

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,

Petitioner,

v.

CELLULAR SOUTH, INC.,

Patent Owner.

Case No. IPR2025-00876

U.S. Patent No. 10,218,954

PATENT OWNER'S PRELIMINARY RESPONSE

TABLE OF CONTENTS

I.	INTRODUCTION	1
	Background	3
II.	OVERVIEW OF THE '954 PATENT.....	6
III.	OVERVIEW OF THE CITED PRIOR ART	16
	A. Overview of Kritt.....	16
	B. Overview of Fontana.....	24
	C. Overview of Lau	27
IV.	LEVEL OF ORDINARY SKILL IN THE ART	27
V.	CLAIM CONSTRUCTION	27
VI.	THE PETITION FAILS TO SHOW THAT THERE IS A REASONABLE LIKELIHOOD THAT PETITIONERS WILL PREVAIL ON A SINGLE CLAIM.....	28
	1. Ground 1: The Combination of Kritt and Fontana Fails to Render the Claim 1-8, 11 and 12 Obvious	28
	a. Kritt + Fontana Fail to Render Obvious “determining a contextual topic from the image files as claimed in 1[e].....	28
	b. Kritt + Fontana Fails to Render Obvious “Assigning a Probability of Accuracy to the Identified Object Based on the Contextual Topic” as claimed in 1[f].....	32
	c. Kritt + Fontana Fails to Render Obvious “converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic” as claimed in 1[g]	35
	d. Kritt + Fontana Fails to Render Obvious “cross-referencing the text and the video data with the video to determine the contextual topics” as claimed in 1[h]	36
	e. Kritt + Fontana Fails to Render Obvious “generating a contextual text, an image, or an animation based on the determined contextual topics” as claimed in 1[i]	39

f.	Kritt + Fontana Fails to Render Obvious “generating a content-rich video based on the generated text, image, or animation” as claimed in 1[j]...41
g.	A Person of Ordinary Skill in the Art Would Not be Motivated to Combine Kritt and Fontana.....42
h.	Houh’s Expert Testimony Is Entitled to Little or No Weight 45
2.	Ground 2: The Combination of Kritt + Fontana + Lau Does Not Render Obvious Claims 2-3 and 9-10. 47
3.	Ground 3: Petitioner Fails to Demonstrate that Claim 13 is Anticipated by Kritt.....47
a.	Kritt does not disclose “extracting topics from the text and video data” as required by limitation 13[e].....48
b.	Kritt does not disclose “assigning probabilities of accuracy for each of the identified objects, where the probabilities are based on the extracted topics” as required by limitation 13[f].....50
c.	Kritt does not disclose “cross-referencing the text, the video data, and the topics with the video to determine contextual topics” as required by limitation 13[g] 50
d.	Kritt does not disclose “generating a contextual text, an image, or an animation based on the determined contextual topics” as required by limitation 13[h] 50
e.	Kritt does not disclose “generating content-rich video based on the generated contextual text, image, or animation” as required by limitation 13[i].....50
VII.	CONCLUSION.....50

TABLE OF AUTHORITIES

Page(s)

Cases

<i>Cellular South, Inc. v. Google, LLC</i> , Case No. 4:25-cv-01487-YGR (N.D. Cal. Aug. 5, 2025)	v
<i>Cellular South, Inc. v. Google, LLC</i> , Case No. 6:24-cv-00245 (W.D. Tex. May 9, 2024)	v
<i>Kinetic Techs., Inc., Petitioner, v. Skyworks Sols., Inc.</i> , IPR2014-00529, 2014 WL 4787238 (P.T.A.B. Sept. 23, 2014)	44
<i>Upjohn Co. v. Mova Pharm. Corp.</i> , 225 F.3d 1306 (Fed. Cir. 2000)	44
<i>Wellman, Inc. v. Eastman Chem. Co.</i> , 642 F.3d 1355 (Fed. Cir. 2011)	27
<i>Xerox Corp. v. Bytemark, Inc.</i> , IPR2022-00624, 2022 WL 3648989 (P.T.A.B. Aug. 24, 2022).....	44

Statutes

35 U.S.C. § 313	1
-----------------------	---

Other Authorities

37 C.F.R. 42.6(e).....	50
37 C.F.R. § 42.24	51
37 C.F.R. § 42.65(a).....	44
37 C.F.R. § 42.107	1

PATENT OWNER'S EXHIBIT LIST

Exhibit No.	Document
2001	U.S. Patent No. 10,218,954
2002	U.S. Patent No. 9,940,972
2003	U.S. Patent No. 11,126,853
2004	Article, Vū Digital Wins Coveted TMC Communications Solutions 2015 Product of the Year Award
2005	Article, NAB Show Unveils Final 10 Participants Selected for SPROCKIT 2015
2006	Notice of References Cited in U.S. Patent Application No. 15/444,054 (Oct. 30, 2019)
2007	Information Disclosure Statement by Applicant in U.S. Patent Application No. 17/423,623 (July 20, 2021)
2008	CSI's Complaint for Patent Infringement against Google in <i>Cellular South, Inc. v. Google, LLC</i> , Case No. 6:24-cv-00245 (W.D. Tex. May 9, 2024)
2009	Proof of Service of Complaint
2010	Joint Case Management and Pretrial Order as Modified in <i>Cellular South, Inc. v. Google, LLC</i> , Case No. 4:25-cv-01487-YGR (N.D. Cal. Aug. 5, 2025)
2011	Transcript of Conference held July 7, 2025 in <i>Cellular South, Inc. v. Google, LLC</i> , Case No. 4:25-cv-01487-YGR (N.D. Cal. Aug. 5, 2025)

Exhibit No.	Document
2012	Preliminary Infringement Contentions in <i>Cellular South, Inc. v. Google, LLC</i> , Case No. 6:24-cv-00245 (W.D. Tex.) (served Dec. 13, 2024)
2013	Comparison of portions of Paper 2 with Ex. 1002

I. INTRODUCTION

Pursuant to 35 U.S.C. § 313 and 37 C.F.R. § 42.107, Patent Owner Cellular South, Inc. (“Patent Owner” or “CSI”) submits this Preliminary Response to the Petition for *Inter Partes* Review (“Petition”) filed by Google LLC (“Petitioner” or “Google”) and challenging claims 1-13 of U.S. Patent 10,218,954 (the “’954 Patent”).

