

Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner,

v.

CELLULAR SOUTH, INC.,
Patent Owner.

Case IPR2025-00876
Patent 10,218,954 B2
Issue Date: February 26, 2019

Title: VIDEO TO DATA

**PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 10,218,954 B2**

Table of Contents

	Page
I. MANDATORY NOTICES UNDER §42.8(A)(1)	1
A. Real Party-In-Interest under §42.8.(b)(1).....	1
B. Related Matters under §42.8(b)(2)	1
C. Lead and Back-Up Counsel under §42.8(b)(3)	1
D. Service Information	2
II. FEE PAYMENT	2
III. REQUIREMENTS UNDER §§ 42.104 AND 42.108 AND CONSIDERATIONS UNDER §§ 314(A) AND 325(D).....	2
A. Grounds for Standing	2
B. Identification of Challenge and Statement of Precise Relief Requested	3
C. Considerations Under §§ 314(a) and 325(d).....	3
IV. OVERVIEW OF THE PATENT.....	5
A. Level of Ordinary Skill	5
B. Specification Overview	5
V. CLAIM CONSTRUCTION	10
VI. THE CHALLENGED CLAIMS ARE UNPATENTABLE.....	10
A. Overview of Grounds.....	10
B. Prior Art Status of Relied-Upon References	11
C. Ground 1: Claims 1-8 and 11-12 Are Obvious Over Kritt in view of Fontana	11
1. Independent Claim 1: “A method to generate video data from a video comprising:” (Claim 1[pre]).....	11
(a) “generating audio files and image files from the video;” (Claim 1[a]).....	13

Table of Contents

(continued)

	Page
(b) “distributing the audio files and the image files across a plurality of processors and processing the audio files and the image files in parallel;” (Claim 1[b])	21
(c) “converting audio files associated with the video to text;” (Claim 1[c])	24
(d) “identifying an object in the image files;” (Claim 1[d])	24
(e) “determining a contextual topic from the image files;” (Claim 1[e]).....	26
(f) “assigning a probability of accuracy to the identified object based on the contextual topic;” (Claim 1[f])	29
(g) “converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic;” (Claim 1[g])	32
(h) “cross-referencing the text and the video data with the video to determine contextual topics;” (Claim 1[h])	36
(i) “generating a contextual text, an image, or an animation based on the determined contextual topics;” (Claim 1[i]).....	40
(j) “generating a content-rich video based on the generated text, image, or animation.” (Claim 1[j])	42
2. Claim 2: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”	44
3. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to extract a topic.”	46

Table of Contents
(continued)

	Page
1. Claim 4: “The method according to claim 1, further comprising: processing the image files to extract additional text.”	46
2. Claim 5: “The method according to claim 4, wherein the additional text is generated by segmenting the image files before processing the image files in parallel.”	47
3. Claim 6: “The method according to claim 1, further comprising: processing the image files to extract an object and determine a motion associated with the object.”	50
4. Claim 7: “The method according to claim 1, wherein the text is generated by segmenting the audio before processing the audio files in parallel.”	50
5. Claim 8: “The method according to claim 6, wherein the audio files and the image files are segmented at spectrum thresholds.”	53
6. Claim 11: “The method according to claim 1, further comprising: generating metadata.”	54
7. Claim 12: “The method according to claim 11, wherein the metadata comprises a number of frames, a range of said frames and/or timestamp references.”	55
D. Ground 2: Claims 2-3 and 9-10 Are Obvious Over Kritt in view of Fontana and Lau	56
1. Claim 2: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”	56
2. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to extract a topic.”	60
3. Claim 9: “The method according to claim 1, further comprising: generating an advertisement based on the text and the video data.”	60

Table of Contents
(continued)

	Page
4. Claim 10: “The method according to claim 9, further comprising: placing the advertisement in the video at a preferred time.”	63
E. Ground 3: Claim 13 Is Obvious Over Kritt.....	64
1. Independent Claim 13: “A system for extracting data from a video, comprising:” (Claim 13[pre]).....	64
(a) “a computer processor; and” (Claim 13[a]).....	64
(b) “a non-transitory computer readable medium containing instructions directing the system to execute the steps of:” (Claim 13[b]).....	64
(c) “converting audio associated with the video to text;” (Claim 13[c])	65
(d) “converting images associated with the video to video data, wherein the video data comprises identified objects;” (Claim 13[d])	65
(e) “extracting topics from the text and the video data;” (Claim 13[e]).....	65
(f) “assigning probabilities of accuracy for each of the identified objects, wherein the probabilities are based on the extracted topics;” (Claim 13[f]).....	67
(g) “cross-referencing the text, the video data, and the topics with the video to determine contextual topics;” (Claim 13[g]).....	67
(h) “generating a contextual text, an image, or an animation based on the determined contextual topics; and” (Claim 13[h]).....	68
(i) “generate a content-rich video based on the generated contextual text, image, or animation.” (Claim 13[i])	68
VII. CONCLUSION.....	68
CERTIFICATE OF SERVICE	70

Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

List of Exhibits

Exhibit No.	Description of Document
1001	U.S. Patent No. 10,218,954 B2 to Naeem Lakhani (PCT filed February 7, 2015, issued February 26, 2019) (“ 954 ” or “ 954 patent ”)
1002	Declaration of Henry Houh, Ph.D. in Support of Petition for <i>Inter Partes</i> Review
1003	U.S. Patent App. Pub. No. 2014/0181668 A1 to Barry A. Kritt et al. (filed Dec. 20, 2012, published Jun. 26, 2014) (“ Kritt ”)
1004	U.S. Patent App. Pub. No. 2012/0078712 A1 to James A. Fontana et al. (filed Sept. 27, 2010, published Mar. 29, 2012) (“ Fontana ”)
1005	U.S. Patent App. Pub. No. 2007/0112630 A1 to Wai Kit Lau et al. (filed Nov. 7, 2006, published May 17, 2007) (“ Lau ”)
1006	Prosecution History of U.S. Patent No. 10,218,954 B2
1007	U.S. Patent No. 9,697,230 to Henry Houh et al. (filed Mar. 31, 2006, issued Jul. 4, 2017)
1008	U.S. Patent No. 8,706,664 to Oliver B. Downs et al. (filed May 3, 2012, issued Apr. 22, 2014) (“ Downs ”)
1009	U.S. Patent No. 9,245,254 to Richard T. Lord et al. (filed Feb. 15, 2012, issued Jan. 26, 2016) (“ Lord ”)
1010	Proof of Service of Complaint

I. MANDATORY NOTICES UNDER §42.8(A)(1)

A. Real Party-In-Interest under §42.8.(b)(1)

Google LLC (“Petitioner”) is the real party-in-interest to this IPR petition.¹

B. Related Matters under §42.8(b)(2)

The ’954 patent was the subject of pending litigation involving Petitioner: *Cellular South, Inc. v. Google LLC*, Case No. 4:25-cv-01487-YGR (N.D. Cal.). Petitioner was served on May 10, 2024. (EX1010, p.001.) The case was originally filed in the Western District of Texas and was transferred to the Northern District of California on February 12, 2025.

C. Lead and Back-Up Counsel under §42.8(b)(3)

Petitioner provides the following designation of counsel.

LEAD COUNSEL	BACK-UP COUNSEL
Heidi L. Keefe (Reg. No. 40,673) hkeefe@cooley.com COOLEY LLP ATTN: Patent Group 1299 Pennsylvania Ave. NW, Suite 700 Washington, DC 20004 Tel: (650) 843-5001 Fax: (650) 849-7400	Andrew C. Mace (Reg. No. 63,342) amace@cooley.com Mark R. Weinstein (Admission <i>pro hac</i> <i>vice</i> to be requested) mweinstein@cooley.com

¹ Google LLC is a subsidiary of XXVI Holdings Inc., which is a subsidiary of Alphabet Inc. XXVI Holdings Inc. and Alphabet Inc. are not real parties in interest to this proceeding.

LEAD COUNSEL	BACK-UP COUNSEL
	Reuben Chen (Admission <i>pro hac vice</i> to be requested) rchen@cooley.com Alexandra D. Leeper (Admission <i>pro hac vice</i> to be requested) aleeper@cooley.com COOLEY LLP ATTN: Patent Group 1299 Pennsylvania Ave. NW, Suite 700 Washington D.C. 20004

D. Service Information

This Petition is being served by Federal Express to the attorney of record for the '954 patent, 27890 - STEPTOE LLP/DC, 1330 CONNECTICUT AVENUE, N.W., WASHINGTON, DC 20036. Petitioner consents to electronic service at the addresses provided above for lead and back-up counsel.

II. FEE PAYMENT

Petitioner requests review of 13 claims, with a \$51,875 payment.

III. REQUIREMENTS UNDER §§ 42.104 AND 42.108 AND CONSIDERATIONS UNDER §§ 314(A) AND 325(D)

A. Grounds for Standing

Petitioner certifies that the '954 patent is available for IPR and that Petitioner is not barred or otherwise estopped.

B. Identification of Challenge and Statement of Precise Relief Requested

Petitioner requests IPR institution based on:

Ground	Claims	Basis for Challenge under §103
1	1-8, 11-12	Kritt in view of Fontana
2	2-3, 9-10	Ground 1 prior art in further view of Lau
3	13	Kritt

Submitted with this Petition is the Declaration of Henry Houh, Ph.D. (EX1002) (“Houh”), a qualified technical expert. (EX1002, ¶¶1-15, App. A.)

C. Considerations Under §§ 314(a) and 325(d)

Petitioner hereby stipulates that if this IPR is instituted, then Petitioner will not pursue in the related pending litigation the specific grounds of invalidity that were raised or that reasonably could have been raised under 35 U.S.C. §§ 102 or 103 on the basis of prior art patents or printed publications in this IPR. To avoid any doubt, if the PTAB declines institution or rescinds institution of IPR, then Petitioner reserves the right to pursue any grounds of invalidity, including but not limited to the grounds raised or that reasonably could have been raised in this IPR, in the related pending litigation.

Petitioner respectfully submits that there is no section 314(a) or 325(d) issue that would warrant discretionary denial of the Petition.

Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

§314(a): The *General Plastic* factors are not relevant; this is the first and only IPR petition filed by Petitioner with respect to the '972 patent.

Nor do the *Fintiv* factors support discretionary denial under §314(a). As noted above, the pending litigation involving Petitioner was recently transferred to the Northern District of California. Before transfer, the pending litigation was still in an early pre-Answer stage, with no substantive discovery or claim construction having taken place. Since being transferred to the Northern District of California, the litigation has not substantively progressed. An initial case management conference is currently set for June 20, 2025. There is no schedule or trial date. Petitioner also intends to move to stay the litigation pending resolution of IPR.

§325(d): *Advanced Bionics* does not apply to Grounds 1 and 3, which rely exclusively on art that was not presented during prosecution. As to Ground 2, Lau was not distinguished by the applicant, relied upon in any ground of rejection, cited in combination with any references, nor considered in combination with Kritt and Fontana. It was mentioned by the Examiner in passing solely as “rendering advertisements with rich media,” consistent with its use in this Petition. (EX1007, p.0069.)

In accordance with the Director's March 26, 2025 Interim Processes for PTAB Workload Management memorandum and the related FAQs², Petitioner reserves the right to address and respond to any assertions that Patent Owner may raise regarding discretionary factors in a Discretionary Denial Brief or otherwise.

IV. OVERVIEW OF THE PATENT

A. Level of Ordinary Skill

A person of ordinary skill would have possessed a bachelor's degree in electrical engineering, computer science, or similar field, with two years of experience in developing and implementing computer software for processing and/or analyzing multimedia content, such as audio, video, or image data. A person could also have qualified as a person of ordinary skill in the art with some combination of (1) more formal education (such as a master's of science degree) and less technical experience, or (2) less formal education and more technical or professional experience. (EX1002, ¶¶18-23.)

B. Specification Overview

The '954 patent purports to offer one solution to the common problem of how to reduce audio and visual information to written form, in order to provide additional

² See <https://www.uspto.gov/patents/ptab/faqs/interim-processes-workload-management>.

content-matched audio-visual information.

The '954 patent describes its purpose at a very high level. It states as its “Technical Field” that it “relates to a method and a system for generating various and useful data from videos.” ('954, 1:18-19.)³ And the summary of the invention confirms the broad brush the patent purports to take: “The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.” ('954, 1:36-38.)

