpainting each of the plurality of text characters with a predefined color.

[007] In another embodiment, a system for extracting text from a video stream is disclosed. The system may a processor and a memory which stores a plurality of instructions. The plurality of instructions, upon execution by the processor, may cause the processor to determine at least one visual text region in each image of a plurality of images. The video stream may include the plurality of images in sequential order. The at least one visual text region may include a plurality of text characters. Determining the at least one visual text region may be based on analysis of one of: lines and curves associated with each of the plurality of text characters of the at least one visual text region, a blob associated with the plurality of text characters along an axis, distribution of the plurality of text characters along a major axis (i.e. Horizontal axis or Vertical axis), and a common attribute associated with the plurality of text characters. The plurality of instructions, upon execution by the processor, may cause the processor to segment the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region, and upon segmenting, inpaint each of the plurality of text characters with a predefined color.

2.0

3.0 BRIEF DESCRIPTION OF THE DRAWINGS

[008] The accompanying drawings, which are incorporated in and constitute a part of this disclosure, illustrate exemplary embodiments and, together with the description, serve to explain the disclosed principles.

[009] **FIG. 1** illustrates an exemplary computing system that may be employed to implement processing functionality for various embodiments;

[010] **FIG. 2** is a functional block diagram of a text extraction device for extracting text from a video stream, in accordance with an embodiment of the present disclosure;

[011] **FIGs. 3A-3C** are snapshots of example video frames comprising text, in accordance with an embodiment;

[012] **FIG. 4** illustrates a process of extracting text from an example video frame of a video stream, in accordance with an embodiment of the present disclosure;

[013] **FIG. 5** is a flowchart of a method of extracting text from a video stream, in accordance with an embodiment of the present disclosure;

[014] **FIG. 6** is a flowchart of a method of analyzing lines and curves associated with each of the plurality of text characters of at least one possible visual text region, in accordance with an embodiment; and

[015] **FIG. 7** is a flowchart of a method of segmenting a plurality of text characters within at least one visual text region from a background associated with the at least one visual text region, in accordance with an embodiment.

4.0

5.0 DETAILED DESCRIPTION OF THE DRAWINGS

[016] Exemplary embodiments are described with reference to the accompanying drawings. Wherever convenient, the same reference numbers are used throughout the drawings to refer to the same or like parts. While examples and features of disclosed principles are described herein, modifications, adaptations, and other implementations are possible without departing from the spirit and scope of the disclosed embodiments. It is intended that the following detailed description be considered as exemplary only, with the true scope and spirit being indicated by the following claims. Additional illustrative embodiments are list.

[017] It should be noted that the textual information is assumed to include text characters on a background of a high contrasting color. It is further assumed that the text may possess one or more special features, for example properly shaped edges and corners (as the construction of script is majorly using lines and curves), text blobs (the text characters are stable connected components and distributed with respect to eight major geographical axes, i.e. the axes aligned with geographical East, West, South, North, North-East, North-West, South-East, and South-West direction), axial uniformity (text is homogenously distributed along a major axis), and that the text characters associated with each other may be written altogether.

[018] Referring now to **FIG. 1**, an exemplary computing system 100 that may be employed to implement processing functionality for various embodiments (e.g., as a SIMD device, client device, server device, one or more processors, or the like) is illustrated. Those skilled in the relevant art will also recognize how to implement the invention using other computer systems or architectures. The computing system 100 may represent, for example, a user device such as a desktop, a laptop, a mobile phone, personal entertainment device, DVR, and so on, or any other type of special or general-purpose computing device as may be desirable or appropriate for a given application or environment. The computing system 100 may include one or more processors, such as a processor 102 that may be implemented using a general or special purpose processing engine such as, for example, a microprocessor, microcontroller or other control logic. In this example, the processor 102 is connected to a bus 104 or other communication medium.

[019] The computing system 100 may also include a memory 106 (main memory), for example, Random Access Memory (RAM) or other dynamic memory, for storing information and instructions to be executed by the processor 102. The memory 106 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by the processor 102. The computing system 100 may likewise include a read only memory ("ROM") or other static storage device coupled to bus 104 for storing static information and instructions for the processor 102.

[020] The computing system 100 may also include storage devices 108, which may include, for example, a media drive 110 and a removable storage interface. The media drive 110 may include a drive or other mechanism to support fixed or removable storage media, such as a hard disk drive, a floppy disk drive, a magnetic tape drive, an SD card port, a USB port, a micro USB, an optical disk drive, a CD or DVD drive (R or RW), or other removable or fixed media drive. A storage media 112 may include, for example, a hard disk, magnetic tape, flash drive, or other fixed or removable medium that is read by and written to by the media drive 110. As these examples illustrate, the storage media 112 may include a computer-readable storage medium having stored therein particular computer software or data.

[021] In alternative embodiments, the storage devices 108 may include other similar instrumentalities for allowing computer programs or other instructions or data to be loaded into the computing system 100. Such instrumentalities may include, for example, a removable storage unit 114 and a storage unit interface 116, such as a program cartridge and cartridge interface, a removable memory (for example, a flash memory or other removable memory module) and memory slot, and other removable storage units and interfaces that allow software and data to be transferred from the removable storage unit 114 to the computing system 100.

[022] The computing system 100 may also include a communications interface 118. The communications interface 118 may be used to allow software and data to be transferred between the computing system 100 and external devices. Examples of the communications interface 118 may include a network interface (such as an Ethernet or other NIC card), a communications port (such as for example, a USB port, a micro USB port), Near field Communication (NFC), etc. Software and data transferred via the communications interface 118 are in the form of signals which may be electronic, electromagnetic, optical, or other signals capable of being received by the communications interface 118. These signals are provided to the communications interface 118 via a channel 120. The channel 120 may carry signals and may be implemented using a wireless medium, wire or cable, fiber optics, or other communications medium. Some examples of the

channel 120 may include a phone line, a cellular phone link, an RF link, a Bluetooth link, a network interface, a local or wide area network, and other communications channels.

[023] The computing system 100 may further include Input/Output (I/O) devices 122. Examples may include, but are not limited to a display, keypad, microphone, audio speakers, vibrating motor, LED lights, etc. The I/O devices 122 may receive input from a user and also display an output of the computation performed by the processor 102. In this document, the terms "computer program product" and "computer-readable medium" may be used generally to refer to media such as, for example, the memory 106, the storage devices 108, the removable storage unit 114, or signal(s) on the channel 120. These and other forms of computer-readable media may be involved in providing one or more sequences of one or more instructions to the processor 102 for execution. Such instructions, generally referred to as "computer program code" (which may be grouped in the form of computer programs or other groupings), when executed, enable the computing system 100 to perform features or functions of embodiments of the present invention.