On its merits, Google’s Petition is fundamentally flawed and Dr. Houh’s carbon-copy declaration cannot save it. What’s more, despite filing on the one-year deadline, the Petition is cobbled together, ignores key claim limitations, and does not warrant allocation of the Board’s limited resources. Because the Petitioner cannot show (and does not show) that there is a reasonable likelihood of prevailing on at least one claim, the Petition must be denied.

Petitioner’s Ground 1 fails for three primary reasons. First, Kritt’s disclosure ends where the ’954 Patent’s novelty begins. Even if Kritt’s “tags” meet the claimed “contextual topic” (they do not), Kritt’s video analysis ends there. In contrast, the crux of the ’954 Patent claims starts with the contextual topic as an initial input to generate new contextual information. As shown below, Claim 1 recites five additional steps—going well beyond Kritt’s teachings—that are either directly or indirectly based on the “*determined contextual topic*” and not taught or disclosed Kritt:

- 1[e] assigning a probability of accuracy to the identified object *based on the contextual topic*;
- 1[f] converting the image files associated with the video to video data, wherein the *video data comprises the object, the probability, and the contextual topic*;
- 1[g] cross-referencing the text and the video data with the video to *determine contextual topics*;
- 1[h] generating a contextual text, an image, or an animation *based on the determined contextual topics*;
- 1[i] generating a content-rich video *based on the generated text, image, or animation*.

In short, Petitioner's mapping of Kritt fails early and its problems only snowball from there. With each consecutive limitation, the disconnect between the prior art and the '954 Patent grows, as the references lack not only the foundational features but also the advanced functionalities that build upon them. The result is a cascade of failures, each more pronounced than the last.

Second, Petitioner concedes that Kritt's "containers" do not meet the "generating audio and image files from the video" limitation and relies on Fontana to fill this gap. *See* claim 1[a]. But the idea that Fontana's low-resolution thumbnail image could be used in Kritt to achieve the necessary object detection for extracting context, topics, and semantics is wrong and entirely unsupported. This is classic hindsight, not a legitimate rationale for obviousness.

Third, Dr. Houh's declaration is nothing more than a clone of the Petition. He offers no additional support—let alone evidence—as to how a POSA would understand Kritt, Fontana or their combination. In fact, the additions to Dr. Houh's declaration largely consist of replacing citations with the cited text itself. Grounds 2 and 3 fail for many of the same reasons as Ground 1. In view of these repeat failures, the Petition should be denied institution.

BACKGROUND

Patent Owner is a diversified telecommunications and technology company that focuses on delivering connectivity and technology solutions to advance the field and its community. While Patent Owner is a relatively small company compared to its competitors in the telecommunications industry, it has always been an innovator. For example, Patent Owner has worked to diversify its product offerings to remain innovative and competitive in the telecommunications and technology marketplace. As a result of these efforts, Patent Owner, through its subsidiary Vu Digital, created and offered "Video-to-Data" (or "V2D"), a cloud-based digital profiling and analytics system that could scan and turn video into searchable data faster than the real-time viewing of the video by using multiple servers that run simultaneously and in parallel.

V2D breaks down a video frame-by-frame and identifies the objects—such as text, audio, images, faces, locations, and logos—within each frame. V2D then

creates a chronological record of all the objects identified within the video. V2D was a transformative product that made video content as easy to search as text, thereby providing users with unprecedented video classification and clustering capabilities, as well as significantly enhanced search engine indexing, content personalization, and targeted advertising capabilities.

The applications for V2D are far reaching. V2D could be an incredibly useful technology for video data processing and content personalization. For example, V2D could be used by law enforcement to process body camera footage which is becoming more widely adopted.

Patent Owner launched its V2D product in May 2015. Mere months later, Patent Owner's V2D product was recognized by TCM—a global integrated media conglomerate and leading source of news and information for the communications and technology industries worldwide—as the “best-of-the-best technology solutions available on the market today.” V2D received TCM's coveted Communications Solutions 2015 Product of the Year Award that recognizes exceptional voice, data, and video communications products and services. Ex. 2004.

Patent Owner was also selected in 2015 to be part of the National Association of Broadcasters' year-long SPROCKIT accelerator program for innovative startups in media and entertainment. Ex. 2005. SPROCKIT is a global innovation platform

created to help large media, entertainment, and technology companies meet emerging tech start-ups to fast-track investment, acquisition, and partnerships between the two groups on new products and services. Petitioner is a corporate partner of SPROCKIT. *Id.*

SPROCKIT Sync is an invitation-only event put on three times a year by SPROCKIT that includes a series of private meetings between selected startups and corporate players in the media and entertainment industry. Following the launch of its V2D product, Patent Owner was invited to attend the SPROCKIT Sync conference that was held on June 18, 2015, at Google Tech Corners in Sunnyvale, California. There, Patent Owner presented its V2D technology at the conference. Google was one of the corporate partners for this event.

Patent Owner protects its intellectual property. As is relevant here, Patent Owner is the assignee of three patents related to its V2D technology including the '954 patent (Ex. 1001) at issue in this IPR, as well as U.S. Patent Nos. 9,940,972 ("the '972 patent") and 11,126,853 ("the '853 patent"), each entitled "Video to data." The '954 patent and '972 patents are related and Google filed IPRs challenging these two patents as well as the '853 patent. *See* IPR2025-00875 (challenging the '972 patent), and IPR2025-00877 (challenging the '853 patent.) At a high level, these patents are directed toward an improvement in the functionality of machine learned

video recognition and classification systems, particularly with regard to specific techniques for improving the accuracy of predictions made using image, audio, text, and other video data.

II. OVERVIEW OF THE '954 PATENT

The '954 patent addresses technological challenges in extracting meaningful metadata from video content, which is a persistent problem in video classification systems. Prior to the claimed invention in the '954 patent, video content was largely opaque to search engines and analytics platforms. While reverse image search and audio-to-text transcription existed, they operated as independent silos and lacked semantic, contextual understanding. The '954 patent addresses these limitations by integrating distributed processing, advanced image and audio analysis, and natural language processing to create a unified, and context-rich video data extraction engine.