Audio-visual data such as videos long predates the '954 patent, and prior artists similarly had long been working on various ways to analyze, search, and generate video content. By the time of the alleged priority date (August 15, 2013), the fields of audio and image analysis for video content were well-developed. Indeed, the Background section of the '954 patent acknowledges known image searching techniques to identify matching and similar images, as well as audio-to-text algorithms for transcribing text from audio. ('954, 1:23-32.)

The specification of the '954 patent provides primarily a series of high-level, functional explanations for how to implement the alleged invention, with scant information on how to carry out any of its steps. For example, Figure 1 of the '954 patent, shown below, “illustrates an embodiment of present invention,” and

³ All underlining added to quoted material unless noted.

specifically “[a]n embodiment of video-to-content engine operation”:

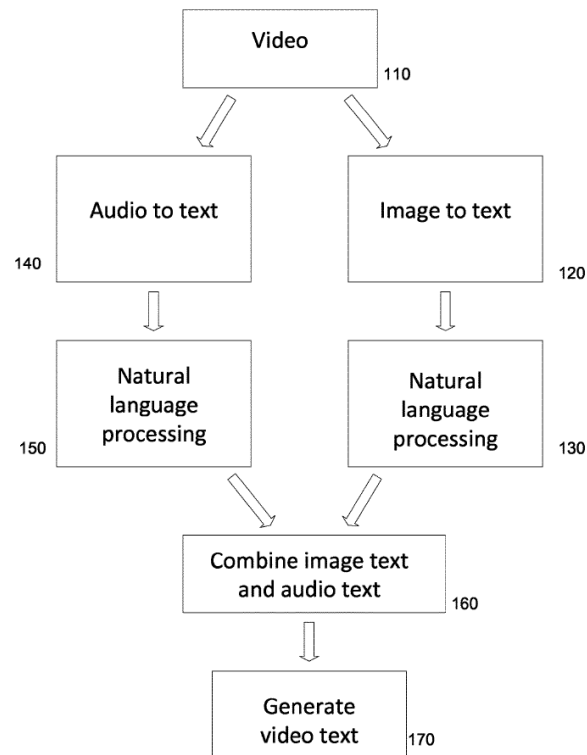


FIG. 1

(’954, Fig. 1, 2:24, 4:9-10.) At step **110**, “a video stream is presented.” (’954, 4:10.) “The engine can extract audio data and image data (e.g. images or frames forming the video) from the video stream.” (’954, 4:16-17.) At step **120**, “the video-to-content engine performs image data processing on the video stream,” such as symbol (or object) detection and identification, such as identifying a dog in the video image data. (’954, 4:22-23, 4:28-35.) Then, “[b]ased on the identified symbol, a plurality of instances of the symbol can be compared to a topic database to identify a topic (such as an event)[,]” such as identifying the dog as running or jumping. (’954, 4:44-47.) Thus, “text describing a symbol of the video and topic relating to the symbol

may be generated[.]” (’954, 4:50-52.) At step **130**, “natural language processing can be applied to the image text.” (’954, 6:31-32.) “For example, based on dictionary, grammar, and a knowledge database, the text extracted from video images can be modified as the video-to-content engine selects primary semantics from a plurality of possible semantics.” (’954, 6:32-36.) Applying dictionary definitions and grammatical rules to text was nothing new in 2013.

“In parallel, at **140**, the video-to-content engine can perform audio-to-text processing on audio data associated with the video.” (’954, 6:41-43.) At step **150**, “natural language processing can be applied to the text” and the text “can be given context, an applied sentiment, and topical weightings.” (’954, 7:36-40.) Again, the specification is largely silent as to how to implement these natural language processing tools, which would have been well-known to a POSA. (’954, 1:23-32 (acknowledging audio-to-text was known).)

At step **160**, “the topics generated from an image or a frame and the topics extracted from audio can be combined. The text can be cross-referenced, and topics common to both texts would be given additional weights.” (’954, 7:41-44.) In other words, this is the common step of finding a way to turn an apples-to-oranges comparison (audio and video) into apples-to-apples (give everything a text descriptor or “topic” that can be combined and compared).

At step **170**, “the video-to-content engine generates video text, such as text

describing the content of the video, using the result of the combined texts and cross reference.” (’954, 7:44-47.) This is just summarizing the video, such as what a theater critic or museum curator would do for a play or a work of multimedia art. “For example, key words indicating topic and semantic that appear in both texts can be selected or emphasized.” (’954, 7:47-49.) An “advertisement (text, images, or animation)” can then be generated based on the video text and added to the video stream. (’954, 8:60-63.)

In connection with identifying objects in an image, the ’954 patent further discloses using context, such as information related to an audio soundtrack or a surrounding sequence of frames, to assist with accurate identification:

As one example, if the audio soundtrack of a video is an announcer calling a football game, then identification of ball in a given frame as a basketball can be assigned a reduced probability or weighting. As another example of using context, if a given series of image frames from a video is positively or strongly identified as a horse race, then identifying an object to be a mule or donkey can be given a reduced weight. . . .

[T]he linear sequencing of frames from a video—as opposed to a stand-alone image—avails itself to a set images {images x-y} from which context can be derived. This contextual methodology can be viewed as systematic detection of probable image false positives by identifying an object from one video frame (or image) as an anomaly when compared to and associated with a series of image frames both prior and

subsequent to the purported anomaly. According to the objects, faces, words, etc. of a given set of frames (however defined), a probability can be associated with an identified anomaly to determine whether an image may be a false positive and, if so, what other likely results should be.

(’954, 5:8-16, 5:22-34.) As the above quotation shows, the concept of identifying “context” in the ’954 patent is functional, rather than based on a specific technical implementation.

V. CLAIM CONSTRUCTION

Petitioner does not believe express claim construction is necessary at this time. The prior art cited herein, as demonstrated below, renders the challenged claims obvious under any reasonable construction. Petitioner therefore respectfully submits that, for purposes of this IPR, express constructions are not required.

VI. THE CHALLENGED CLAIMS ARE UNPATENTABLE

A. Overview of Grounds

The Petition primarily relies on Kritt (**EX1003**) and Fontana (**EX1004**), which like the ’954 patent disclose techniques for parallel processing of the audio and image data from video content and using the resulting analysis to generate additional content to add to the video. (Houh, ¶¶55-59, 62-64.) The Petition cites an additional reference, Lau (**EX1005**), in **Ground 2**, which describes matching advertisements to video content. (Houh, ¶¶67-69.)

B. Prior Art Status of Relied-Upon References

Because the '954 patent claims priority to an earliest application filed August 15, 2013, AIA law applies to the challenged claims. All the references cited in the Grounds qualify as prior art under §102(a)(2) because they were each patent applications filed before August 15, 2013. Fontana and Lau are additionally prior art under §102(a)(1) because they were published before August 15, 2013.

C. Ground 1: Claims 1-8 and 11-12 Are Obvious Over Kritt in view of Fontana

1. Independent Claim 1: “A method to generate video data from a video comprising:” (Claim 1[pre])

Assuming the preamble provides a claim limitation, Kritt discloses it. Kritt discloses **generat[ing] video data from a video**, such as tags and other meta-data describing the video generated from performing audio and/or video processing of the video.

For example, Kritt discloses that “a viewer selects a video, and human, nonhuman, and audio objects of the video are identified.” (Kritt, ¶0017.) Kritt implements its method using, for example, a “video processing module **152**, audio processing module **154**, and control module **156**,” as illustrated in Kritt’s Figure 1:

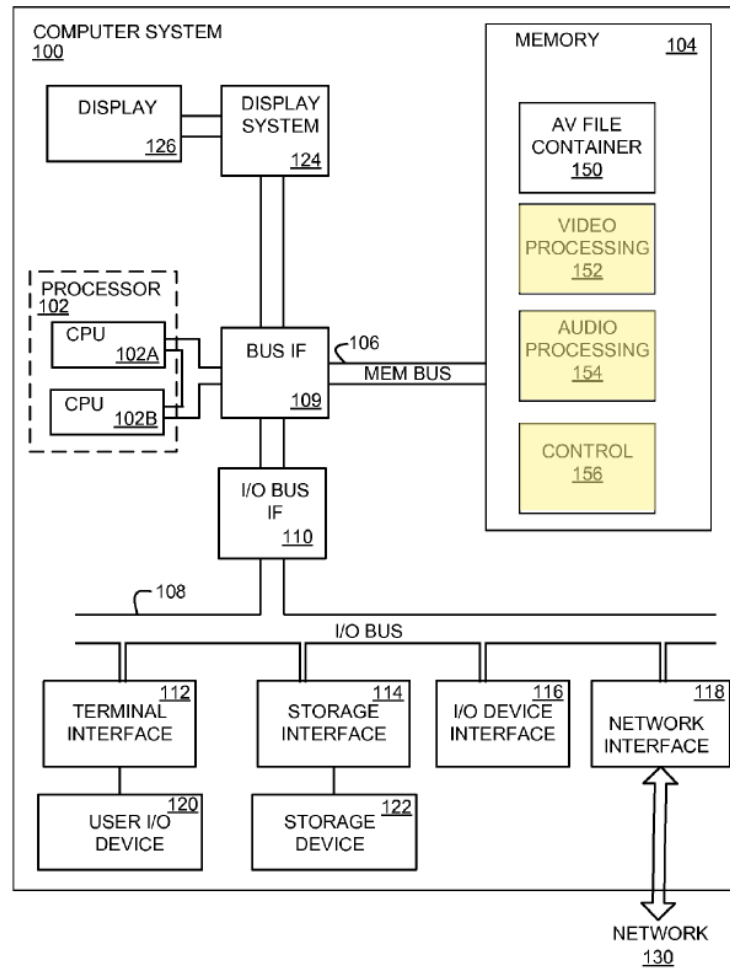


FIG. 1

(Kritt, ¶0023, Fig. 1) The video and audio processing modules are used to “generate visual tags” and “generat[e] audio and key word tags” describing the content of the video. (Kritt, ¶0024.) Kritt’s “visual tags” and “audio and key word tags” are each data about a video, generated from the video, making them “**video data from a video.**”

Therefore, Kritt discloses “**a method to generate video data from a video.**”
(Houh ¶¶72-74.) And as explained below, the method as claimed is obvious over
Kritt and Fontana.

(a) “**generating audio files and image files from the video;**”
(Claim 1[a])

Claim 1[a] is disclosed by and obvious over Kritt in view of **Fontana**. (Houh
¶¶75-88.)

Kritt teaches that the video for processing may be stored “using a container or
wrapper file format,” which includes “video files, audio files, subtitle files, chapter-
information files, metadata, and other files,” as illustrated in Kritt’s Figure 2. (Kritt,
¶0002.)

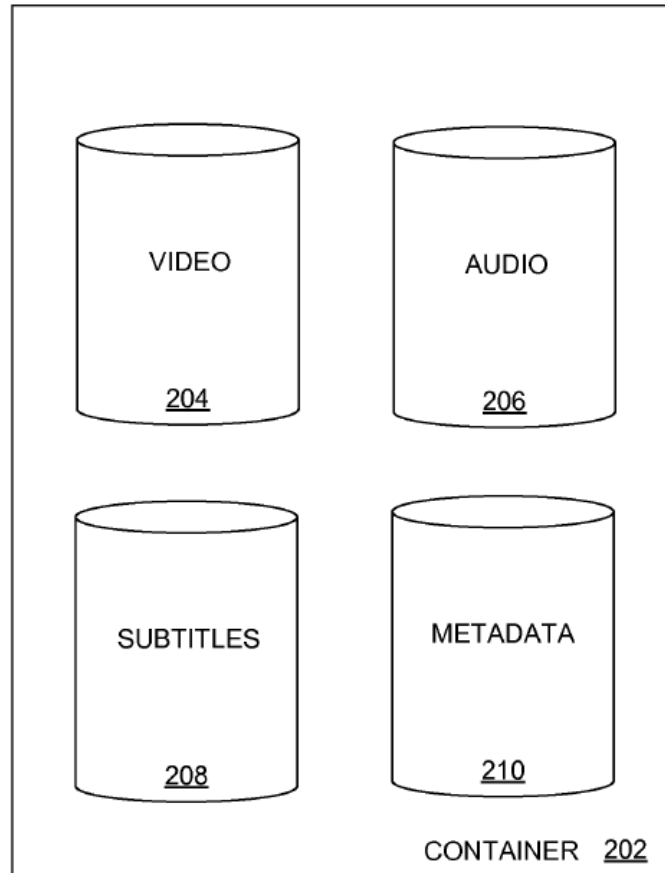


FIG. 2

(Kritt, ¶0031, Fig. 2.) Such containers were well-known as they were a common way to store the various components of multimedia content such as videos. (Houh ¶76.)