[024] In an embodiment where the elements are implemented using software, the software may be stored in a computer-readable medium and loaded into the computing system 100 using, for example, the removable storage unit 114, the media drive 110 or the communications interface 118. The control logic (in this example, software instructions or computer program code), when executed by the processor 102, causes the processor 102 to perform the functions of the invention as described herein.

[025] It will be appreciated that, for clarity purposes, the above description has described embodiments of the invention with reference to different functional units and processors. However, it will be apparent that any suitable distribution of functionality between different functional units, processors or domains may be used without detracting from the invention. For example, functionality illustrated to be performed by separate processors or controllers may be performed by the same processor or controller. Hence, references to specific functional units are only to be seen as references to suitable means for providing the described functionality, rather than indicative of a strict logical or physical structure or organization.

[026] Referring now to **FIG. 2**, a block diagram of a text extraction device 200 for extracting text from a video stream is illustrated, in accordance with an embodiment of the present disclosure. The text extraction device 200 may include a visual text region determining module 202, a segmenting module 204, and an inpainting module 206.

[027] As will be understood, a video stream includes a plurality of video frames. Further, a set of frames from the plurality frames may include a text, for example movie credits (which may include a role and name of a person who performed that role, e.g. the role is "Director of

photography" and the name of the person is "John Smith"). It should be noted that the set of frames may be in image format (as such, the term "frame" and "image" may have been used interchangeably in this disclosure).

[028] The text extraction device 200 is configured extract the above-mentioned text from the set of images. To this end, the visual text region determining module 202 of the text extraction device 200 may identify a set of images which include text, from the plurality of images. In other words, the visual text region determining module 202 may identify the images which include text that may be processed further to extract the text, and leave the rest of the images which do not include any text.

[029] Once the set of images including text are identified, the visual text region determining module 202 may further determine at least one visual text region in each image. As mentioned earlier, the video stream may include the plurality of images in sequential order, and the at least one visual text region may include a plurality of text characters. For example, the plurality of text characters may include alphabets, numerical digits, special characters, etc. Further, the plurality of text characters may belong to any one or a combination of two or more scripts such as Roman script, Chinese script, Arabic script, Devanagari script, etc.

[030] Further, it should be noted that in an image containing text, the text may be present on a part of the image and confined within a visual text region. Further, the text may be infused with background which may be acting as a part of the text. It is, therefore, desirable to accurately identify the visual text region within the image to avoid analyzing the entire image for extracting the test. By first singling out one or more visual text region within the image, only the relevant regions within image are analyzed and therefore the efficiency of the process of text extraction is improved.

[031] Referring now to FIGs. 3A-3C, various snapshots of example video frames comprising text are illustrated, in accordance with an embodiment. **FIG. 3A** illustrates a first snapshot 300A comprising multiple visual text regions 302-1, 302-2, 302-3, 302-4, as marked in rectangular boxes. As will be understood, each of the visual text regions 302-1, 302-2, 302-3, 302-4 comprises text characters in Roman (English) script. Further, as shown in FIG. 3A, the visual text regions 302-1, 302-2 are present on a background which is substantially plain and White in color. Furthermore, the visual text regions 302-3, 302-4 are present on a background which is substantially plain and Black in color. **FIG. 3B** illustrates a second snapshot 300B comprising a visual text region 304, as marked in a rectangular box. The visual text region 304 comprises text characters in a Brahmic script. The text characters of the text region 304 are present on a graphical background, as a result of the which these text characters are infused with background, with the background acting as a part of the text characters. **FIG. 3C** illustrates a third snapshot 300C comprising a visual text

region 306, as marked in a rectangular box. The visual text region 306 comprises text characters in Roman (English) script, with the text characters of the text region 306 mostly being present on plain background White in color.

[032] Referring back to FIG. 2, in some embodiments, determining the at least one visual text region is based on analysis of one of: lines and curves associated with each of the plurality of text characters of the at least one visual text region, a blob associated with the plurality of text characters along an axis, distribution of the plurality of text characters along a major axis (i.e. Horizontal axis or Vertical axis), and a common attribute associated with the plurality of text characters. To this end, in some embodiments, the visual text region determining module 202 may include a lines and curves analyzing module 202A, a blob detecting and analyzing module 202B, a text characters distribution analyzing module 202C, and a text characters attribute analyzing module 202D.

[033] The lines and curves analyzing module 202A may analyze the lines and curves associated with each of the plurality of text characters of the at least one visual text region. It is assumed that the text in an image may possess shaper edges, since the construction of any script is by way of using lines and curves. Further, the text characters may be formed with intersection of lines and curves thereby forming corners. By way of an example, the lines and curves analyzing module 202A may analyze pixel intensity and average axial distance between these corners to estimate a visual text region. In particular, the lines and curves analyzing module 202A may detect one or more corners formed with intersection of the lines and curves associated with each of the plurality of text characters. Further, the lines and curves analyzing module 202A may determine pixel intensity and average axial distance between the one or more corners. Furthermore, the lines and curves analyzing module 202A may determine the at least one visual text region based on the pixel intensity and average axial distance. In some embodiments, an expression for detecting corners and checking axial uniformity in the possible visual text region may be as follows:

$$E(u,v) = \sum_{x,y} \underbrace{w(x,y)}_{\text{window function}} \underbrace{[I(x+u,y+v)}_{\text{shifted intensity}} - \underbrace{I(x,y)}_{\text{intensity}}]^2$$

... Equation (1)

$$\text{E (u, v)} \sim [\text{u} \quad \text{v}] \sum_{x,y} w(x,y) \begin{bmatrix} IxIx & IxIy \\ IxIy & IyIy \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

... Equation (2)

$$R = \det(M) - k(\text{trace}(M))^2$$

where,

$\det(M) = \lambda 1 \lambda 2$,

$\text{trace}(M) = \lambda 1 + \lambda 2$,

$\lambda 1$ and $\lambda 2$ are the eigen values of Matrix (M)

[034] Using the above expression, difference in intensities for a displacement of (u, v) may be calculated. It should be noted that window function may either be rectangular or gaussian window. The eigen values ($\lambda 1$ and $\lambda 2$) may determine whether a region is a corner, an edge, or a flat region.

[035] The blob detecting and analyzing module 202B may identify blob(s) of gray connected stable components within the image. For example, the blob detecting and analyzing module 202B may analyze the blob associated with the plurality of text characters along an axis based on maximally stable extremal visual text regions (MSER) model. It is assumed that the text characters within the visual text region are stable connected components and distributed with respect to eight major geographical axes (i.e. the axes aligned with geographical East, West, South, North, North-East, North-West, South-East, and South-West direction). As will be appreciated by those skilled in the art, the MSER model may be used as a method of blob detection in images, based on finding correspondences between image elements from two images with different viewpoints.