The process described in the '954 patent first receives a video file and extracts both its audio track and individual image frames. These audio and image components are then distributed across multiple processors, enabling parallel analysis for efficient, near real-time processing. The audio is segmented, and each segment is transcribed into text using speech recognition techniques. Simultaneously, the image frames are processed to identify objects, symbols, brands, faces, and other features through various image analysis methods. Both the transcribed audio and the

descriptive text generated from images are further refined using natural language processing, which helps to extract context, topics, and semantic meaning, and can assign weights to identified elements based on their relevance. The information from both audio and image analyses are combined and cross-referenced, giving additional emphasis to topics or keywords that appear in both sources. This integrated data is used to generate a comprehensive textual description of the video's content, along with metadata such as timestamps, frame references, and topical tags. The final output can include text, metadata, images, or animations that describe the video. The key is the ability to extract semantic meaning from both the audio and visual components of the same video and then cross-reference them to develop a more holistic understanding of the video's content.

Figure 1 of the '954 patent, shown below, illustrates an embodiment of a video-to-content engine:

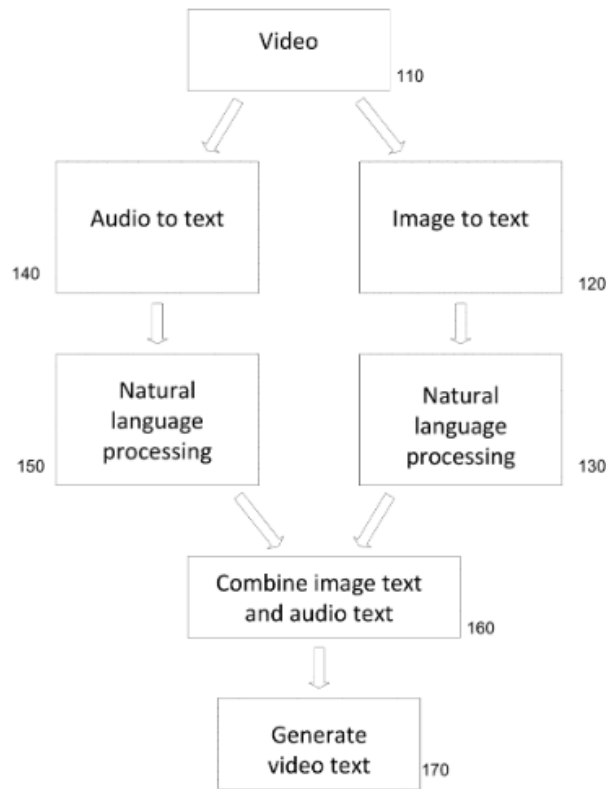


Figure 1

Starting with step 110, the video-to-content engine embodied on a computer, a server, or servers, extracts audio data and image data (e.g. images or frames forming the video) from the video stream in parallel processes. Ex. 1001 at 4:9-21. At step 120, the video-to-content engine performs image data processing on the image data extracted from the video stream by distributing the image files across a plurality of processors for parallel processing. *Id.* at 4:22-27. An example of the distributed image data processing is illustrated in Figure 2, shown below:

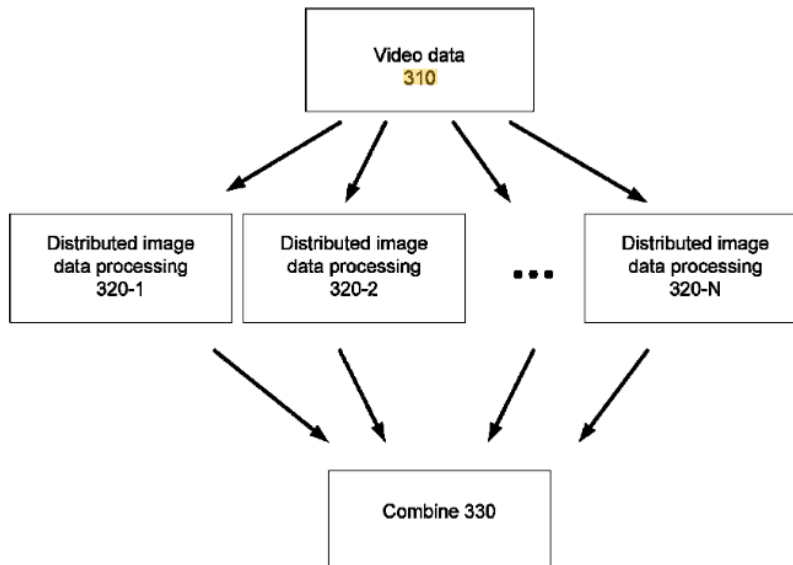


Figure 2

Ex. 1001 at Fig. 2.

As shown in Figure 2, the image data is segmented into image files and each image file is processed in parallel (e.g., by distributing image files to the plurality of processors for distributed processing 320-1 to 320-N). *Id.* Processing the image files comprises extracting one or more objects and identifying one or more objects. For example, using image processing techniques (such as color edge detection, image extraction, high-level vision and symbol detection, figure-ground separation, depth and motion perception, etc.) and an object template database, an object or symbol of an image can be extracted, isolated, and identified in the form of text describing the object. *Id.* at 4:28-56.

At step 130, natural language processing is applied to the image text to derive semantic information from the identification of the objects. *Id.* at 6:31-40. The image text can be modified as the video-to-text engine selects primary semantics from a plurality of possible semantics. *Id.* at 6:31-36. The video-to-content engine generates topical meta-data or topical descriptors that describe the content of the image files and applies this topical meta-data to the original video. *Id.* For example, if the object identified in the image is a dog, when compared to the topic database, the dog may be further described as running or jumping. *Id.* at 4:45-49. Importantly, because this topical meta-data generation process—using the distributed/parallel processing architecture described above—is repeated for additional symbols/objects within an image frame, the process develops granular contextual details about what is happening in each image of the video. As a result, the dog initially identified from the video frame, when compared to the topics database, might further associate the dog with the topical meta-data of “running” and “jumping.” *Id.* at 4:44-49. The image’s topical metadata are stored for the entire video or each image segment, resulting in several topic descriptors. *Id.* at 4:61-62. From this topical metadata, it is possible to generate a more detailed contextual representation of the video as showing a dog, which is running, on the beach, with people, by a resort. *Id.* at 4:57-

5:2. In other words, the contextual topics are not just a series of images, but rather, they explain what is happening in the image.