Kritt discloses **generating image files from the video**, in the form of “shot files” and extracted video frames, each of which is an image file. (Houh ¶¶77-78.) As to “shot files,” Kritt explains that “a shot may be a continuous sequence of frames captured without interruption by a camera oriented in a single direction or camera

angle.” (Kritt, ¶0039.) A POSA would recognize a frame as an image, and would further appreciate that the disclosed “shot files” are image files, in that they consist of a sequence of frames (that is, images). (Houh ¶77.) Generating shot files from videos was well-known, as Kritt acknowledges: “A video file **204** may be parsed into shot files according to any known method.” (Kritt, ¶0040 (describing one method using histograms).)

Kritt additionally discloses extracting video frames from the video, which are image files. (Houh ¶78.) In the example above for parsing shot files, Kritt analyzes “frames” to generate the shot files. (See Kritt, ¶0040.) Kritt further discloses extracting a “key frame” “for each shot file.” (Kritt ¶0041.) “The key frame may be deemed to be representative of all frames in the shot, permitting descriptive data for the shot to be determined only for the key frame and not for every frame of the shot.” (Kritt, ¶0041.)

Kritt also renders obvious **generating audio files from the video**. Kritt describes storing an audio file as part of the video container file, and separate modules for processing video and audio. (Krit ¶0022.) Kritt further discloses that the audio and video processing modules “may be on different computer systems and may be accessed remotely.” (Kritt, ¶0022.) A POSA would have understood and found it obvious that audio files and image files would be generated from the video in order to distribute them to the appropriate modules for processing. (Houh ¶79.)

Kritt further explains that “one or more audio features or audio signal descriptors may be extracted from an audio file.” (Kritt, ¶0056.) To extract audio features or audio signal descriptors from an audio file, the audio file must first be extracted from the container. (Houh ¶79.)

Furthermore, the combination of Kritt with Fontana renders obvious **generating audio files and image files from the video.** (Houh ¶¶80-84.)

Fontana teaches “processing of multimedia content at an audio processing module 606, [and] a video processing module 608” where “[e]ach of these modules can be executed concurrently (e.g., in parallel), with jobs associated with each module operating on one or more computing systems.” (Fontana ¶0090.) “The audio processing module 606 is configured to process audio content associated with the multimedia content” while the “video processing module 608 is configured to process the video portion(s) of multimedia content[.]” (Fontana ¶¶0091, 0094.) Fontana discloses that the audio and video may be separately processed by “segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.)[.]” (Fontana ¶0135.)

Fontana further teaches that its audio processing can be run on “**audio files**” **generated** from the video. (Houh ¶¶82-83.) Specifically, Fontana teaches that “an audio separation operation 1128 strips, or extracts, the audio from the multimedia content.” (Fontana ¶0164.) Stripping or extracting the audio from the multimedia

content generates an audio file. (Houh, ¶83.) Fontana additionally teaches that “a number of sources provide speech to text conversion program.” (Fontana ¶0169.) One such program disclosed in Fontana is “large vocabulary continuous speech recognition (LVCSR) engines” that “depend on a language model that includes a vocabulary/dictionary for speech-to-text conversion of audio files.” (Fontana ¶0169.) A POSA would thus have appreciated that Fontana’s disclosures encompass “**generating audio files ... from the video.**” (Houh, ¶83.)

Fontana additionally teaches “a thumbnail extraction module” that “is arranged to generate thumbnails In some embodiments, the thumbnail extraction module 618 generates a series of thumbnails representing scenes throughout the multimedia content.” (Fontana ¶0095.) A thumbnail is an image. (Houh, ¶84.) Such images were commonly stored as image files in 2013, such as in gif, jpg, or tiff formats, for example, which are common formats for graphical image files. (*Id.*) A POSA would thus have appreciated that by extracting thumbnail images from the video, Fontana’s disclosures satisfy “**generating ... image files from the video.**” (*Id.*)

Rationale and Motivation to Combine (Kritt with Fontana): A POSA would have found it obvious to combine Kritt with Fontana. (Houh ¶¶85-88.) Kritt and Fontana are analogous references to the ’954 patent. Like the ’954 patent, both Kritt and Fontana are addressed to the field of processing and generating multimedia

content, including video, and both specifically address how to search within such multimedia content. ('954, 1:36-38 (“The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.”); '954, p.1 (Abstract); Kritt, ¶0006 (“One embodiment is directed to a method that visually summarizes the types and locations of particular types of content in a video in a way that facilitates understanding by a viewer.”); Fontana ¶0001 (“[T]he present disclosure relates to systems and methods for processing and delivery of multimedia content.”); Fontana ¶0041.) Kritt and Fontana both teach the use of audio and video processing modules to analyze the audio and video, respectively, of a video file. (Kritt, ¶¶0022-0024 (describing “the video processing module 152, audio processing module 154”); Fontana ¶0091 (“The audio processing module 606 is configured to process audio content associated with the multimedia content.”); Fontana ¶0094 (“The video processing module 608 is configured to process the video portion(s) of multimedia content”).)

A POSA would have appreciated that Kritt’s method of generating video data from a video could incorporate aspects of Fontana’s audio and video processing. (Houh ¶¶86-88.) In particular, Kritt teaches that “a key frame may be determined for a shot file.” (Kritt, ¶0041.) Kritt explains that “[o]ther known methods for determining a key frame may be employed.” (Kritt, ¶0041.) By contrast, Fontana discloses that a thumbnail can be generated to correspond to the location of an object

of interest; this would make the thumbnail a better choice of a “key frame” to deem “representative of all frames in the shot” because it would include the relevant object of interest. (Houh ¶86.) A POSA would therefore have been motivated to combine Kritt with Fontana, to use Fontana’s disclosures of “generat[ing] thumbnails at possible locations the content provider would like to create an object of interest,” to identify the key frame based on likely locations of objects of interest and take advantage of Fontana’s benefits in object identification within thumbnails. (See Fontana ¶0095; *see also, e.g.*, Fontana ¶¶0139, 0141-0142, 0148 (describing identification of thumbnails corresponding to objects of interest); Houh ¶86.)

Likewise, Kritt teaches that “visual objects in a shot may be identified by application of one or more known image recognition processes to the shot.” (Kritt, ¶0043, ¶0045.) But Kritt provides few details on the substance of these known methods for object identification. Fontana provides multiple approaches to object identification. (See Fontana ¶0139 (describing known “object recognition programs, including computer vision programs”); *see also, e.g.*, Fontana ¶¶0084, 0087-0089, 0094-0095, 0139-0145, Figs. 9-10 (describing identification of objects of interest).) A POSA would also have been motivated to combine Kritt with Fontana, to use Fontana’s disclosures around object identification within thumbnail images to take advantage of Fontana’s improvements in object identification. And a POSA would have had a reasonable expectation of success in the combination given Kritt’s

disclosure that “[o]ther known methods for determining a key frame may be employed,” and the well-known nature of the computer vision tools disclosed by Fontana. (*See e.g.*, Kritt, ¶0041; Fontana ¶0139; Houh ¶87.)

As to audio, Kritt discloses identifying “key words” and “an audio transcript” from the audio files but does not explain how to perform speech-to-text recognition. (*See, e.g.*, Kritt, ¶¶0016-0017, 0065.) Instead, Kritt discloses that this transcript may be created “using any known technique.” (Kritt, ¶0065.) Fontana provides additional details on well-known algorithms to perform speech-to-text recognition as part of audio processing. (*See* Fontana ¶¶0169-0171 (describing known speech-to-text conversion programs); *see also* Fontana Figs. 11-12, ¶¶0121, 00164-0174 (describing speech-to-text algorithms).) Therefore, a POSA would also have been motivated to combine Fontana’s disclosures on performing audio processing, including speech-to-text recognition, to implement Kritt. (Houh ¶88.) As above, a POSA would have had a reasonable expectation of success in the combination in light of the analogous nature of the references, Kritt’s instruction to use “any known technique” for speech recognition, and the well-known nature of the speech-to-text recognition algorithms disclosed in Fontana. (*See, e.g.*, Kritt, ¶0065; Fontana ¶¶0169-0171; Houh ¶88.)

(b) **“distributing the audio files and the image files across a plurality of processors and processing the audio files and the image files in parallel;” (Claim 1[b])**

Claim 1[b] is disclosed by and obvious over Kritt in view of Fontana. (Houh, ¶¶89-97.)

As discussed above, Kritt describes using separate “video processing” and “audio processing” modules, which “may be on different computer systems.” (Kritt, ¶0022.) Kritt’s “video processing module 152 may include various processes that generate visual tags... The audio processing module 154 may include various processes for generating audio and key word tags.” (Kritt, ¶0024.) Each of the “video processing module” and “audio processing module” comprises a processor(s), such that Kritt’s distribution of the audio files and image files for processing by the video processing module and audio processing module **distribut[es] the audio files and the image files across a plurality of processors.** (Houh ¶90.) This is confirmed by Kritt’s disclosure that its embodiments can be provided using “cloud computing,” such as “virtual machines hosted by a cloud service provider.” (See Kritt, ¶¶0078-0079 (describing cloud computing implementation).) A POSA would have known that a cloud computing infrastructure with “a shared pool of configurable computing resources” and “virtual machines hosted by a cloud service provider,” such as that taught by Kritt, distributes files for processing across **a plurality of processors.** (See *id.*; Houh ¶91.)

Kritt further acknowledges that its disclosed processes may be “executed substantially concurrently,” which a POSA would have recognized as **parallel processing** of those functions. (Kritt ¶0080; Houh ¶92.) A POSA would further have recognized that Kritt’s video processing module and audio processing module—blocks 152 and 154, respectively, in Kritt’s Figure 1, shown in claim 1[pre] above—were two such blocks that could be executed in parallel, such that the audio processing and video processing occur concurrently. This results in **processing the audio files and the image files in parallel**. (Houh ¶92.)

It would further have been obvious to a POSA to combine Kritt with Fontana to further **distribute the audio files and the image files across a plurality of processors and process the audio files and the image files in parallel**. (Houh ¶¶93-97.) Whereas Kritt discloses at a high level the use of a plurality of processors and parallel processing, Fontana provides further implementation information on how to carry out video and audio processing in parallel within a distributed computing environment, including the cloud-based environment disclosed by Kritt. A POSA would therefore have been motivated by Kritt’s disclosures to combine with Fontana in order to implement parallel processing of audio and video files, which as Fontana explains “allow[s] multimedia content to be processed efficiently.” (Fontana ¶0055; Houh ¶93.)

Fontana implements its method using “a plurality of distributed computing

systems[.]” (Fontana ¶0045.) A distributed computer system is a form of “parallel” processing. (Houh, ¶94.) Fontana’s distributed computing system allows tasks, including “video, image or audio algorithms[.]” to be “separated and performed across multiple computing systems in parallel.” (Fontana ¶0053; *see also id.* ¶0090.) Fontana’s distributed computing system includes both parallel processing of image files separate from audio files, and processing multiple image files in parallel. (Houh, ¶94.) Therefore, Fontana discloses that its distributed computing system **“distribut[es] the image files across a plurality of processors and process[es] the image files in parallel.”**

For example, Fontana teaches a “workflow server” that “distributes one or more portions of jobs associated with each data processing request to the integration framework 204 and the storage network 206.” (Fontana ¶0050.) The result “allow[s] computationally intensive jobs (e.g., video and audio content processing) to be distributed across a number of computing systems.” (*Id.*) Thus, the workflow server divides computationally intensive jobs, such as video and image processing, into portions and distributes those portions across multiple computing systems. The result of this system is that these processing tasks, such as image processing, would be divided up and performed in parallel. (Houh, ¶¶95-96.)

Fontana teaches that its method is implemented using “a plurality of computing systems, illustrated as servers 302a-c.” (Fontana ¶0052.) These servers

“are interfaced to inbound work ... for coordination and communication of data for processing.” (*Id.*) Fontana further teaches that “one or more of the servers 302a-c can include specific graphical processing units for processing lower level video, image or audio algorithms” and are “configured to share processing jobs, such that tasks can be ... separated and performed across multiple computing systems in parallel.” (Fontana ¶0053.) As a result, “any of those computing systems can perform all **or a portion** of a processing job.” (Fontana ¶0055.) A POSA would have appreciated that the described servers that share processing jobs carry out parallel processing. (Houh, ¶97.) In the context of image processing, it would have been obvious to a POSA that image processing within Fontana’s distributed computing system could **process the image files in parallel** by, for example, distributing the image files for processing across multiple servers. (*Id.*)

**(c) “converting audio files associated with the video to text;”
(Claim 1[c])**

Kritt discloses claim 1[c]. (Houh ¶98.) Kritt determines “an audio transcript” that “may include all of the words spoken in the video” using “any known technique.” (Kritt ¶0065.) Kritt’s audio transcript of the words spoken in the video **convert[s] audio files associated with the video to text.**

(d) “identifying an object in the image files;” (Claim 1[d])

Kritt discloses “**identifying an object in the image files.**” (Houh ¶¶99-102.)