[036] The MSER features may be extracted from the images by identifying localized seed (Minima) and locating the maximum change in pixel value in multiple directions. The resultant points can be encapsulated in a rectangle. A rectangular blob region may be filtered based on histogram of height of the rectangle. The regions with consistent height may be within the configurable threshold limit. Using a set of points qualified as corners and text blobs (or MSER blobs), a gradient in x direction is determined for finding axial symmetry, as follows:

$$\text{gradient} = \nabla R(x,y) = \frac{\partial R}{\partial x}(x, y) \quad \ldots \text{Equation } (3)$$

[037] From the Equation (3), it may be found that the points are at equal distance axially. Further, points having similar heights in y direction may be determined using Equation (4) and Equation (5), as below:

$$H = \lim_{0 \to i} gradient(y_i) - gradient(y_{i+1}) \quad \ldots \text{Equation } (4)$$

$$\text{Qualified points } Q = \text{histogram}(\text{sort}(H)) \quad \ldots \text{Equation } (5)$$

[038] The text characters distribution analyzing module 202C may analyze homogenous distribution of the text along a major axis (i.e. Horizontal axis or Vertical axis). As will be understood, corners formed around the visual text may be distributed with very low gradient along

the major axis. Therefore, based on the analysis of the axial gradient distribution, corner points describing the visual text region may be identified. Once the corner points are identified, the visual text region may be determined.

[039] The text characters attribute analyzing module 202D may analyze the common attributes associated with the plurality of text characters. In some embodiments, the common attribute may be a font, a size, or a spacing associated with each of the plurality of text characters. It is assumed that text characters sharing common attributes are related to each other. In particular, for example, text characters which are written together (i.e., similar spacing) with same font size and color may be associated with each other. As such, for example, the text characters, i.e. alphabets of the term "Director of photography" of the above example may share common attributes in terms of font, size, or spacing. Therefore, the text characters attribute analyzing module 202D may determine a visual text region based on detection of occurrence of common attributes of the text characters. It should be noted that the text characters attribute analyzing module 202D may not necessarily identify the text characters, however, simply by analysing the presence of the above-mentioned common attributes, the text characters attribute analyzing module 202D may be able to determine a visual text region where text is present in an image. As such, the text characters attribute analyzing module 202D may cluster pixels with symmetrically axial distribution. The above is further explained in conjunction with FIG. 4.

[040] Referring now to FIG. 4, a process diagram of a process 400 of extracting text from a frame (image) 402 of a video stream is illustrated, in accordance with an embodiment. At step 404, a visual text region 410, as marked in a rectangular box, is determined, and at step 406, the visual text region 410 is extracted, i.e. separated from the rest of the image 402. As will be understood, the steps 404 and 406 may be performed by the visual text region determining module 202 of the text extraction device 200.

[041] Returning back to FIG. 2, once the one visual text region is identified by the visual text region determining module 202, the segmenting module 204 may segment the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region. In some embodiments, the segmenting may be performed based on k-means clustering. In particular, the process of segmenting may include classifying each pixel associated with the at least one visual text region into a predefined category of one or more predefined categories based on pixel intensity and color. Further, the segmenting may include generating one or more clusters corresponding to the one or more predefined categories, and identifying, from the one or more clusters, a cluster corresponding to text characters distinct from clusters corresponding to background.

[042] As will be appreciated, the visual text region may include a two-dimensional array of pixels. Further, in some instances, the text characters may be infused with the background, with the background acting as a part of the text characters. The segmenting module 204 may classify pixels associated with the text characters as distinct from the pixels associated with the background. Therefore, in order to segment, i.e. distinguish the text characters from the background, the pixels may be categorized into two or more categories, using k-means clustering. In some example embodiment, the pixels may be categorized into five categories (i.e. k = 5). As will be understood, some of these categories may correspond to the text while remaining categories may correspond to the background.

[043] Further, in some embodiments, a category associated with a pixel may be determined using averaging method. To this end, coordinates of corners of a probable visual text region may be determined. Thereafter, an average value of the coordinates may be obtained. The pixels around the coordinates of the corners may be categorized based on the average value. As such, one or more clusters of pixels corresponding to text characters may be determined. The one or more clusters corresponding to text characters may constitute the at least one visual text region. For example, as shown in FIG. 4, once the visual text region 410 is determined and extracted, text characters 412 from the visual text region 410 may be segregated, at step 408 by the segmenting module 204 of the text extraction device 200.

[044] Referring back to FIG. 2, once the plurality of text characters within the at least one visual text region are segmented from the background, each of these plurality of text characters may be inpainted with a predefined color. By way of an example, the each of the plurality of text characters may be inpainted with Black color on white canvas (background). In particular, the pixels categorized into a category corresponding to the text characters may be inpainted with Black color. Similarly, the pixels categorized into a category corresponding to the background may be inpainted with White color. Therefore, the plurality of text characters may be inpainted with Black color while the background region may be inpainted with White color.

[045] Referring now to FIG. 5, a flowchart of a method 500 of extracting text from a video stream is illustrated, in accordance with an embodiment of the present disclosure. The method 500 may be performed by the text extraction device 200. It should be noted that the method 500 may be directed to detecting a visual text region, and then estimating the background associated with the visual text regions and reconstructing the image with minimal contribution of original background. A video stream which includes a plurality of frames and from which text is to be extracted may be received by the text extraction device 200. As mentioned earlier, a set of frames from the plurality frames may include a text, for example movie credits (which may include a role

and name of a person who performed that role, e.g. the role is "Director of photography" and the name of the person is "John Smith"). The method 500 may first identify the set of frames from the plurality frames that may include text.

[046] In some embodiments, the set of images which include text may be identified from the plurality of images. Once the set of images including text are identified, at step 502, at least one visual text region from each image of a plurality of images may be determined. As will be understood, the video stream may include the plurality of images in a sequential order. The at least one visual text region may include a plurality of text characters. In some embodiments, determining the at least one visual text region may involve additional steps 502A-502D where an analysis of the images may be carried.

[047] At step 502A, lines and curves associated with each of the plurality of text characters of the at least one possible visual text region may be analyzed. The step 502A is further explained in detail in conjunction with FIG. 6.

[048] Referring now to **FIG. 6**, a flowchart of a method 600 of analyzing lines and curves associated with each of the plurality of text characters of the at least one possible visual text region is illustrated, in accordance with an embodiment. At step 602, one or more corners formed with intersection of the lines and curves associated with each of the plurality of text characters of the at least one visual text region may be detected. At step 604, pixel intensity and average axial distance between the one or more corners may be determined. At step 606, the at least one visual text region may be determined based on the pixel intensity and average axial distance.