At step 140, the audio files generated from step 110 are processed and converted into text. *Id.* at 6:41-44. At step 150, and as with the image processing of step 130, natural language processing can be applied to the text generated from step 110 to derive semantic information from the audio files. *Id.* at 7:36-37. For example, the video-to-content engine generates topical meta-data or topical descriptors that describe the content of the audio files. *Id.* at 7:37-40. Continuing with the above example concerning topical metadata of a dog object with actions of “running” and “jumping” on a beach identified from the image files, suppose the audio text from the video contains the following dialogue:

Person 1: “Look at Buddy go! He's chasing that frisbee like a pro.”

Person 2: “I know, he loves playing fetch on the beach. The kids are having so much fun with him today.”

Person 1: “Yeah, the waves crashing in the background make it the perfect summer day.”

In this example, the video-to-content engine could derive additional topical metadata from the semantic information, such as: subject: dog (named Buddy); action: chasing frisbee; location: beach; activity: playing fetch; participants: kids, people (2 individuals); and environment: waves, summer.

At step 160, the text, video data, and contextual topics are combined and cross-referenced to determine topics common to both the text of the audio files and the video data. *Id.* at 7:41-44. When the topical metadata generated from the semantic information of an image is combined with the topical metadata generated from the semantic information of the audio, topics common to both the image and the audio are determined and weighted. *Id.* By cross-referencing the topical topics derived from the image analysis and the audio text, the video-to-content engine can identify the common topics and themes that describe the overall content of the video—e.g. common topics: dog, beach, playing/activity. The cross-referencing allows the video-to-content engine to validate and strengthen the contextual topics, since the visual and audio components both *independently* point to a video centered around a family's dog playing at the beach. In this example, the audio text provides additional context and details that complement the visual information (e.g., identifying the dog's name, the presence of kids, and the summer beach setting) through cross-referencing.

At step 170, the video-to-content engine generates video text describing the content of the video using the cross-reference of the combined text and video data. *Id.* at 7:44-47. For example, key words indicating topic and semantic that appear in both texts can be selected and/or emphasized. *Id.* at 7:47-49. The video-to-content

engine generates text, image, or animation based on the video text and places the generated text, image, or animation in the video stream. *Id.* at 7:44-52.

The output data from the combined image and audio generated topics is then applied in various ways to create content-rich video and enrich user experience or user defined outputs including “search engine optimization, video categorization, recommendation engines, advertisement targeting, content personalization, analytics, etc.” *Id.* at 8:15-28.

Building on the use of context—derived from both audio and visual data—to enable more reliable classification, identification, and personalization of video content, the '954 patent further describes leveraging this contextual information to improve its accuracy and minimize false positives. The '954 patent describes that:

Identifying various objects in an image can be a difficult task. For example, locating (segmenting) and *positively identifying an object in a given frame or image can yield false positives—locating but wrongfully identifying an object.* **Therefore, present embodiments can be utilized to eliminate false positives, for example, by using context.**

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.

Id. at 5:3-16

The '954 patent also describes how context in addition to a linear sequence of frames can be used to improve image recognition accuracy in difficult situations such as partial visible or self-occluded objects, lack of objects, faces or words, or in an overly cluttered image. *Id.* at 5:18-22. Here, the above contextual methodology is applied to linear sequencing of video frames instead of a single stand-alone image to derive context. In this way, an object identified in one frame can be detected as an anomaly when compared to a sequence of frames just before and after the purported anomaly. As a result, a lower probability of accuracy can be associated with the object anomaly to determine whether the image may be a false positive. The '954 patent describes various examples of how establishing probabilities of accuracy based on context of the video frames and audio transcription is used to improve results.

In certain instances, identification of an individual can be a difficult task. For example, facial recognition can become difficult when an individual's face is obstructed by another object like a football, a baseball helmet, a musical instrument, or other obstructions. An advantage of some embodiments described herein can include the ability to identify an individual without identification of the individual's face. **Embodiments can use contextual information such**

as association of objects, text, and/or other context within an image or video.

As one example, a football player scores a touchdown but rather than identifying the player using facial recognition, the play can be identified by object recognition of, for example, the player's team's logo, text recognition of the player's jersey number, and by cross referencing this data with that team's roster (as oppose to another team, which is an example of why the logo recognition can be important). Such embodiments can further learn to identify that player more readily and save his image as data.

Id. at 5:35-52.

Similarly, the audio transcript of a video can be used to derive certain context helpful in identifying and correcting or eliminating image false positives. In this way, an image anomaly or anomalies identified in a given video frame(s) are associated with time (time stamped) and correlated with a time range from the transcribed audio to establish certain probabilities of accuracy.

Id. at 5:53-59.

In summary, the '954 patent is directed to a method and system for distributed, parallel processing of video content, extracting and synthesizing data from both audio and image components, leveraging contextual information to improve the accuracy of object and topic identification, and generating rich metadata and content representations that make video as searchable and actionable as text.

Claim 1 of the '954 patent is recited below:

1[pre] A method to generate video data from a video, comprising:

1[a] generating audio files and image files from the video;

1[b] distributing the audio files and image files across a plurality of processors and processing them in parallel;

1[c] converting audio files associated with the video to text; identifying an object in the image files;

1[d] determining a contextual topic from the image files;

1[e] assigning a probability of accuracy to the identified object based on the contextual topic;

1[f] converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic;

1[g] cross-referencing the text and the video data with the video to determine contextual topics;

1[h] generating a contextual text, an image, or an animation based on the determined contextual topics;

1[i] generating a content-rich video based on the generated text, image, or animation.

Id., Claim 1.