As discussed for claim 1[a], Kritt teaches that “key frames” for a shot (i.e.,

image files) “may be deemed to be representative of all frames in the shot, permitting descriptive data for the shot to be determined only for the key frame and not for every frame of the shot.” (Kritt, ¶0041.) Kritt thus teaches that its analysis of shots may be performed on the key frames. (Houh ¶100.)

Kritt further discloses identifying “visual objects,” such as faces or facial expressions, in the shots and frames (i.e., **“in the image files”**). For example, Kritt discloses that “visual objects in a shot may be identified and tagged,” “a human visual object may be identified by identifying a face (“human facial object”) in a frame,” and “an extracted human facial object may be further processed to determine the identity of the viewer or to determine a facial expression of the viewer.” (Kritt, ¶0043.) Kritt further discloses identifying the human objects (e.g., faces) with particular individuals: “a scene may include a first shot that is a long shot of a particular person and a second shot that is a medium shot of the same person” or “a sequence of four consecutive shots in which the first and third shots have a first human object and the second and fourth shots have a second human object.” (Kritt, ¶0051.) A human object (a face, or its associated person) is **an object in the image files**, such that Kritt discloses **identifying an object in the image files**. (Houh ¶101.)

Kritt further discloses processing the “human facial object ... to determine a facial expression of the viewer.” (Kritt, ¶0043.) Kritt alternately describes this as

determining an “emotion”: “a human facial object may be processed to determine the emotion expressed on the facial object.” (Kritt, ¶0047.) This emotion may be determined using “facial features and their orientation, and a support vector machine to determine an emotion corresponding with detected facial features.” (Kritt, ¶0047; *see also* Kritt, ¶0045 (describing facial identification methods that “may describe the face as a whole or the facial features separately”).) Then, “a sequence of frames in which a facial expression morphs from one emotion to another may be analyzed to determine an emotional category of a human facial object.” (Kritt, ¶0047.) Furthermore, Kritt discloses that a tag may “identif[y] a human object displaying a particular emotion.” (Kritt, ¶0034.) A POSA would appreciate that this tag corresponds to a facial expression. (Houh ¶102.) The determined facial expression, and the facial features that make up that facial expression, are likewise **objects identified in the image files.**

**(e) “determining a contextual topic from the image files;”
(Claim 1[e])**

Kritt discloses “**determining a contextual topic from the image files**” in several ways: (1) identifying a shot or a scene as a sequence of semantically related images that is determined by evaluating the images; and (2) identifying an emotion for a shot or scene by evaluating facial expressions across a sequence of image files. (Houh ¶¶103-110.)

Scenes/Shots: Kritt discloses **determining a contextual topic** in the form of

a “scene” or a “shot.” Shots and scenes are each a sequence of related images, and they are determined by evaluating the **image files** (frames).

A shot “may be a continuous sequence of frames captured without interruption by a camera oriented in a single direction or camera angle.” (Kritt, ¶0039.) Kritt explains that the “video file 204 may be parsed into shot files according to any known method,” and provides as an example determining a shot by comparing histograms for consecutive frames of the video: “[A] histogram may be computed for each frame of the video file and the histograms for consecutive frames compared. If the histogram intersection of first and second consecutive frames is greater than a threshold, it may be inferred that the frames are similar, and consequently that the two frames are part of the same shot.” (Kritt ¶0040.) Based on this comparison, Kritt “pars[es] the video file so that sequential frames between determined shot boundaries are grouped togetherI [sic] or otherwise identified or tagged as being associated with a particular shot.” (*Id.*) “Sequential frames associated with a particular shot” are then “a shot file.” (*Id.*) By determining which frames belong to a given shot, and tagging those frames as associated with a particular shot, Kritt determines a **contextual topic** (the shot) **from the image files** (the analyzed frames). (Houh ¶105.)

Similarly, Kritt also discloses identifying scenes, which are “semantic segment[s]” comprised of two or more frames (and one or more shots) and which

may be “intended to convey common emotional content, that is intended to convey an element of a story, that is accompanied by a common audio segment, or some combination of the foregoing” (Kritt, ¶0050.) “A visual scene may include one or more camera shots and one or more human and nonhuman objects.” (Kritt, ¶0051.) Kritt’s determination of a scene is a **contextual topic**. (Houh ¶¶106-107.)

In particular, Kritt’s disclosure that a scene may “convey an element of a story” corresponds to the example topic of an “event” described in the specification of the ’954 patent. (’954, 4:44-47 (“[A] plurality of instances of the symbol can be compared to a topic database to identify a topic (such as an event). For example, the result may identify the dog (symbol) as running or jumping.”).) Kritt explains that “visual scenes may be determined if a preceding and following shot include related visual objects. For example, the first shot may include a particular person, the second shot may include another person, and two may be interacting.” (Kritt, ¶0052.) This determination in Kritt is performed by analyzing frames—which are **image files**—and/or shots—which are a sequence of frames, so also **image files**. (See Kritt, ¶¶0050-0052.)

Emotions: Additionally, Kritt discloses **determining a contextual topic** in the form of “an emotion, mood, or theme for one or more of the scenes.” (Kritt, ¶0017.) An “emotion, mood, or theme” is a “contextual topic” because it describes the context (in this case, the general feeling) associated with the shot or scene. (Houh

¶¶108-110.)

Furthermore, the '954 patent explains a “topic” with an example of identifying an event or an action performed by an object—such as a dog running and jumping: “[A] plurality of instances of the symbol can be compared to a topic database to identify a topic (such as an event). For example, the result may identify the dog (symbol) as running or jumping.” ('954, 4:44-47.)

This is analogous to Kritt. Kritt teaches that it identifies “an emotional category of a human facial object” to assign an “emotion tag” by analyzing the actions performed by the face—or as Kritt explains, “by comparing motion vectors” for the facial expression across “a sequence of frames” (ie, image files) in which “a facial expression morphs from one emotion to another.” (Kritt ¶0047.) “[M]otions that may be determined may include amusement, joy, anger, disgust, embarrassment, fear, sadness, surprise, and a neutral state.” (*Id.*) Kritt thus teaches identifying the actions (motions) performed by a face from a series of frames (image files) to determine an emotional category or mood (eg, “amusement, joy, anger, disgust, embarrassment, fear, sadness, surprise”), from which an emotion tag is generated. This corresponds to the '954 patent’s **determining a contextual topic from the image files**. (Houh ¶¶109-110.)

(f) **“assigning a probability of accuracy to the identified object based on the contextual topic;” (Claim 1[f])**

Kritt discloses and renders obvious claim 1[f]. As discussed for claim 1[d],

the identified objects include human objects (faces and/or people) as well as facial features and facial expressions. Kritt **assigns a probability of accuracy** to these **identified objects based on a contextual topic**, which, as explained for claim 1[e], include shots, scenes, and emotions. (Houh ¶¶111-117.)

Human objects (faces/people): Kritt assigns a probability that a human object (“**identified object**”) was correctly identified by using other information about the shot or scene (“**the contextual topic**”) in which the human object was identified. The probability is therefore assigned **based on the contextual topic**. (Houh ¶¶112-113.)

Specifically, Kritt discloses “a visual tag” associated with a shot or scene to “indicate that a particular human object appears in a shot.” (Kritt, ¶0033.) Kritt teaches that this “visual tag” can be compared with “one or more other tags associated with the same shot or scene” in order “to validate that the human object was correctly identified.” (Kritt, ¶0033.) Kritt discloses that this validation process may involve assessing “if a **probability or confidence parameter** associated with the particular tag is above a threshold,” in which case “it may be determined that the object was correctly identified.” (Kritt, ¶0033.) (Kritt, ¶0033.) A POSA would have appreciated that assessing “if a probability or confidence parameter associated with the particular tag is above a threshold” requires **assigning a probability of accuracy to the identified object**. (Houh ¶113.) In the context of Kritt, this

probability is assessed based on the above-described comparison of the visual tag for the human object to other tags associated with the same shot or scene—which are **contextual topics**. Thus, the probability of accuracy is assigned **based on the contextual topic**.

Facial expressions: Kritt describes a “confidence level” that a perceived emotion (a **contextual topic**) “accurately represents a ground truth emotion.” (Kritt, ¶0034.) (Houh ¶¶114-116.)

As described in claim 1[e], Kritt uses the identified facial expressions (*see* claim 1[d]) to detect emotions and generate an emotion tag associated with a scene. (Kritt ¶0047 (“a human facial object may be processed to determine the emotion expressed on the facial object..... the operation 510 may include generating an emotion tag that is associated with the scene of the video in which the facial emotion was detected.”).) Kritt then discloses assigning to that emotion “a confidence level that the perceived emotion accurately represents a ground truth emotion.” (Kritt, ¶0034.) A POSA would have appreciated that a confidence level corresponds to a **probability of accuracy**. (Houh ¶116.) A POSA further would have found it obvious that, because Kritt identifies the emotion based on the facial expression, a probability of accuracy could be assigned to the underlying facial expression based on the probability of accuracy (ie, confidence level) assigned to the emotion (the **contextual topic**). (*Id.*)

Thus, a POSA would have found it obvious in light of Kritt to **assign a probability of accuracy** to the facial expression (**object**) based on the emotion (**contextual topic**).

- (g) “**converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic;**” (Claim 1[g])

Kritt discloses claim 1[g]. Kritt discloses **converting the image files associated with video** (i.e., the frames and shots) **to video data** in the form of tags, tag files, and other metadata about a video, through use of video and image analysis. These tags and tag files **comprise objects** including human objects, faces, and facial expression information; **the probability** of accuracy or confidence levels; and **the contextual topic** including the shot/scene information and any emotion tag associated with the shot/scene. (Houh ¶¶118-124.)

Beginning with **the object** (*see* claim 1[d]), Kritt discloses generating a “visual tag” that “corresponds with visual objects, such as human and nonhuman objects.” (Kritt, ¶0055.) These “visual tags may be associated or set for each scene.” (Kritt, ¶0055.) This process of generating visual tags associated with scenes is depicted in Figure 5:

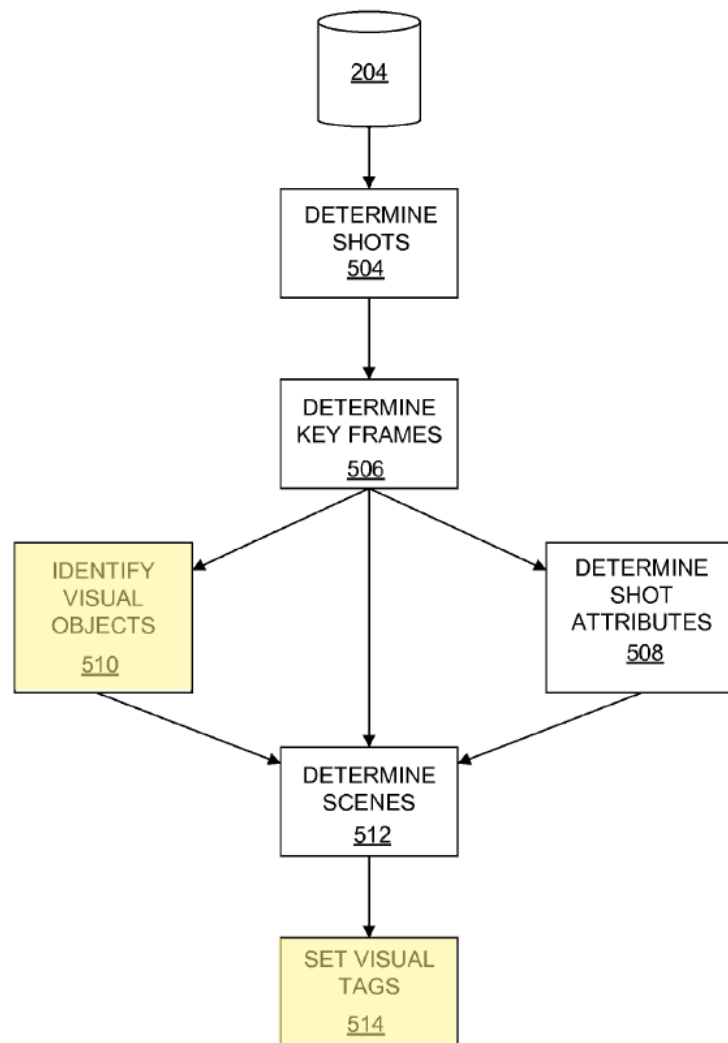


FIG. 5

(Kritt, Fig. 5 (annotated).) These visual tags could include facial expressions. Kritt explains that a tag may “identif[y] a human object displaying a particular emotion.” (Kritt, ¶0034.) A POSA would have appreciated that a tag for “a human object displaying a particular emotion” in the context of Kritt is a tag for a facial expression,

because Kritt teaches determining emotions by analyzing facial expressions. (Houh ¶120.) This would be an obvious implementation given Kritt’s disclosure of analyzing changes in facial features/facial expression across a series of frames. (See Kritt, ¶0047.)