[049] Returning back to FIG. 5, at step 502B, a blob associated with the plurality of text characters along an axis may be analyzed. The blob associated with the plurality of text characters along an axis may be analyzed based on maximally stable extremal visual text regions (MSER) model. At step 502C, distribution of the plurality of text characters along a major axis (i.e. Horizontal axis or Vertical axis) may be analyzed. At step 502D, a common attribute associated with the plurality of text characters may be analyzed. The common attribute may be one of a font, a size, or a spacing associated with each of the plurality of text characters.

[050] It should be noted that for any image, at least one of the above steps 502A-502D may be performed. In other words, in some images, it may be possible to determine the at least one visual text region by performing any one of the steps 502A-502D, while some images may require a combination of two or more of the steps 502A-502D. Furthermore, the sequence of performing the steps 502A-502D may not be restricted to any one particular sequence, and as such, the steps 502A-502D may be performed in any sequence.

[051] At step 504, the plurality of text characters within the at least one visual text region may be segmented from a background associated with the at least one visual text region. In some embodiments, the segmenting may be based on k-means clustering. The step 504 of segmenting is further explained in detail in conjunction with FIG. 7.

[052] Referring now to **FIG. 7**, a flowchart of a method 700 of segmenting the plurality of text characters within the at least one visual text region from the background associated with the at least one visual text region is illustrated, in accordance with an embodiment. At step 702, each pixel associated with the at least one visual text region may be classified into a predefined category of one or more predefined categories based on pixel intensity and color. At step 704, one or more clusters corresponding to the one or more predefined categories may be generated. At step 706, a cluster corresponding to text characters distinct from clusters corresponding to background may be identified from the one or more clusters.

[053] Returning back to FIG. 5, at step 506, upon segmenting, each of the plurality of text characters may be inpainted with a predefined color. By way of an example, the each of the plurality of text characters may be inpainted with Black color on a White background.

[054] The present disclosure discusses various techniques for accurately extracting visual texts from a video stream by removing complex background. The techniques provide optimized, cost effective, and one shot (i.e. with very few or no iterations required) techniques for removing the background and segmenting the text, by using clustering techniques, such as, but not limited to, K-means clustering. The above techniques are useful in media application, such as credit scenes digitization, subtitle verification, media library, metadata tagging, and the like.

[055] It is intended that the disclosure and examples be considered as exemplary only, with a true scope and spirit of disclosed embodiments being indicated by the following claims.

What We Claim Is:

1. A method of extracting text from a video stream, the method comprising:

determining at least one visual text region in each image of a plurality of images, wherein the video stream comprises the plurality of images in sequential order, wherein the at least one visual text region comprises a plurality of text characters, and wherein determining the at least one visual text region is based on analysis of one of:

lines and curves associated with each of the plurality of text characters of the at least one visual text region;

a blob associated with the plurality of text characters along an axis;

distribution of the plurality of text characters along a major axis; and

a common attribute associated with the plurality of text characters;

segmenting the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region; and

upon segmenting, inpainting each of the plurality of text characters with a predefined color.

2. The method as claimed in claim 1, wherein the analysis of lines and curves comprises:

detecting one or more corners formed with intersection of the lines and curves associated with each of the plurality of text characters of the at least one visual text region;

determining pixel intensity and average axial distance between the one or more corners; and

determining the at least one visual text region based on the pixel intensity and the average axial distance.

3. The method as claimed in claim 1, wherein the blob associated with the plurality of text characters along an axis is analyzed based on maximally stable extremal visual text regions (MSER) model.

4. The method as claimed in claim 1, wherein the common attribute is one of a font, a size, a color, or a spacing associated with each of the plurality of text characters.

5. The method as claimed in claim 2, wherein the segmenting is based on k-means clustering, and wherein the segmenting comprises:

classifying each pixel associated with the at least one visual text region into a predefined category of one or more predefined categories based on the pixel intensity and color;

generate one or more clusters corresponding to the one or more predefined categories; and

identifying, from the one or more clusters, a cluster corresponding to text characters distinct from clusters corresponding to background.

6. The method as claimed in claim 1, wherein each of the plurality of text characters is in painted with Black color.

7. A system for extracting text from a video stream, the system comprising:

a processor; and

a memory storing a plurality of instructions, wherein the plurality of instructions, upon execution by the processor, cause the processor to:

determine at least one visual text region in each image of a plurality of images, wherein the video stream comprises the plurality of images in sequential order, wherein the at least one visual text region comprises a plurality of text characters, and wherein determining the at least one visual text region is based on analysis of one of:

lines and curves associated with each of the plurality of text characters of the at least one visual text region;

a blob associated with the plurality of text characters along an axis;

distribution of the plurality of text characters along a major axis; and

a common attribute associated with the plurality of text characters;

segment the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region; and

upon segmenting, in paint each of the plurality of text characters with a predefined color.

8. The system as claimed in claim 7, wherein the analysis of lines and curves comprises:

detecting one or more corners formed with intersection of the lines and curves associated with each of the plurality of text characters of the at least one visual text region;

determining pixel intensity and average axial distance between the one or more corners; and

determining the at least one visual text region based on the pixel intensity and the average axial distance.

9. The system as claimed in claim 7,

wherein the blob associated with the plurality of text characters along an axis is analyzed based on maximally stable extremal visual text regions (MSER) model; and

wherein the common attribute is one of a font, a size, or a spacing associated with each of the plurality of text characters.

10. The system as claimed in claim 8, wherein the segmenting is based on k-means clustering, and wherein the segmenting comprises:

classifying each pixel associated with the at least one visual text region into a predefined category of one or more predefined categories based on the pixel intensity and color;

generate one or more clusters corresponding to the one or more predefined categories; and

identifying, from the one or more clusters, a cluster corresponding to text characters distinct from clusters corresponding to background.

AMENDED CLAIMS
received by the International Bureau on 09 May 2022 (09.05.2022)

[Claim 1] (Amended) A method of extracting text from a video stream, the method comprising:
determining, by the text extraction device (200), at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) in each image of a plurality of images (300A, 300B, 300C) from the video stream based on analysis of:
lines and curves associated with a plurality of text characters of the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306);
a blob associated with the plurality of text characters along an axis;
distribution of the plurality of text characters along a major axis; and
a common attribute associated with the plurality of text characters;
segmenting, by the text extraction device (200), the plurality of text characters within the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) from a background associated with the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306); and
upon segmenting, inpainting, by the text extraction device (200), each of the plurality of text characters with a predefined color.