III. OVERVIEW OF THE CITED PRIOR ART

A. Overview of Krit

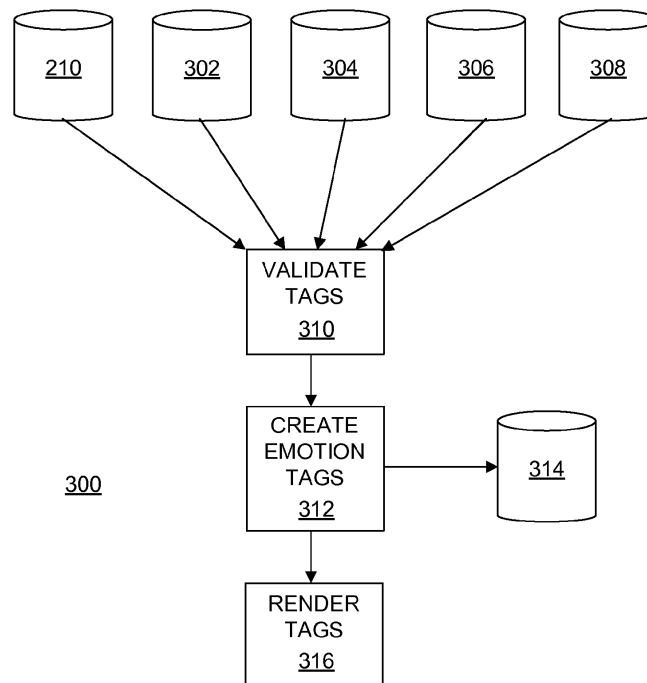
Kritt is fundamentally different from the '954 patent. Unlike the '954 patent's context-rich video data extraction engine, Krit discloses a simple user interface that visually summarizes the video through tags thereby allowing a user to understand

what is happening in the video at a point in time (i.e. bad language, music, sound effects). Kritt's approach is limited to the identification and static display of tags associated with objects and features in a video. Importantly, Kritt lacks any mechanism for extracting, weighting, or cross-referencing contextual topics as described and claimed in the '954 patent.

Kritt acknowledges that a typical video ranges in length from 20 minutes to over two hours and includes many scenes, each corresponding with a particular segment of the video. Scenes typically include many frames and may include one or more camera shots. Ex. 1003 at [0005]. Kritt defines a shot to "be a continuous sequence of frames captured without interruption". Ex. 1003 at [0033]. Kritt describes a segment as "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[.]" Ex. 1003 at [0050]. In sum, Kritt describes a scene as including one or more shots, where a shot represents a collection of frames captured by a video camera, and a segment represents two or more video frames featuring common objects and audio and/or conveying common content.

Though Kritt and the '954 patent both discuss the segmentation of video, they approach segmentation in different ways. Kritt discusses segmenting a pre-stored

audio/visual file container into shots and scenes, while the '954 segments the video and audio by generating respective images for multiple distributed and parallel processing. *Id.*; Ex. 2001 at 1:56-61. Figure 3 of Kritt, shown below, provides “a block diagram of its process for visually summarizing a video in a way that facilitates quick understanding by a viewer of the locations of particular types of content according to an embodiment.” Ex. 1003 at [0032].



Ex. 1003 at Fig. 3.

Kritt goes on to describe that “[t]he process may receive as input a visual tag file 302, an audio tag file 304, a key word tag file 306, an attribute tag file 308, and a metadata file 210.” The visual tag file corresponds to perceivable objects; the

audio tag file corresponds to audible objects; the key word tag file corresponds to key word objects; and the attribute tag file corresponds to attribute objects. Each tag may also be associated with a time stamp indicating start and stop time. Kritt provides that “[e]xemplary embodiments for automatically determining tags are described [] with reference to FIGS. 5-6 [and] in some embodiments, tags of one or more types may be wholly or partially determined using manual methods.” Ex. 1003 at [0032].

Kritt’s tag-based analysis of the segmented video is fundamentally object-centric and timestamp-driven. The processing in Kritt can further be characterized as a mechanical and static approach to video analysis. The ’954 patent, by contrast, describes a much more dynamic system, teaching methods for extracting topics from each segment, and then weighting those topics according to contextual cues—such as the co-occurrence of objects and audio events, or the semantic relationships between detected features and spoken words. This process enables the ’954 system to generate metadata and outputs that reflect not just the presence of objects or keywords, but the underlying context and topics of the video, which can be used for advanced applications. While Kritt offers some classification for the objects of a video, the ’954 is designed to produce a contextually meaningful output not simply a classification and timestamp.

Kritt describes operation 310 of Figure 3 to “include comparing a tag with one or more other tags associated with the same shot or scene for consistency[,]” and discloses that it “may include modifying a tag determined to be inconsistent with other tags associated with the same shot.” Ex. 1003 at [0033]. Alternatively, Kritt states that “if a probability or confidence parameter associated with the particular tag is above a threshold, it may be determined that the object was correctly identified and that the shot or scene includes multiple objects.” *Id.* As made clear from the above description, however, the probability or confidence level proposed by Kritt is merely associated with a particular tag, or the identification of objects. The “confidence parameter[s]” in Kritt operate differently from the “probability of accuracy” described in the ’954 patent, as the latter is not merely assigned to the identified object based on an arbitrary threshold. The ’954 patent weighs the probability of accuracy of the identified object based on the contextual topic determined from the image file. The ’954 patent illustrates by 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.” Ex. 1001 at 5:8-12. Or in another example, “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.” *Id.* at

5:13-16. The '954 patent does not stop there—the probability of accuracy is further utilized to convert the image files to video data, whereas Kritt's confidence parameters are limited solely to object identification. *See id.* at 14:10-12. In other words, Kritt's probability mechanism operates solely for the purpose of assigning tags to the segments of a video, and unlike the '954 patent, neglects to extend to any contextual analysis thereof.

Kritt goes on to disclose that “[i]n 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[,]” and finally that “[i]n a rendering operation 316, one or more tags of the tag files 302,304, 306, 308, and 314 may be rendered as one or more indicia on a display device according to known techniques.” Ex. 1003 at [0034]. This rendering is limited to visual summary and user interface display, however, without any underlying semantic or contextual synthesis or output. By contrast, the '954 does not merely render tags or icons for user selection, but generates a context-rich, semantically weighted output that can be used for advanced applications such as search engine optimization, recommendation engines, and targeted advertising, all based on the extracted and weighted contextual topics.

As discussed above, Kritt indicates that exemplary embodiments for automatically determining tags are described in Figures 5 and 6. Figure 5 of Kritt, shown and annotated below, illustrates a process for generating visual tags. Ex. 1003 at [0039].