As for the **contextual topic**, both the contextual topics of shots/scenes and emotions are stored as **video data** in the form of tags. Kritt teaches that determining a shot can result in sequential frames being “tagged as being associated with a particular shot.” (Kritt, ¶0040.) This is illustrated in Figure 5:

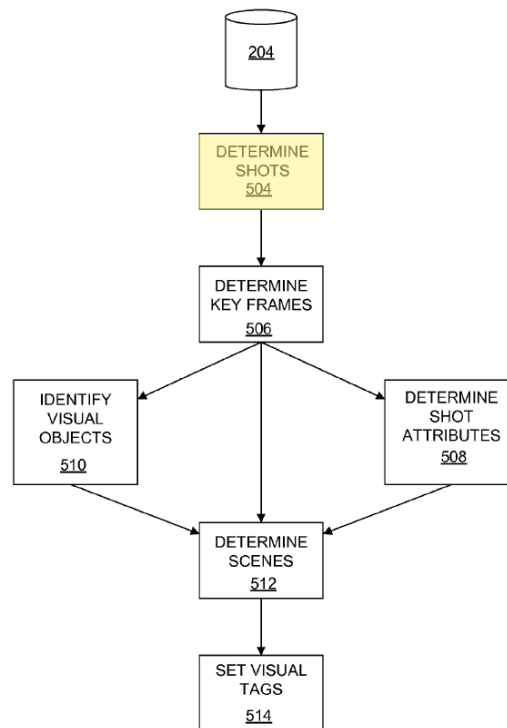


FIG. 5

(Kritt, Fig. 5 (annotated).) A POSA would have found it obvious that the frames or

shots belonging to a determined scene could likewise be tagged, in operation 512 of the same Figure 5. (Houh ¶121.) This is consistent with Kritt’s use of tags to identify a frame as belonging to a particular shot (Kritt, ¶0040) and Kritt’s use of tags in connection with scenes: “One or more of the various tags described herein may be associated with a particular scene or semantic segment if the particular tag is determined from content in the scene.” (Kritt, ¶0050.)

Kritt additionally teaches generating an “emotion tag file 314” that “includes tags that are associated with emotion objects,” which are a **contextual topic**. (Kritt, ¶0034.) In particular, “a tag may be generated for an emotion object,” where an “emotion object may be associated with an emotion, mood, or theme that a typical viewer might be expected to perceive.” (*Id.*; Houh ¶122.)

Kritt likewise discloses including the **probability** as part of the video data of tags. As discussed in claim 1[f], Kritt discloses or renders obvious assigning a probability of accuracy to the human objects (e.g., faces) and facial expressions (determined via emotions). Kritt teaches use of “a probability or confidence parameter associated with the particular tag” for visual objects, including human objects (like faces or people), to determine whether the object was correctly identified. (Kritt ¶0033.) Kritt further discloses “adding an indication to the tag that it may be relied on to a particular extent.” (*Id.*) A POSA would have appreciated that the confidence level (**probability**) associated with a human object is one such

indication of the extent to which an object identification can be relied upon, and that the probability could accordingly be added to the tag. (Houh ¶123.)

As to facial expressions and emotions, Kritt teaches that the emotion tag file (**video data**) can include “a confidence level that the perceived emotion accurately represents a ground truth emotion.” (Kritt, ¶0034.) As discussed for claim 1[f], it would have been obvious to a POSA that the confidence level (that is, **probability**) assigned to an emotion could likewise be assigned to the facial expression/facial features used to determine that emotion. (Houh ¶124.)

(h) “cross-referencing the text and the video data with the video to determine contextual topics;” (Claim 1[h])

Kritt discloses claim 1[h]. (Houh ¶¶125-132.) In Kritt, **contextual topics** (emotions/moods and shots/scenes) are determined by **cross-referencing the text** (audio transcript and key words) **and the video data** (visual tags) **with the video**.

Emotions: As discussed for claim 1[e], Kritt **determines a contextual topic** in the form of an emotion, mood, or theme for a shot or scene by evaluating facial expressions across a sequence of image files. Kritt further explains that the image files are not the only input that factors into determining an emotion. Rather, Kritt teaches that objects, key words (which are derived from the **text**, *see* claim 1[a]), and other data from the **video** are together used to determine the emotion: “the objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.”

(Kritt, ¶0017.) The objects include visual objects, which are **video data** (*see* claim 1[g]). (Kritt, ¶0043.) The “key words may be determined from the audio transcript” (**the text**). (Kritt, ¶0066.) And, as discussed further below, the other data includes information such as metadata files, time stamps, and scene information that is derived from **the video**. For example, Kritt teaches that “[e]ach tag may be associated with a time stamp that indicates the start and stop time in which the object or attribute is rendered or otherwise associated.” (Kritt, ¶0017; *see also* Kritt 0032 (visual tag file includes tags for “visually perceivable objects” and key word tag file includes tags for “key word objects”).) As noted below, these time stamps are used as part of determining an emotion.

Kritt teaches how each of the visual objects (**video data**), key words from the text (**text**), and other information from the video (e.g., metadata files, time stamps) are **cross-referenced** to determine an emotion:

In operation 312, an emotion tag file 314 may be created from the attribute tag file 308 and the consistency-corrected visual tag 302, audio tag 304, key word tag 306, and metadata 210 files. The emotion tag file 314 includes tags that are associated with emotion objects. In one embodiment, an emotion object may be associated with an emotion, mood, or theme that a typical viewer might be expected to perceive or that the creators of a video intended the audience to perceive. Each emotion object may be of a predefined type and associated with a time stamp. ... Moreover, an emotion object may be generated in

operation 312 by identifying patterns of visual, audio, key word, and attribute tags that correspond or correlate with an emotion object.

(Kritt, ¶0034.) By compiling “the attribute tag file 308 and the consistency-corrected visual tag 302, audio tag 304, key word tag 306, and metadata 210 files” to determine emotion tag files, Kritt cross-references information from the text (e.g., as summarized in the audio tag and key word tag), the video data (the visual tags corresponding to video objects), and the video (such as the attribute tag, metadata files, and time stamps). (Houh ¶¶127-128.)

Kritt further explains that this determination of an emotion may occur within the context of one or more scenes—that scene information is an additional cross-referencing to information about the video:

According to various embodiments, a viewer selects a video, and human, nonhuman, and audio objects of the video are identified. In addition, key words that are spoken by human objects in the video are identified. ... The objects and key words are then associated with the scenes of the video. Further, the objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.

(Kritt, ¶0017; Houh ¶¶129-130.)

Kritt therefore discloses **cross-referencing the text** (key words) **and the video data** (visual tags for, eg, faces and other human objects) **with the video**

(metadata, time stamps, and scenes) **to determine contextual topics** (emotions).

Scenes: Kritt also **determines a contextual topic** in the form of a shot or scene, as discussed for claim 1[e]. As with emotions, Kritt discloses that the scenes are determined by **cross-referencing the text and the video data with the video**. The determination of scenes derives from cross-referencing multiple types of information about the video. Kritt explains that a scene may be “a segment of two or more video frames that includes common objects, that is intended to convey common emotional content, that is intended to convey an element of a story, that is accompanied by a common audio segment, or some combination of the foregoing.” (Kritt ¶0050.) Likewise, “scenes may be determined by grouping together consecutive shots having visual or audio objects corresponding with the same ground truth.” (Kritt ¶0051.) As before, the objects include visual objects, which are **video data** (see claim 1[g]). (See, e.g., Kritt, ¶0043.) The “common audio segment” and “audio objects” include **the text** (see claim 1[c]), such as spoken words identified in an audio transcript. (Houh ¶131; see also Kritt, ¶¶0016-0017 (“Human, nonhuman, and audio objects may be used to classify a particular segment of a video as a scene. The objects and key words are then associated with the scenes of the video.”).) Finally, the **video** frames “intended to convey common emotional content” and “intended to convey an element of a story” are from the **video**, as are the determined emotions. (See Kritt, ¶0050; see also claim 1[e] (describing

determining an emotion); Houh ¶131.)

Kritt thus teaches **cross-referencing the text** (audio objects, such as spoken words) **and the video data** (visual tags for, e.g., faces and other human or nonhuman objects) **with the video** (including video frames, emotions) **to determine contextual topics** (scenes).

- (i) **“generating a contextual text, an image, or an animation based on the determined contextual topics;” (Claim 1[i])**

Kritt discloses claim 1[i] with regard to each of the contextual topics (scenes, and emotions). (Houh ¶¶133-137.)

Emotions: Kritt teaches **generating a contextual text, an image, or an animation based on the determined contextual topics** of emotions in the form of “an emotion or mood denoting icon” such as “an ‘emoticon.’” (Kritt, ¶0038.) These emotion denoting icons are “representative of an emotion tag.” (Kritt, ¶0038.) As explained for claim 1[e], the emotion tag corresponds to a **contextual topic**, and as explained in claim 1[h], may be further based on information about the scene. Thus, the emotion denoting icon is **based on the determined contextual topics**.

The emotion denoting icon is **an image, or an animation**. For example, “an emotion or mood denoting icon 416 may be an ‘emoticon’,” or alternately “a colored or gray-scale icon.” (Kritt, ¶0038.) Exemplary emotion denoting icons are shown in Figure 4:

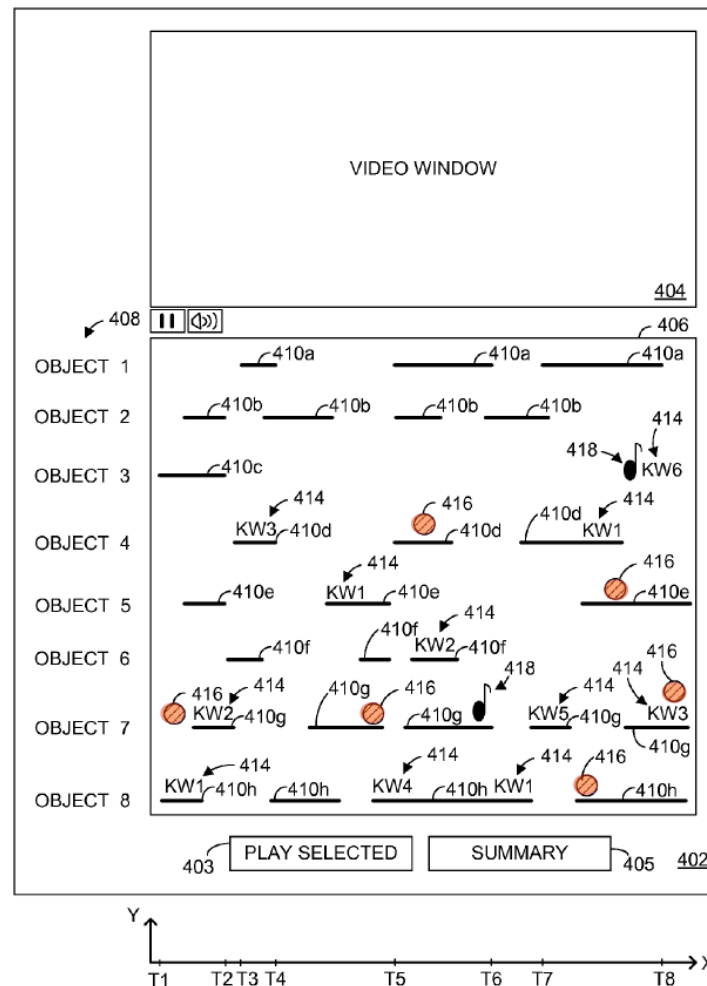


FIG. 4

(Kritt, Fig. 4 (annotated).) Additionally, the emotion denoting icon is **contextual** because it describes the content of the scene it is describing. (Houh ¶135.)