[Claim 2] (Amended) The method as claimed in claim 1, wherein the analysis of lines and curves comprises:
detecting, by the text extraction device (200), one or more corners formed with intersection of the lines and curves associated with each of the plurality of text characters of the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306);
determining, by the text extraction device (200), pixel intensity and average axial distance between the one or more corners; and
determining, by the text extraction device (200), the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) based on the pixel intensity and the average axial distance.

[Claim 3] (Original) The method as claimed in claim 1, wherein the blob associated with the plurality of text characters along an axis is analyzed based on maximally stable extremal visual text regions (MSER) model.

[Claim 4] (Original) The method as claimed in claim 1, wherein the common attribute is one of a font, a size, a color, or a spacing associated with each of the plurality of text characters.

[Claim 5] (Amended) The method as claimed in claim 2, wherein the segmenting

is based on k-means clustering, and wherein the segmenting comprises: classifying, by the text extraction device (200), each pixel associated with the at least one visual text region into a predefined category of one or more predefined categories based on the pixel intensity and color; generating, by the text extraction device (200), one or more clusters corresponding to the one or more predefined categories; and identifying, by the text extraction device (200), from the one or more clusters, a cluster corresponding to text characters distinct from clusters corresponding to background.

[Claim 6] (Amended) The method as claimed in claim 1, wherein each of the plurality of text characters is inpainted with black color.

[Claim 7] (Amended) A system 100 for extracting text from a video stream, the system comprising:

a processor 102; and

a memory 106 storing a plurality of instructions, wherein the plurality of instructions, upon execution by the processor 102, cause the processor 102 to:

determine at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) in each image of a plurality of images (300A, 300B, 300C), wherein the video stream comprises the plurality of images (300A, 300B, 300C) in sequential order, wherein the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) comprises a plurality of text characters, and wherein determining the at least one visual text region is based on analysis of:

lines and curves associated with each of the plurality of text characters of the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306);

a blob associated with the plurality of text characters along an axis;

distribution of the plurality of text characters along a major axis; and

a common attribute associated with the plurality of text characters;

segment the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region; and

upon segmenting, inpaint each of the plurality of text characters with a predefined color.

[Claim 8] (Amended) The system as claimed in claim 7, wherein the analysis of lines and curves comprises:

detecting one or more corners formed with intersection of the lines and

curves associated with each of the plurality of text characters of the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306);
determining pixel intensity and average axial distance between the one or more corners; and
determining the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) based on the pixel intensity and the average axial distance.

[Claim 9] (Original) The system as claimed in claim 7,
wherein the blob associated with the plurality of text characters along an axis is analyzed based on maximally stable extremal visual text regions (MSER) model; and
wherein the common attribute is one of a font, a size, or a spacing associated with each of the plurality of text characters.

[Claim 10] (Amended) The system as claimed in claim 8, wherein the segmenting is based on k-means clustering, and wherein the segmenting comprises:
classifying each pixel associated with the at least one visual text region (302-1, 302-2, 302-3, 302-4, 304, 306) into a predefined category of one or more predefined categories based on the pixel intensity and color;
generating one or more clusters corresponding to the one or more predefined categories; and
identifying, from the one or more clusters, a cluster corresponding to text characters distinct from clusters corresponding to background.

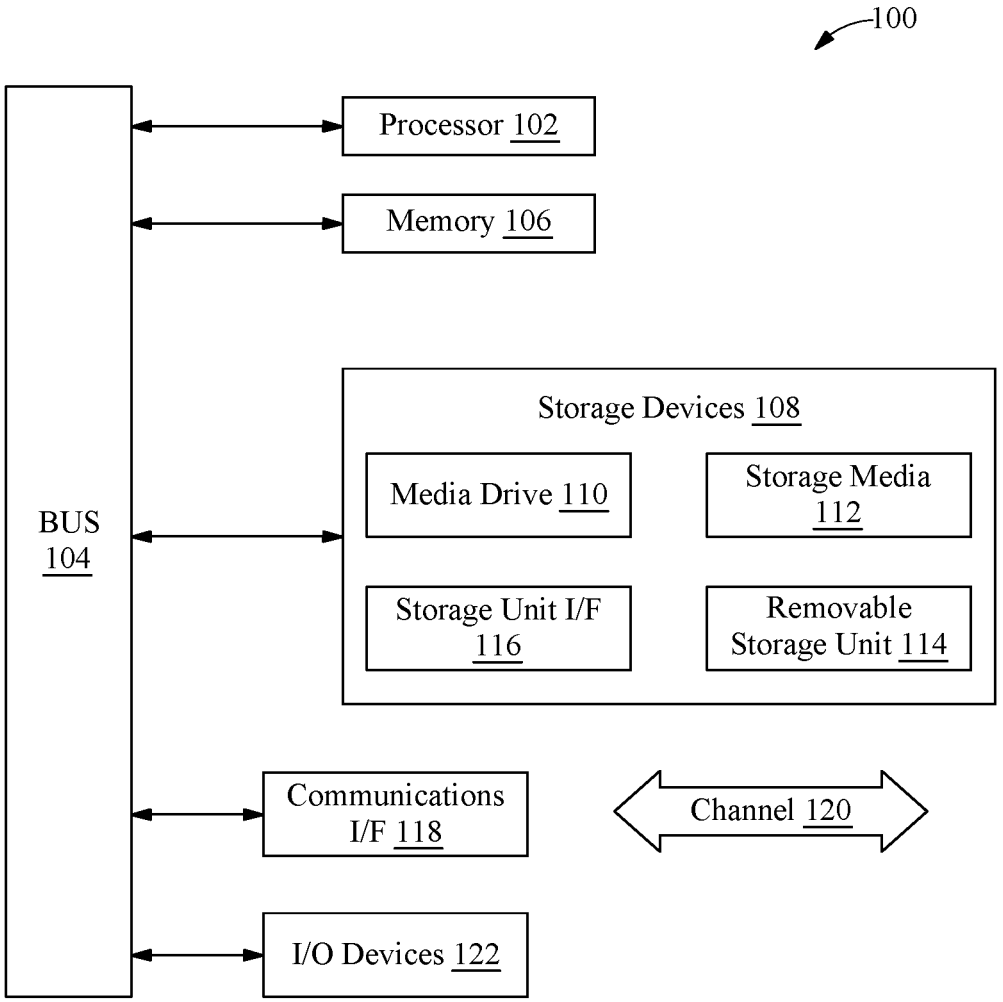

100
Processor 102
Memory 106
Storage Devices 108
Media Drive 110
Storage Media 112
Storage Unit I/F 116
Removable Storage Unit 114
BUS 104
Communications I/F 118
Channel 120
I/O Devices 122