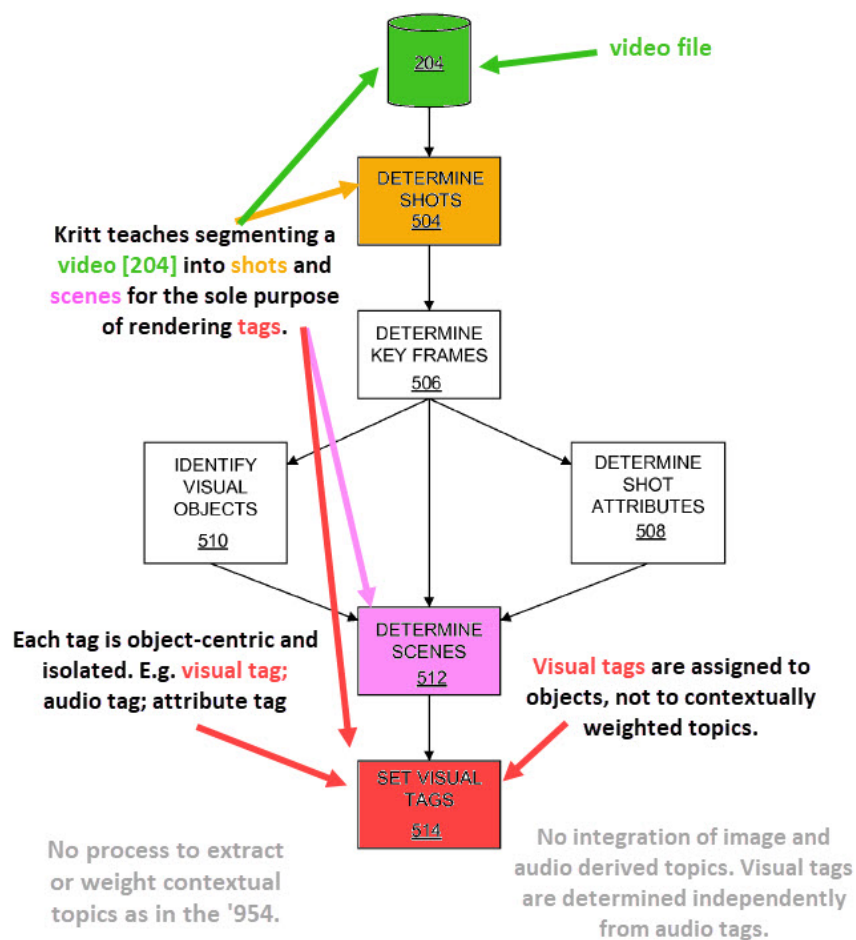


FIG. 5

Ex. 1003 at Fig. 5 (annotated).

Kritt states that “[t]he video file 204 may be parsed into shot files according to any known method.” Ex. 1003 at [0040]. Once the identification of visual objects and shot attributes are completed, Kritt states that “[i]n operation 512, a visual scene may be determined according to any known method[,]” *Id.* at [0051], and that “[t]he determination of a visual scene in operation 512 may include associating the scene with a probability or confidence parameter that is a measure of how likely the identified and grouped shots define a scene according to a ground truth specifying the temporal boundaries of a scene.” *Id.* at [0052]. Kritt also states that “[t]he determination of a visual scene in operation 512 may include associating an attribute tag with the scene.” *Id.* at [0054]. Finally, Kritt describes that “[i]n operation 514, visual tags may be associated or set for each scene.” *Id.* at [0055]. As shown above, Kritt’s application of object identification techniques is limited to the assignment of corresponding “tags” and does not extend to the extraction or synthesis of contextual topics as described by the ’954 patent. The ’954 system is designed to use such object and attribute identification as a mere input for a more advanced process of topic extraction, context assignment, and semantic weighting, which is entirely absent from Kritt. In other words, the ’954 patent uses object identification solely as a precursor to topic extraction and contextual weighting, **not as an end in itself**.

Again, Kritt describes a fundamentally tag-based and object-centric approach to video analysis that lacks any mechanism for extracting, weighting, or cross-referencing contextual topics as defined and claimed in the '954 patent. The differences between these patents are not merely in degree, but in kind, with the '954 patent providing a context-driven, topic-centric solution that is absent from Kritt's tag-based summarization.

B. Overview of Fontana

Petitioner relies on Fontana solely for its teaching of “generating audio files and image files” from a video and using distributed and parallel processing to process those files.

The core of the Fontana is directed to a system for receiving, processing, and delivering multimedia content by associating external metadata with the content. *See* Ex. 1004 at [0041]. To process multimedia content, Fontana teaches using an audio processing module 606 and a video processing module 608 and describes that “each 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]. For audio processing, Fontana teaches that “an audio separation operation 1128 strips, or extracts, the audio from the multimedia content” to generate an audio file. For video processing, Fontana relies on a thumbnail extraction module that

generates low-resolution thumbnail images for quick browsing and scene selection, rather than for detailed object identification or recognition. Ex. 1004 at Fig. 6 (620, 608), [0095].

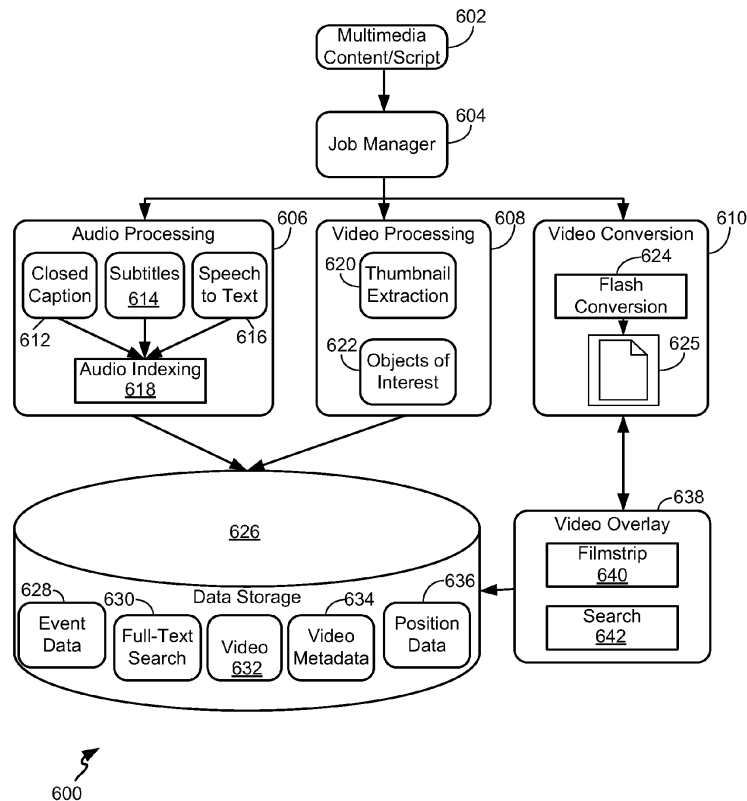


FIG. 6

Ex. 1004 at Fig. 6.