Kritt additionally teaches **generating a contextual text, an image, or an animation based on the determined contextual topics** of scenes in the form of “visual information summarizing all of the scenes of a video,” such as “visual representations of key words, and human, nonhuman, audio, and emotion objects.” (Kritt, ¶0016.) These visual summaries “describe the scenes of a video to a viewer”

and enable the viewer “to generally understand the scene.” (Kritt, ¶0016; *see also* Kritt, ¶0017.)

The visual information is **contextual**, because it “describe[s] the scenes of a video.” (Kritt, ¶0016; Houh ¶¶136-137.) Furthermore, “visual information” includes **text, an image, or an animation**. (See Kritt, ¶0035 (“A variety of text, symbols, lines, and icons (‘indicia’) for summarizing the video may be rendered”).) And the visual information summary is **based on the determined contextual topics** because it describes a scene, which is a contextual topic. (See claim 1[e].) Additionally, emotion denoting icons are one way of summarizing the scenes, which are additionally based on the contextual topic of emotion. (See, e.g., Kritt, ¶0038 (“display of emotion denoting icons 416 facilitates a quick understanding by a viewer of the types and locations of various emotions, moods, or themes in the video”).)

(j) **“generating a content-rich video based on the generated text, image, or animation.” (Claim 1[j])**

Kritt discloses claim 1[j]. (Houh ¶¶138-139.)

Kritt teaches **generating a content-rich video** by displaying the video alongside images and text generated to summarize the scenes, including the emotion denoting icons. Kritt teaches “a display screen 402” that displays a “user interface” with “windows 404 and 406.” (Kritt ¶0035.) “The video may be played in the window 404” while “[a] variety of text, symbols, lines, and icons (‘indicia’) for

summarizing the video may be rendered in the window 406.” (*Id.*) The “indicia” in window 406 also include “emotion, mood, or theme denoting icons 416.” (Kritt ¶0038.) This display of the generated **content-rich video** is depicted in Figure 4:

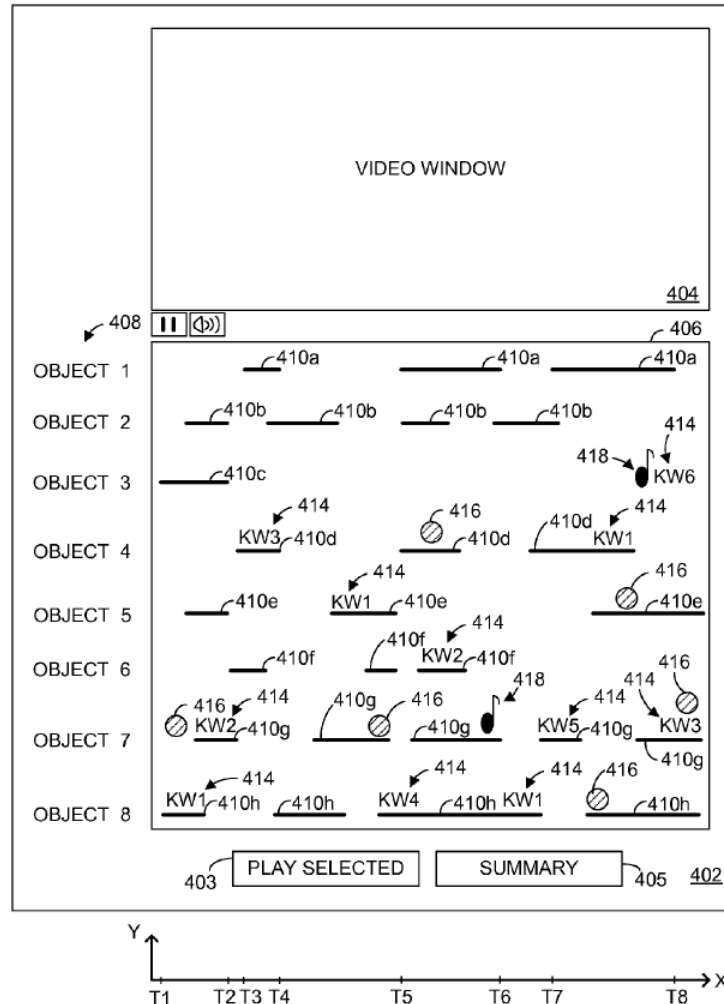


FIG. 4

(Kritt, Fig. 4.) This two-window display of the content-rich video depicts the **contextual text, image, or animation** generated in claim 1[i]—namely, the visual information summary, and the emotion denoting icons. (Houh ¶¶138-139.) This aligns with the '954 patent specification. The '954 patent gives as one example of

content-rich video that an advertisement may be generated based on the video content, and the advertisement “can be placed at a specific time in the video” and “can also be placed at a specific time as part of the video player, e.g., side panel, and also may be placed on a second screen.” (’954, 2:5-10.) So too here, the content-rich video is generated by placing the generated contextual text, image, or animation in a side panel at a specific time in the video player. (Houh ¶139.)

2. Claim 2: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”

The additional limitations of claim 2 are obvious over the combination of Kritt with Fontana. (Houh ¶¶140-144.)

Kritt describes determining “key words” from “the audio transcript” (“**the text**”), including key words that provide “significant information content about a scene in a video” (“**context**”), such as “a concept or idea that is central to a plot or story.” (Kritt, ¶0066.) Key words can be determined “by counting the frequency of occurrences of words, the most frequently occurring verbs and nouns being determined to be key words.” (Kritt, ¶0066.) A POSA would have recognized this as a rudimentary form of **natural language processing**. (Houh ¶142.) Such natural language processing techniques were well-known by the priority date of the ’954 patent, and a POSA would have found it obvious to build on Kritt’s disclosure by implementing one, such as term frequency/inverse document frequency (TF-IDF),

which is a form of natural language processing that identifies salient words by multiplying term frequency by inverse document frequency. (*See, e.g.*, EX1008, Downs, 3:53–4:49 (describing TF-IDF); EX1009, Lord 12:4-15 (listing term frequency-inverse document frequency among known techniques for natural language processing); Houh ¶142.) Because a “key word may be a name of an actor that appears in a scene,” the key words determined by Kritt are a **context associated with the video**. (*See* Kritt, ¶0066.)

Furthermore, as discussed for claim 1[a], it would have been obvious to combine Kritt with the audio processing methods described in Fontana. (Houh ¶143.) One such audio processing technique discussed in Fontana is **natural language processing**. (Fontana ¶¶0184, 0100 (describing searching within a transcript “based on natural language processing”).) A POSA combining Kritt with the audio processing techniques of Fontana thus would have found it obvious to additionally implement natural language processing, such as to identify key words. (Houh ¶143.)

Fontana determines the topic of the multimedia content—the **context associated with the video**—using natural language processing. Specifically, Fontana teaches that search queries, including those leveraging the above natural language processing, can be used **to determine context associated with the video** in the form of keywords or relevant portions of the multimedia content: “A content

request operation 814 receives a request related to multimedia content... such as a search query related to keywords appearing in one or more fields of metadata associated with the content (e.g., titles, authors, producers, genre, etc.) or in the transcript or other text associated with one or more pieces of content.” (Fontana ¶0126.) Additionally, “[t]he provide metadata operation 816 selects at least a portion of the metadata associated with the content (e.g., including ... transcript information,...) for inclusion with the content during playback.” (Fontana ¶0127.) A POSA would have appreciated that identifying metadata, portions of the transcript, and portions of the multimedia data related to keywords is **determining context associated with the video**, and could be combined with Kritt to determine key words using **natural language processing**. (Houh ¶144.)

3. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to extract a topic.”

The additional limitations of claim 3 are obvious over the combination of Kritt with Fontana for the same reasons as claim 2. A POSA would have appreciated that Kritt’s key words, and Fontana’s keywords, are **topics**. (Houh ¶145.)

1. Claim 4: “The method according to claim 1, further comprising: processing the image files to extract additional text.”

The additional limitations of claim 4 are disclosed by Kritt. (Houh ¶¶146-147.) Kritt explains that “another type of nonhuman object that may be shown in a

video frame is text.” (Kritt, ¶0004.) Kritt teaches that these “nonhuman...objects of the video are identified.” (Kritt, ¶0017.) Kritt further teaches **extracting** the identified nonhuman objects, such as text: “Both human and nonhuman visual objects may be identified in operation 510. ... The operation 510 may include extracting an identified object for further processing.” (Kritt, ¶0043.)

2. Claim 5: “The method according to claim 4, wherein the additional text is generated by segmenting the image files before processing the image files in parallel.”

The additional limitations of claim 5 are disclosed and rendered obvious by the combination of Kritt with Fontana. (Houh ¶¶148-154.)

Kritt discloses that visual objects (including the additional text, *see* claim 4) may be identified and extracted (**generated**) from “a shot” by “operation 510.” (Kritt, ¶0043 (“In operation 510, visual objects in a shot may be identified and tagged.”).) Kritt illustrates operation 510’s placement in the logical flow in Figure 5, showing operation 510 (“identify visual objects”) after “determine shots” (operation 504):

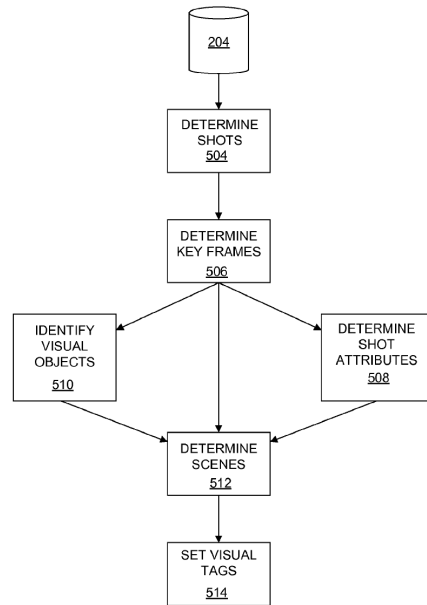


FIG. 5

(Kritt, Fig. 5.)

As explained in claim 1[a], the process of generating the shot files **segments the image files** for the video. This segmentation to determine shots takes place **before processing the image files**, as shown in Figure 5 above. And as discussed for claim 1, a POSA would have found it obvious over the combination of Kritt with Fontana to process the image files **in parallel**. (Houh ¶¶149-150.)

The combination with Fontana for image processing also yields additional means by which Kritt with Fontana **segments the image files before processing the image files in parallel**. This segmentation would occur before processing the image files, so before Kritt identifies and extracts the **additional text** as discussed for claim 4. (Houh ¶¶151-154.)

Fontana teaches segmenting the image files in several ways. As described for claim 1[b], Fontana teaches a distributed computing system that “allow[s] for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing and resulting usability for content providers.” (Fontana ¶0135.) Fontana’s workflow server “distributes one or more portions of jobs associated with each data processing request.” (Fontana ¶0050.) The result “allow[s] computationally intensive jobs (e.g., video and audio content processing) to be distributed across a number of computing systems.” (*Id.*) This distributed computing system **segments** tasks including image processing, prior to processing the images in parallel. (Houh ¶152.)

Fontana also teaches **segmenting image files** into “a series of thumbnails representing scenes throughout the multimedia content” prior to performing the object recognition. (Fontana ¶0095.) “In certain embodiments, the candidate object generation operation 906 splits the multimedia content into a plurality of sections, and generates a thumbnail image associated with each of those sections for preview by the content provider.” (Fontana ¶0139.) Fontana also discloses using “computer vision programs” such as “OpenCV,” which includes “segmentation” tools, as well as using “a video or other multimedia-editing web service” to “segment, edit, and

reprocess the content.” (Fontana ¶¶0139; *id.* ¶¶0058.) Each of these **segments** the image files before parallel processing. (Houh ¶¶153-154.)

3. Claim 6: “The method according to claim 1, further comprising: processing the image files to extract an object and determine a motion associated with the object.”

Kritt discloses the additional limitations of claim 6. (Houh ¶¶155-157.)

Kritt discloses identifying objects including human objects, such as faces and facial features. (*See* claim 1[d].) Kritt further discloses that “operation 510 may include extracting an identified object for further processing.” (Kritt, ¶¶0043.) Kritt additionally teaches **determining a motion associated with the object.** Specifically, Kritt identifies emotions by “comparing motion vectors” that “may be based on deformation of facial features”—in other words, by determining a motion associated with the face and facial features. (Kritt, ¶¶0047; Houh ¶¶157.)