FIG. 1

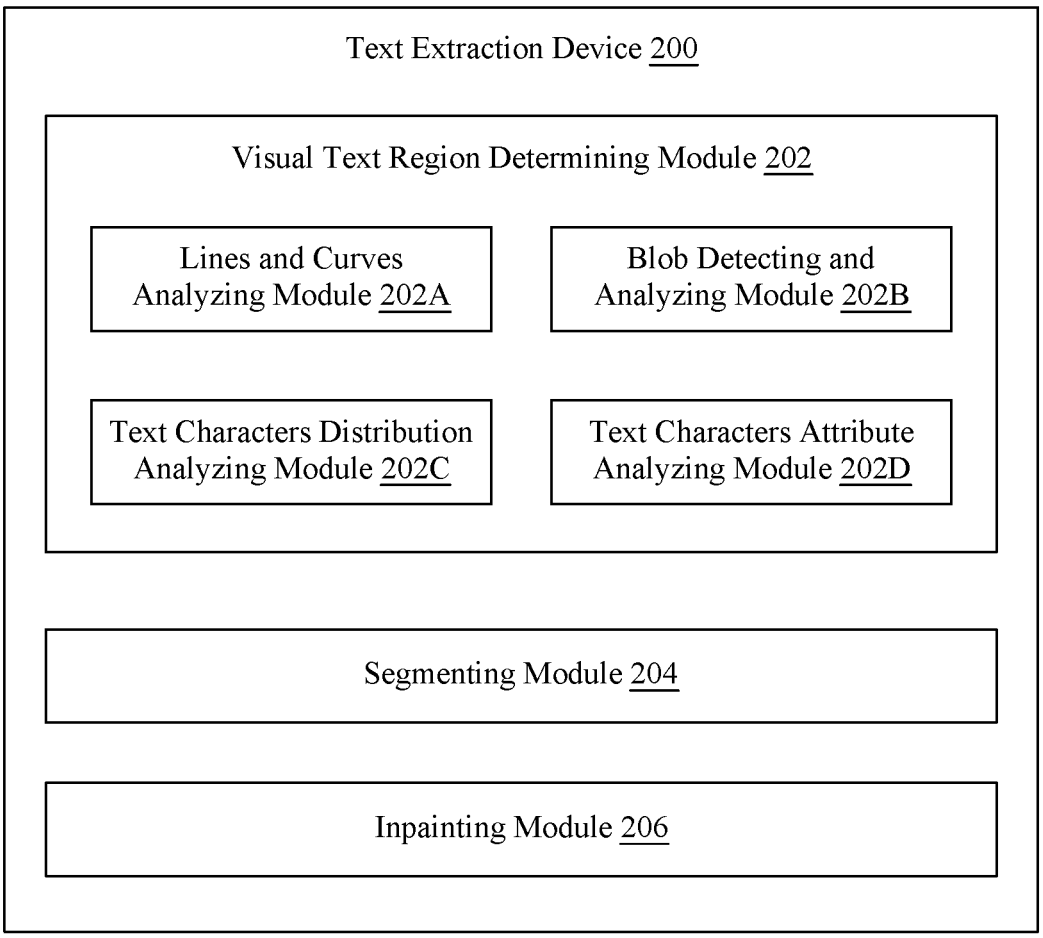

Text Extraction Device 200
Visual Text Region Determining Module 202
Lines and Curves Analyzing Module 202A
Blob Detecting and Analyzing Module 202B
Text Characters Distribution Analyzing Module 202C
Text Characters Attribute Analyzing Module 202D
Segmenting Module 204
Inpainting Module 206