Thumbnails in Fontana are designed for rapid navigation and visual summarization, not for extracting granular object details. Ex. 1004 at [0095], [0139]. As Fontana itself teaches, “[t]he thumbnail extraction module 618 is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest (for example a first frame, a last frame, and

immediately following major scene or sound changes in the content).” *Id.* at [0095]. This language underscores thumbnails are created as simple visual markers tied to broad content transitions, not as tools for detailed object analysis. Fontana does not describe any process for extracting or preserving fine-grained visual information; rather, it teaches that thumbnails are summary representations intended to facilitate navigation of the video— not to enable semantic or contextual understanding at the object level.

Fontana’s thumbnail generation process is inherently limited by its focus on efficiency, prioritizing the creation of small, low-fidelity images that serve only as simple visual cues for users navigating video content. Ex. 1004 at [0095], [0139]. The reduction in image resolution inherent in thumbnail generation results in significant loss of definition, rendering it impractical for accurate object detection. *Id.* Fontana’s image processing method and its focus on generating low-resolution thumbnails is contrary to the ’954 Patent’s use of high-resolution image frames extraction for generating synthesized semantic context from the image. Ex. 1004, [0087]-[0097]. Thus, as explained below, a POSA would not look to Fontana’s low-resolution image files for use with Krittr’s video summarization user-interface.

C. Overview of Lau

Petitioner relies on Lau as a tertiary reference that—when combined with Kritt and Fontana—allegedly renders obvious placing advertisements in videos based on the results of audio and image processing. However, only dependent claims are challenged using Lau and because Lau does not fill any of the holes left by the faulty combination of Kritt and Fontana for independent claim 1, a deep dive into Lau's teachings cannot save Petitioner's combination.

IV. LEVEL OF ORDINARY SKILL IN THE ART

For the limited purpose of this Preliminary Response, Patent Owner does not contest Petitioner's definition of a person of ordinary skill in the art, but it reserves the right to do so if trial is instituted.

V. CLAIM CONSTRUCTION

Similarly, for the limited purpose of this Preliminary Response, Patent Owner does not contend that any claim terms need construction to arrive at the conclusion that the Petition is substantively deficient. *See Wellman, Inc. v. Eastman Chem. Co.*, 642 F.3d 1355, 1361 (Fed. Cir. 2011) (terms “need only be construed ‘to the extent necessary to resolve the controversy’”). Patent Owner reserves the right to construe terms if trial is instituted.

VI. THE PETITION FAILS TO SHOW THAT THERE IS A REASONABLE LIKELIHOOD THAT PETITIONERS WILL PREVAIL ON A SINGLE CLAIM

1. Ground 1: The Combination of Kritt and Fontana Fails to Render the Claim 1-8, 11 and 12 Obvious

The Petition relies on Kritt as teaching all the limitations of independent claim

1. Petitioner further relies on Fontana as a back-up reference for limitations 1[a] (“generating audio files and image files from the video”) and limitation 1[b] (“distributing the audio file and the image files across a plurality of processor and processing the audio files and the image file in parallel.”) For the reasons set forth below, the combination of Kritt and Fontana fail to render obvious independent claim 1 nor would a POSA be motivated to implement Fontana’s low-resolution thumbnail in place of Kritt’s high-resolution key frame for detailed object detection. Finally, Dr. Houh’s declaration is conclusory and fails to offer any supporting evidence.

a. Kritt + Fontana Fail to Render Obvious “determining a contextual topic from the image files as claimed in 1[e]”

Claims 1[d] and 1[e] require “identifying an object in the image files” and then “determining a contextual topic from the image files.” With respect to determining a contextual topic, the ’954 patent explains that using *contextual information regarding an object within an image* helps reduce false positives (Ex. 1001 at 5:4-8; 5:42-44) and that an object (or symbol) from an image is “compared to a topic database to identify a topic (such as an event)”. *Id.* at 4:44-49. For

example, a dog may be identified as an object in the image, and when compared to the topic database, the dog is identified as “running or jumping,” thereby giving context to what the dog is doing. *Id.* The '954 patent further teaches that context can also be applied to a series of images to determine whether an object was accurately identified. For example, if the context of an image is identified as a horse race but the object is identified as a mule or donkey, the probability that the object was correctly identified can be decreased. *Id.* at 5:11-16. Thus, a contextual topic provides context about the object(s) in the image, for example: object = dog; contextual topic = playing fetch.

Petitioner alleges that Kritt's disclosure of “shots/scenes” and “emotions” meets the claimed “*determining a contextual topic from the image files*” limitation. *Pe.* at 26-29. Regarding “shots/scenes,” Kritt explains that a “shot” is a “continuous sequence of frames” and that a “histogram” is used to “infer that [the] frames are similar and consequently that two frames are part of the same shot.” *Ex.* 1003 at [0040]. Histogram intersection is a method, often used in image/video processing and computer vision applications, for comparing histograms to determine potential similarity between two images or frames of video. The histogram from a given frame is computed and compared to the corresponding histogram in the next frame to determine if both frames contain similar content to be included into a given shot—

a series of such shots would form a scene or multiple scenes. For instance, a series of frames for outdoor backyard scenery would contain highly similar content (e.g. grass, trees, etc.) yielding highly similar histograms for each of the frames and thereby potentially grouping these frames into a single shot. However, the histograms, their intersection and/or the classification of the frames into a single shot do not provide any contextual information about the video image itself (i.e. if there is a picnic in a park or someone's backyard.) In other words, the use of histograms and their intersection to classify frames into a given shot does not articulate or determine any contextual topic from the image files as required by limitation 1[e]. By contrast, Kritt's histogram simply determines the boundaries for a particular shot (sequence of frames) and does not provide context as to what is shown in the sequence of frames. According to the Petitioner, "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.) Pet. at 27. But the Petitioner misses the point. Determining shot boundaries based on the start and end of a similar sequence of image frames does not provide context about what is happening in an image or within a frame sequence. It simply identifies when a shot begins and ends.