4. Claim 7: “The method according to claim 1, wherein the text is generated by segmenting the audio before processing the audio files in parallel.”

The additional limitations of claim 7 are obvious over Kritt in view of Fontana. (Houh ¶¶158-164.)

Kritt discloses **segmenting the audio before processing.** For example, Kritt describes segmenting audio by classifying whether a segment is “silence, speech (spoken words), music or ... sound effect.” (Kritt ¶¶0057; *see also* Kritt ¶¶0058-60 (describing use of segments in audio classification).) Then, “spoken words may be

determined from audio features classified as speech” (i.e., “**the text is generated by segmenting the audio**”), including in parallel as described for claim 1[b]. (Kritt ¶0065.) This is illustrated in Figure 6, **segmenting the audio** and classifying the segments occurs in block 604 **before processing the audio files** through the audio transcription process in block 612:

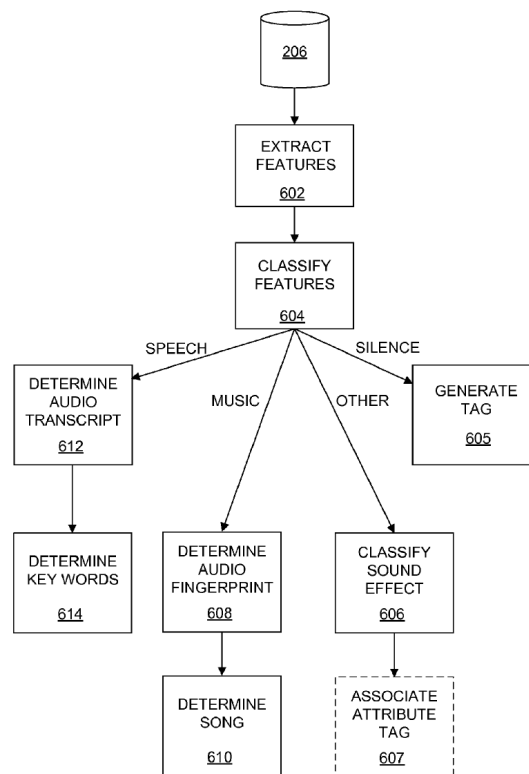


FIG. 6

(Kritt, Fig. 6.)

As discussed for claim 1[b], the combination of Kritt with Fontana renders obvious **processing the audio files in parallel**. Thus, the combination of Kritt and Fontana renders obvious **wherein the text is generated by segmenting the audio before processing the audio files in parallel**. (Houh ¶¶159-160.)

Additionally, a POSA would have found it obvious in light of Fontana's parallel processing disclosures to **"segment[] the audio before processing the audio files in parallel"** in order to take advantage of Fontana's teaching that multiple servers can be used to efficiently carry out tasks like audio processing.

Specifically, Fontana teaches multiple servers with "audio algorithms" that "are configured to share processing jobs, such that tasks can be performed by one or more of the computing systems, or separated and performed across multiple computing systems in parallel." (Fontana ¶0053.) As a result, "any of those computing systems can perform all or a portion of a processing job ... allowing multimedia content to be processed efficiently when necessary." (Fontana ¶0055.) Fontana additionally teaches that segmentation may occur in order to facilitate parallel processing. Fontana teaches that its "distributed computing systems ... allow for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing." (Fontana ¶0135; *see also* Fontana ¶0058 ("instructions provided ... to segment, edit, and reprocess the content").) Based on the foregoing, a POSA would thus have found it obvious that to implement audio processing in parallel, Fontana would **segment the audio files before processing the audio files in parallel**. (Houh ¶¶161-164.) It thus would have been obvious to a POSA based on the combination of Kritt with Fontana to **segment the audio**

before processing the audio files in parallel.

5. Claim 8: “The method according to claim 6, wherein the audio files and the image files are segmented at spectrum thresholds.”

The additional limitations of claim 8 are disclosed by Kritt. (Houh ¶¶165-167.) A **spectrum threshold** is a quantifiable valuation of some aspect of the media content. (Houh ¶166.) Indeed, the ’954 patent gives as an example of segmentation at spectrum thresholds that “the audio data can be processed and converted into a spectrum. Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low audio activities in the audio data.” (’954, 7:11-15.)

Kritt discloses segmenting audio and image files **at spectrum thresholds** in order to, for example, segment the audio and image files based on scene or content changes, and separate spoken word audio segments from other segments such as silence. For example, Kritt discloses segmenting a video by creating shot files (“**the image files**”) that are segmented by comparing the histogram for consecutive frames using thresholds. (Kritt ¶0040.) “If the histogram intersection of first and second consecutive frames is greater than a threshold, it may be inferred that the frames are similar, and consequently that the two frames are part of the same shot. On the other hand, if the histogram intersection of first and second consecutive frames is less than the threshold, it may be inferred that the two frames form a shot boundary.” (*Id.*)

This comparison of the intersection between histograms representing video frames, and using a threshold to determine whether the frames belong to the same shot, **segments the image files at spectrum thresholds.**

Similarly, Kritt discloses segmenting **the audio files** by comparing one or more audio features, such as “energy contour, volume contour, or frequency domain feature, such as short term energy, bandwidth, entropy, spectral centroid, Mel-Frequency Cepstral Coefficients, or a Discrete Wavelet Transform” to distinguish, for example, “speech” segments from “silent” segments. (Kritt ¶¶0056-0059.) A POSA would have understood and found it obvious that such segmenting would involve **spectrum thresholds**, such as a threshold energy or volume, or other information in a frequency domain (such as spectral centroid), that differentiates speech from silence. (*See, e.g.*, claim 7; Houh ¶167.)

This use of silence to segment audio aligns with the '954 Patent's disclosure that, for example, “quiet periods in the audio data can be detected, and the segmentation can be defined by the quiet periods....Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low audio activities in the audio data.” ('954, 7:8-17; Houh ¶167.)

6. Claim 11: “The method according to claim 1, further comprising: generating metadata.”

The additional limitations of claim 11 are obvious over Kritt in view of

Fontana. (Houh ¶168.) This combination **generates metadata**, including tags, tag files, and other metadata about a video. For example, Krittr teaches “various shot attributes may be determined and recorded as metadata. ... Metadata for a shot may include mood, emotion, or theme where another shot attribute is associated with a known cinematic technique for evoking the mood, emotion, or theme.” (Krittr, ¶0042.) A POSA would further appreciate that each of the tags generated by Krittr—including identification of shots and scenes; tags identifying human and nonhuman objects, audio objects, and emotions; and probabilities associated with those tags—is **metadata**. (See, e.g., claims 1[b], 1[f], 1[g], 1[h]; Houh ¶168.)

7. Claim 12: “The method according to claim 11, wherein the metadata comprises a number of frames, a range of said frames and/or timestamp references.”

The additional limitations of claim 12 are disclosed by Krittr. (Houh ¶169.) Krittr discloses including **timestamp references** among the generated **metadata**. For example, as noted for claim 11, each tag generated in Krittr is metadata. Krittr teaches that, for tags for identified objects and attributes, “[e]ach tag may be associated with a time stamp that indicates the start and stop time in which the object or attribute is rendered or otherwise associated.” (Krittr, ¶0032.) Furthermore, “[e]ach emotion object may be of a predefined type and associated with a time stamp.” (Krittr, ¶0034.) Likewise, audio tags are time stamped: “An audio tag with a speech type attribute may be associated with the audio feature, the tag having a

time stamp of the period of speech. ... An audio tag with music type attribute may be associated with the audio feature, the tag having a time stamp of the period of music.” (Kritt, ¶0059.)

D. Ground 2: Claims 2-3 and 9-10 Are Obvious Over Kritt in view of Fontana and Lau

1. Claim 2: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”

Claim 1 is obvious over Kritt in view of Fontana for the same reasons as in Ground 1. The additional limitations of claim 2 are obvious over Kritt in view of Fontana and Lau. (Houh ¶¶170-175.)

Like Fontana, Lau discloses using **natural language processing** to determine context associated with a video. For Lau, natural language processing is performed for purposes of matching the video to an advertisement. Lau teaches that “based on one or more keywords, ad units from the ad matrix are determined.” (Lau ¶0038.)

Lau additionally teaches that this process can build off natural language processing: “Recognition engine 212 receives rich media content that may be accessed by a user and uses correlation recognition detection techniques to recognize the content....In another embodiment, it could be natural language processing engine.” (Lau, ¶0041.)

A POSA would have appreciated that using recognition detection techniques, such as natural language processing, to recognize the content is **applying natural language processing to the text to determine context associated with the video.**

Rationale and Motivation to Combine (Kritt and Fontana with Lau): It would have been obvious to combine Kritt and Fontana with Lau in order to use Lau's disclosures for how to place advertisements within a video to carry out Fontana's disclosures that advertisements can be linked to video content based on the results of the audio and image processing, and a POSA would have had a reasonable expectation of success in the combination. (Houh ¶¶172-175.) Like Kritt and Fontana, Lau is an analogous reference to the '954 patent, addressed to the field of processing and generating multimedia content, including video, and specifically how to search within such multimedia content. ('954, 1:36-38 ("The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information."); '954, p.1 (Abstract) ("The text and the video content can be cross-referenced with the video."); Lau, ¶0002 ("Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content."); Lau ¶0006 ("In one embodiment, an advertisement is matched to subject matter in a portion of rich media content, such as digital video"); Lau ¶0053 ("Content may be searched to determine if the content includes the keywords."); *see* claim 1[a] (analogous analysis for Kritt and Fontana).) Furthermore, both Fontana and Lau disclose linking advertisements to the multimedia content. ('954, 2:5-8 ("An advertisement and/or product or service

recommendations can be placed at a specific time in the video based on the video content and/or section symbol of a video image.”); Fontana ¶0117 (“FIG. 7M illustrates example advertisement data, 716 that can be used in association with multimedia content, to link one or more advertisements with multimedia content during playback.”); Lau ¶0002 (“Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content.”).)

Fontana expressly discloses that advertisement data could be linked to the multimedia content: “FIG. 7M illustrates example advertisement data 716 that can be used in association with multimedia content, to link one or more advertisements with multimedia content during playback.” (Fontana ¶ 0117.) But Fontana does not teach how to identify advertisements to link with multimedia content; rather, “the matching of advertisements and content occurs based on a decision process separate from the content delivery system of the present disclosure.” (Fontana ¶ 0117.) A POSA thus would have appreciated from Fontana’s disclosures that the combination of Kritt and Fontana could be further adapted to deliver an advertisement, and would have been motivated to further combine with an advertisement-matching method, such as that disclosed by Lau, in order to practice Fontana’s disclosures of linking advertisement information with the multimedia content analyzed in Kritt and Fontana. (Houh ¶173.)

A POSA would have had a reasonable expectation of success in combining Kritt and Fontana with Lau. It would have been obvious to a POSA that the text (including audio transcript and key words) and tags (including object tags and emotion tags) generated by Kritt with Fontana could be used as the keywords/content to search for matching advertisements to video content according to Lau. (Houh ¶174.) Indeed, Kritt contemplates that the metadata, including tags, generated by Kritt can be used to identify recommended scenes “that have human, nonhuman, audio, and emotion objects, and key words that are similar to the scene selected by the viewer.” (Kritt, ¶0018.) A POSA would have appreciated from this disclosure that just as the metadata generated by Kritt could be used to identify similar scenes, it could equally be adapted to identify advertisements similar to scenes, as taught by Lau. (Houh ¶174.)

Likewise, a POSA would have appreciated that the audio and language processing algorithms taught in Lau—including, in particular, natural language processing—could be used in the context of Kritt with Fontana. Kritt discloses performing its audio processing for speech-to-text recognition “using any known technique,” and a POSA would have appreciated that the natural language processing described in Lau was one such well-known technique. (Kritt, ¶0065; Houh ¶175.)

2. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to extract a topic.”

The additional limitations of claim 3 are obvious over Kritt in view of Fontana and Lau for the same reasons as explained for claim 2. A POSA would have appreciated that Kritt’s key words, Fontana’s keywords, and Lau’s keywords are **topics** that could be extracted using **natural language processing**, as taught by Lau (*see* claim 2, *supra*; Houh ¶176).

3. Claim 9: “The method according to claim 1, further comprising: generating an advertisement based on the text and the video data.”