FIG. 2

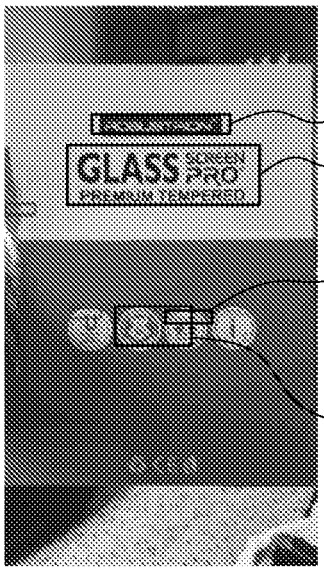


FIG. 3A

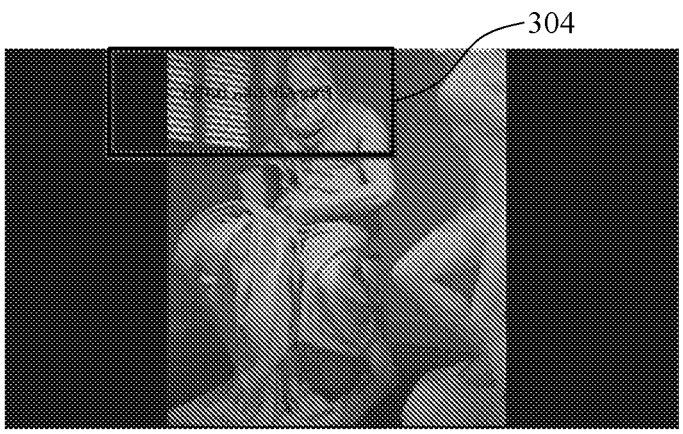


FIG. 3B

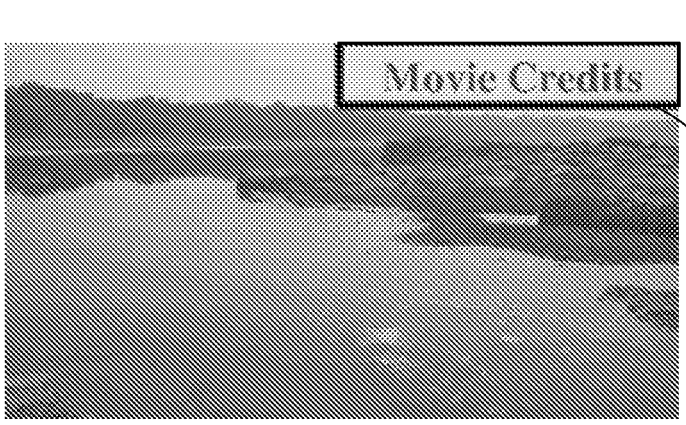


FIG. 3C

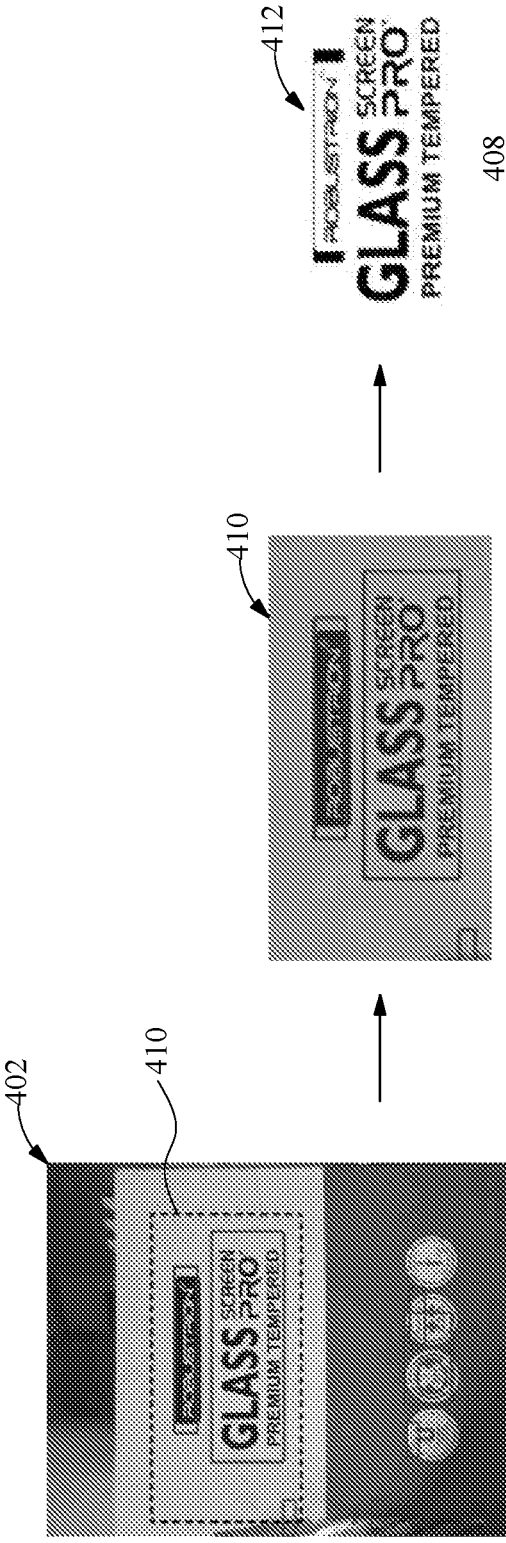

402
410
GLASS PRO
PREMIUM TEMPERED
404
410
406
412
GLASS PRO
PREMIUM TEMPERED
408

FIG. 4

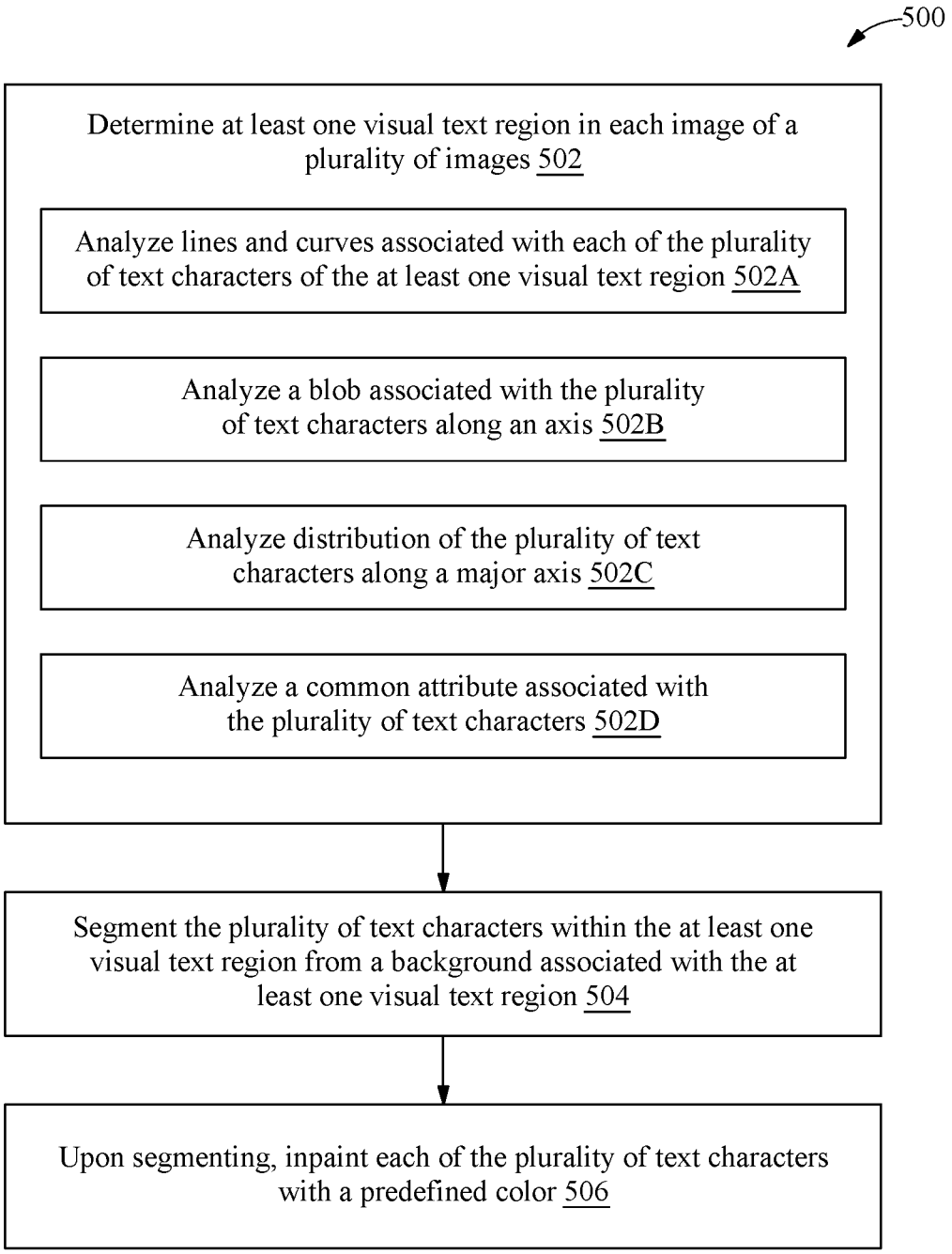

500
Determine at least one visual text region in each image of a plurality of images 502
Analyze lines and curves associated with each of the plurality of text characters of the at least one visual text region 502A
Analyze a blob associated with the plurality of text characters along an axis 502B
Analyze distribution of the plurality of text characters along a major axis 502C
Analyze a common attribute associated with the plurality of text characters 502D
Segment the plurality of text characters within the at least one visual text region from a background associated with the at least one visual text region 504
Upon segmenting, inpaint each of the plurality of text characters with a predefined color 506