The additional limitations of claim 9 are obvious over Kritt in view of Fontana and Lau. (Houh ¶¶177-182.) The combination of Kritt with Fontana and Lau **generates an advertisement based on the text and the video data**. Specifically, the advertisement generated by Lau is created by compiling ad units, each of which is selected by matching the ad units to the video based on the text (ie, the audio transcript and keywords) and the video data (ie, the metadata generated by Kritt with Fontana).

Lau discloses performing a search to determine advertisements whose content and/or topics match the content and/or topics for portions of video content. The search involves counting the number of keyword and/or concept matches for a search term near a particular time in the video content. As previously noted, a POSA would

have found it obvious to combine Kritt and Fontana with Lau to use **the text and the video data** generated by Kritt with Fontana (e.g., the audio transcript, keywords, and metadata from audio and image processing) as keywords and concepts searched by Lau for purposes of identifying ads or ad units to match to the video based on context or subject matter. (*See* claim 2; Houh ¶178.)

Lau teaches matching keywords and concepts in a video to ads based on context or subject matter: “Correlation engine 202 may relate a unique content ID with a time series of keywords and concepts (that advertisers have purchased), and in turn, relate the keywords and concepts to ads submitted by advertisers.” (Lau ¶ 0051.) The “content ID” acts as “[a]n identifier for the content being served.” (Lau ¶0110.) For example, “for each ad, correlation engine 202 finds candidate content that may be relevant. This is done by searching for content in the index to match the keywords, categories, and concepts associated with the ad to information in the content.” (Lau ¶ 0056.)

Lau teaches that the search to match advertisements to video content may be performed by combining the results of different types of analyses, such as text with images as well as conceptual ideas like context: “Advertisers may buy correlation information, such as keywords, phrases or concepts, either through a bidding process or some other means, and submit their ads and related information to correlation engine 202 through correlation assistant 214. ... The phrases may be any

combination of words and other information, such as symbols, images, etc. The concepts may be a conceptual idea of something.” (Lau, ¶0046.) Thus, a POSA would have found it obvious that in the combination of Kritt and Fontana with Lau, correlating advertisements to the video could be **based on the text** (speech-to-text transcript and any derived keywords) **and the video data** (the identified objects and related metadata). (Houh ¶¶179-180.)

For each ad, Lau identifies a set of “candidate times” within the video, then scores each based on the strength of the match: “the more keywords/concepts for the ad matches near the time, the higher the score.” (Lau ¶ 0057.)

Lau additionally discloses that “[a]n advertisement may be broken into ad units” that are each “a subset of a larger advertisement. ... Thus, advertiser system 106 is not restricted to just serving an entire advertisement. Rather the most relevant pieces of the advertisement may be selected from the matrix of ad units.” (Lau ¶ 0023.) As a result, “[c]orrelation engine 202, when determining the advertisement, may determine one or more ad units that correlate to the subject matter,” and “the 330 ad units may be combined and presented to the user” as the final advertisement. (Lau ¶ 0038.) Thus, the combination **generates an advertisement based on the text and the video data**. (Houh ¶182.)

4. Claim 10: “The method according to claim 9, further comprising: placing the advertisement in the video at a preferred time.”

The additional limitations of claim 10 are obvious over Kritt in view of Fontana and Lau. (Houh ¶¶183-186.)

Specifically, Lau teaches that its “advertisements are time aligned to correlate to the subject matter to maximize the revenue.” (Lau ¶ 0045.) A POSA would have appreciated that time-aligned advertisements must be **placed in the video at a preferred time**. (Houh ¶184.) This understanding is consistent with Lau’s teachings on how to place advertisements in the video.

For example, Lau discloses that “the advertisement may be displayed in serial, parallel, or be injected into the rich media content.” (Lau ¶ 0034; *see also* Lau ¶ 0084 (“rendering formatter 204 can determine that an advertisement should be rendered serially relative to the portion of rich media content, in parallel to the portion of rich media content, or injected into the rich media content”); Lau claim 5 (“the advertisement is injected into or laid on top of portion of rich media content”).) A POSA would have appreciated that injecting the advertisement into the video **places the advertisement in the video at a preferred time**. (Houh ¶185.)

Lau likewise discloses that the ad units may be turned into “video that may serve as pre/mid/post-roll.” (Lau ¶0027.) Playing an advertisement as pre-roll, mid-

roll, or post-roll each likewise **place the advertisement in the video at a preferred time.** (Houh ¶186.)

E. Ground 3: Claim 13 Is Obvious Over Kritt

1. Independent Claim 13: “A system for extracting data from a video, comprising:” (Claim 13[pre])

Assuming the preamble provides a claim limitation, Kritt discloses it. (Houh ¶187.) Ground 1 explains several types of data Kritt extracts from a video. (*See, e.g.,* claim 1[pre], 1[a], 1[c], 1[d].) To highlight several examples, Kritt discloses extracting from a video both visual objects (Kritt, ¶ 0043 (“The operation 510 may include extracting an identified object for further processing. For example, an extracted human facial object may be further processed to determine the identity of the viewer or to determine a facial expression of the viewer.”)) and audio features (Kritt, ¶0056, (“one or more audio features or audio signal descriptors may be extracted from an audio file 206.”).)

(a) “a computer processor; and” (Claim 13[a])

Kritt discloses **a computer processor.** (Houh ¶188.) For example, Kritt explains that its system is implement in a “computer system” that includes “one or more processors 102.” (Kritt, ¶0019.)

(b) “a non-transitory computer readable medium containing instructions directing the system to execute the steps of:” (Claim 13[b])

Kritt discloses the additional limitations of claim 13[b]. (Houh ¶189.) Kritt

teaches that its system is implemented in “one or more computer readable medium(s) having computer readable program code embodied thereon,” where “a computer readable storage medium may be any tangible medium that can contain, or store, a program for use by or in connection with an instruction execution system, apparatus, or device.” (Kritt, ¶¶0070-71.)

**(c) “converting audio associated with the video to text;”
(Claim 13[c])**

For the same reasons explained in claim 1[c] in Ground 1, Kritt discloses claim 13[c]. (Houh ¶190.)

(d) “converting images associated with the video to video data, wherein the video data comprises identified objects;” (Claim 13[d])

For the same reasons explained in claims 1[d] and 1[g] for Ground 1, Kritt discloses claim 1[d]. In particular, claim 1[d] explains Kritt’s disclosures around identifying objects in images, and claim 1[g] explains Kritt’s disclosures around converting images associated with the video to video data, wherein the video data comprises objects previously identified in claim 1[d]. (Houh ¶191.)

**(e) “extracting topics from the text and the video data;”
(Claim 13[e])**

Kritt discloses claim 13[e]. (Houh ¶¶192-195.) Kritt discloses **extracting topics from the video data** for the same reasons explained in claim 1[e]. The topics extracted from the video data are shots/scenes, and emotions, moods, or themes.

Kritt additionally discloses **extracting topics from the text**, also in the form of scenes and emotions. (*See also* Ground 1, claim 1[h].)

Kritt teaches analyzing the audio data, including the text, to determine scenes. In particular, “a segment of two or more video frames that includes common objects, that is intended to convey common emotional content, that is intended to convey an element of a story, that is accompanied by a common audio segment, or some combination of the foregoing may be classified as a scene.” (Kritt, ¶0050.) Kritt explains that audio objects, which includes the transcript (**the text**), may be used to determine scenes: “Human, nonhuman, and audio objects may be used to classify a particular segment of a video as a scene.” (Kritt, ¶0017.)

Kritt additionally teaches analyzing the text to determine emotions. For example, Kritt discloses that “the objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.” (Kritt, ¶0017.) The objects include audio objects, such as spoken words in the audio transcript (**the text**). (*See, e.g.*, Kritt, ¶0016 (“A video may include spoken words, ... which may be referred to herein as audio objects.”); Kritt, ¶0065 (“An audio transcript may include all of the words spoken in the video.”).) Additionally, the key words are derived from **the text** (audio transcript). (*See, e.g.*, Kritt, ¶0016 (“The spoken words may include ‘key words.’”); Kritt, ¶0017 (“In addition, key words that are spoken by human objects in the video

are identified.”); Kritt, ¶0066 (“a key word may be determined from the audio transcript by counting the frequency of occurrences of words, the most frequently occurring verbs and nouns being determined to be key words.”).)

- (f) **“assigning probabilities of accuracy for each of the identified objects, wherein the probabilities are based on the extracted topics;” (Claim 13[f])**

For the same reasons discussed above in Ground 1 for claim 1[f], Kritt discloses and renders obvious claim 1[f]. (Houh ¶196.)

- (g) **“cross-referencing the text, the video data, and the topics with the video to determine contextual topics;” (Claim 13[g])**

Kritt discloses claim 13[g]. (Houh ¶197.) As discussed above for claim 1[e] in **Ground 1**, Kritt **determines contextual topics** in the form of emotions/moods and shots/scenes. As discussed for claim 1[e] and 1[h], this determination is based on cross-referencing information derived from the image files, which Kritt stores as visual tags (**video data**, see claim 1[g]), **the text** in the form of the audio transcript and resulting key words, and **the video** (other attributes and metadata about the video). Additionally, as explained in Ground 1, claim 1[h], determining the contextual topic of an emotion additionally cross-references the topic of scene. Likewise, as also explained in Ground 1, claim 1[h], determining the contextual topic of a scene additionally cross-references the topic of emotion. Thus, Kritt

Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

discloses **cross-referencing the text, the video data, and the topic with the video to determine contextual topics.**

- (h) **“generating a contextual text, an image, or an animation based on the determined contextual topics; and” (Claim 13[h])**

For the reasons explained in Ground 1 for claim 1[i], Kritt discloses limitation 13[h]. (Houh ¶198.)

- (i) **“generate a content-rich video based on the generated contextual text, image, or animation.” (Claim 13[i])**

For the reasons explained in Ground 1 for claim 1[j], Kritt discloses limitation 13[i]. (Houh ¶199.)

VII. CONCLUSION

Petitioner respectfully requests IPR institution.

Dated: May 9, 2025

Respectfully submitted,

COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Avenue NW
Suite 700
Washington, DC 20004
Tel: (650) 843-5001
Fax: (650) 849-7400

By: / Heidi L. Keefe /
Heidi L. Keefe
Reg. No. 40,673
Counsel for Petitioner

CERTIFICATE OF COMPLIANCE WITH WORD COUNT

Pursuant to 37 C.F.R. § 42.24(d), I certify that this petition complies with the type-volume limits of 37 C.F.R. § 42.24(a)(1)(i) because it contains 13,579 words, according to the word-processing system used to prepare this petition, excluding the parts of this petition that are exempted by 37 C.F.R. § 42.24(a) (including the table of contents, a table of authorities, mandatory notices, a certificate of service or this certificate word count, appendix of exhibits, and claim listings).

DATED: May 9, 2025

COOLEY LLP
ATTN: Patent Docketing
1299 Pennsylvania Avenue NW
Suite 700
Washington, D.C. 20004
Tel: (650) 843-5001
Fax: (650) 849-7400

/ Heidi L. Keefe /
Heidi L. Keefe
Reg. No. 40,673

CERTIFICATE OF SERVICE

I hereby certify, pursuant to 37 C.F.R. Sections 42.6 and 42.105, that a complete copy of the attached **PETITION FOR INTER PARTES REVIEW OF U.S. PATENT NO. 10,218,954 B2**, including all exhibits (**Nos. 1001-1010**) and related documents, are being served via Federal Express on the 9th day of May, 2025, the same day as the filing of the above-identified document in the United States Patent and Trademark Office/Patent Trial and Appeal Board, upon Patent Owner by serving the correspondence address of record with the USPTO as follows:

27890 - STEPTOE LLP/DC
1330 CONNECTICUT AVENUE, N.W.
WASHINGTON, DC 20036

And, via Federal Express upon counsel of record for Patent Owner in the litigation pending before the U.S. District Court for the Northern District of California entitled *Cellular South, Inc. v. Google LLC*, Case No. 4:25-cv-01487-YGR (N.D. Cal.) as follows:

Robert Sean Hill
Holland & Knight LLP
1722 Routh Street
Suite 1500
Dallas, TX 75201
214-964-9421
Fax: 214-964-9501
Email: robert.hill@hklaw.com

DATED: May 9, 2025

/ Heidi L. Keefe /
Heidi L. Keefe
Reg. No. 40,673
COOLEY LLP
1299 Pennsylvania Ave. NW,
Suite 700
Washington, D.C. 20004
Tel: (650) 843-5001