FIG. 5

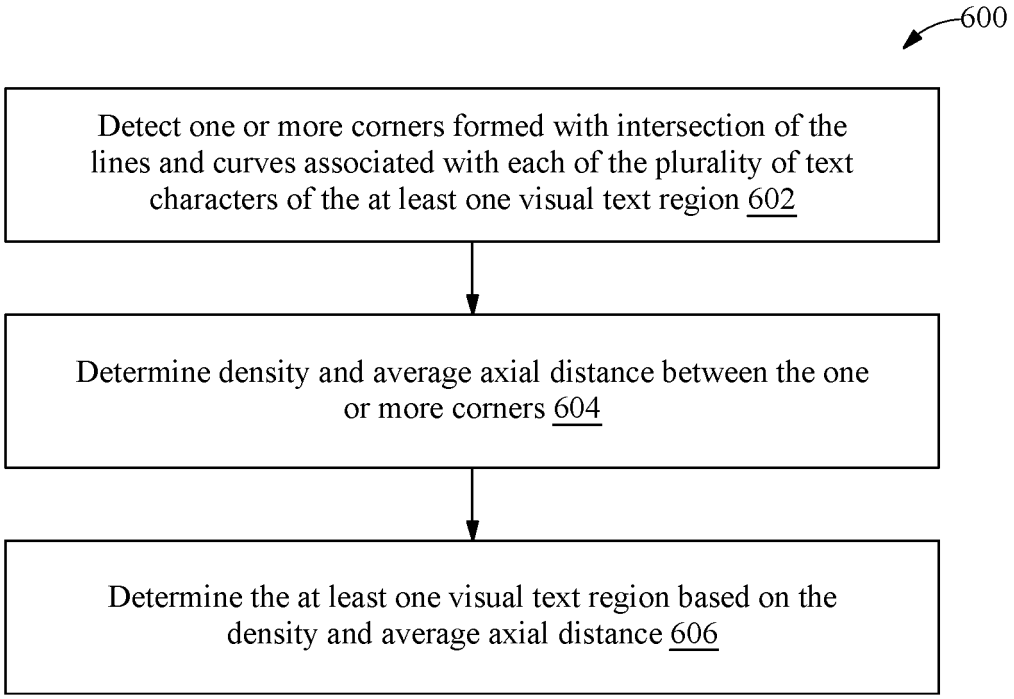

600
Detect one or more corners formed with intersection of the lines and curves associated with each of the plurality of text characters of the at least one visual text region 602
Determine density and average axial distance between the one or more corners 604
Determine the at least one visual text region based on the density and average axial distance 606

FIG. 6

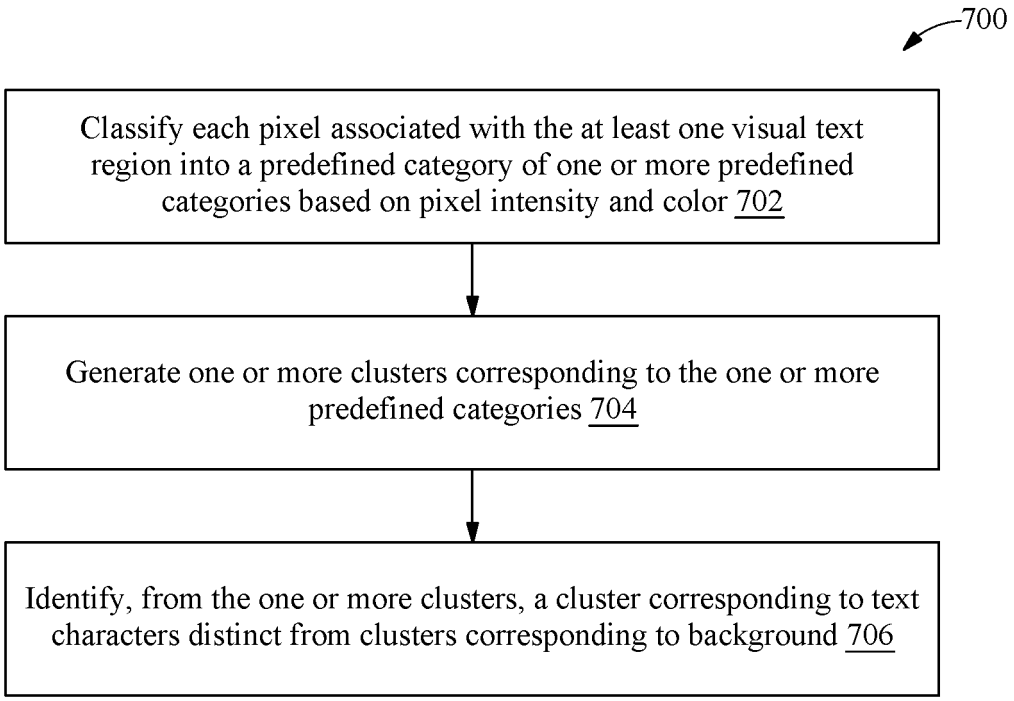

700
Classify each pixel associated with the at least one visual text region into a predefined category of one or more predefined categories based on pixel intensity and color 702
Generate one or more clusters corresponding to the one or more predefined categories 704
Identify, from the one or more clusters, a cluster corresponding to text characters distinct from clusters corresponding to background 706

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2021/060976

A. CLASSIFICATION OF SUBJECT MATTER
G06T7/00,G06V10/40 Version=2022.01

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)

G06T,G06V

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)

Databases accessed:PatSeer, IPO Internal Database
Keywords: image,feature extract,text,blob,lines,curves,colour

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US6473524B1 (Videk Inc) 29 October 2002 (29/10/2002) (whole document)	1-10
Y	US20110298922A1 (EYECUE VISION TECHNOLOGIES Ltd) 08 December 2011 (08/12/2011) (abstract, figures 5, 7, 8, paragraphs 0033, 0045-0050)	1-10

☐ Further documents are listed in the continuation of Box C. ☒ 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
"D" document cited by the applicant in the international application
"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	Date of mailing of the international search report
04-03-2022	04-03-2022

Name and mailing address of the ISA/	Authorized officer
Indian Patent Office Plot No.32, Sector 14,Dwarka,New Delhi-110075 Facsimile No.	Rahul Gahlan Telephone No. +91-1125300200

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2021/060976

Citation	Pub.Date	Family	Pub.Date
US 2011298922 A1	08-12-2011	CN 102713931 A	03-10-2012
		EP 2462537 A1	13-06-2012
		JP 2013501304 A	10-01-2013
		US 2016228772 A1	11-08-2016
		WO 2011017393 A1	10-02-2011