Petitioner also contends that Kritt's identification of "scenes" is the same as "*determining a contextual topic from the image files.*" Kritt describes a scene as a segment of video/audio frames that convey a "common emotional content" or an "element of the story." Ex. 1003 at [0050]. According to the Petitioner, "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 they are interacting." Pet. at 28. However, Kritt's scenes—like Kritt's shot boundaries—concentrate on grouping together consecutive shots having the same visual and audio objects. Ex. 1003 at [0051]. Neither Kritt nor the Petitioner explains how Kritt's scenes (or sequence of frames with similar objects) provide any context regarding the image in the frame or what is happening in the scene based on the image. Nor do Kritt's "scenes" determine a "contextual topic" from the image. Instead, without any further analysis or support from Dr. Houh, Petitioner simply concludes that Kritt's "determination of a scene" is a contextual topic because it is "performed by analyzing frames—which are image files—and /or shots—which are a sequence of frames, so also image files." Pet. at 28. Like shots, Kritt's scenes are another method of parsing a video file and have nothing to do with extracting contextual information regarding the image files.

Petitioner does not rely on Fontana here, and neither Kritt's "shots" nor Kritt's "scenes" teach the "determining a contextual topic from the image files." Therefore, the combination of Kritt and Fontana fails to render obvious claim 1[e].

b. Kritt + Fontana Fails to Render Obvious "Assigning a Probability of Accuracy to the Identified Object Based on the Contextual Topic" as claimed in 1[f]

Limitation 1[f] recites "assigning a probability of accuracy to the identified object based on the contextual topic." As explained by the '954 Patent, using context or the arrangement of objects in a static image to aid recognition accuracy is difficult for multiple reasons. For example, the object could be partially blocked, lack faces and words, or the image could be overly cluttered. Ex. 1001 at 5:17-21. However, the '954 patent teaches that using:

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*" and that the 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. *Id.* at 5:22-34. (emphasis added).

In other words, by comparing the object and its contextual topic with prior and subsequent frames a probability that the object and topic were correctly identified is determined. For example, if "playing fetch" is the contextual topic, the

objects in the image are identified as a dog and a ball, and the surrounding frames also have the contextual topic and identify a dog and a ball, the probability that object is correctly identified as a “dog” and a “ball” is assigned. However, if a frame in the sequence identifies the contextual topic as “playing chase” and the objects as a “dog” and another “dog,” this discrepancy can be used to lower the probability of accuracy that the objects in the image are a “dog” and a “ball”. The key here is that the probability that the object has been correctly identified is based on the “contextual topic.” Because Kritt does not disclose contextual topics, Kritt does not and cannot render limitation 1[f] obvious.

Patent Owner explains in detail why Kritt’s “shots” and “scenes” are not contextual topics as required by limitation 1[e]. Because Kritt does not disclose “contextual topics,” Kritt predictably cannot disclose “assigning a probability of accuracy “based on the contextual topic.”

Nevertheless, Petitioner argues that Kritt’s visual tags “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.’” Pet. at 30 (citing Kritt at [0033].) This mapping fails. First, Petitioner never argues that “visual tags” are contextual topics. In fact, Kritt’s disclosure makes clear that the visual tag is the object identified in the image. Kritt states that “[t]he visual tag file 302 includes tags that

correspond with the visually perceived objects” whereas the audio tags “correspond with aurally perceived objects”) Ex. 1003 at [0032-33] (emphasis added). Kritt further explains that the object tags are compared to see if the tags relate to the same object and, if so, “a probability or confidence parameter” is assigned indicating whether the object was correctly identified. *Id.* Kritt compares an object to an object. There is no suggestion—let alone disclosure—in Kritt of using a contextual topic determined from an image file to assess whether the objects in that image were accurately identified.

Petitioner’s argument regarding “facial expressions” also fails. Even assuming *arguendo* that emotions are a contextual topic, neither Kritt nor Dr. Houh explain how the emotion is used to verify the accuracy of the objects (here, a human face) in the image. Instead, without support or explanation (and using circular logic), Petitioner concludes that a “probability of accuracy could be assigned to the underlying facial expression based on the probability of accuracy assigned to the emotion.” Pet. at 31 (stating that “a POSA 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 (i.e. confidence level) assigned to the emotion (the contextual topic)” (emphasis added)). Notably, Dr. Houh fails to explain why a POSA would find it

obvious. Instead, his declaration simply parrots the petition. *See* Ex. 1002 at 116 (only change in ¶116 compared to Pet at 31 is including text of Kritt's ¶ [0034] instead of citation.)

Petitioner does not rely on Fontana for this limitation and Kritt does not teach “assigning a probability of accuracy to the identified object based on the contextual topic.” Therefore, the combination of Kritt and Fontana fails to render obvious limitation 1[f].

c. Kritt + Fontana Fails to Render Obvious “converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic” as claimed in 1[g]

This limitation requires at least converting the data generated by the preceding steps—specifically, the object (1[d]), the probability (1[f]), and the contextual topic (1[e]) into video data. As explained above, Kritt cannot and does not disclose “contextual topics” or probabilities determined based on “contextual topics.” Not surprisingly, Petitioner also fails to show that Kritt teaches “converting the [video’s] image file” into video data that includes (1) objects, (2) contextual topics and (3) probabilities. Instead, Petitioner simply rehashes its flawed arguments regarding Kritt’s purported disclosure of (1) objects, (2) contextual topics and (3) probabilities. At no point does Petitioner or Dr. Houh argue—let alone describe—how Kritt converts its video image files into video data comprising the object, probability, and

contextual topic. The reason is simple. Kritt's focus is providing a quick timeline summary of what is happening in a video so viewers can visually understand the video and perhaps avoid bad language or possibly jump to their favorite song. Generating this static summary does not require Kritt to convert image files into video data including the object, probability, and contextual object. In contrast (and as explained below), the '954 patent's conversion of the video data enables it to provide real time, dynamic content output that can be used in a variety of applications—from web-crawling and video indexing to trademark monitoring and targeted advertising. Ex. 1001 at 13:5-45.

The Petitioner does not rely on Fontana for this limitation and Kritt fails to teach “converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic.” Therefore, the combination of Kritt and Fontana fails to render obvious claim 1[g].

d. Kritt + Fontana Fails to Render Obvious “cross-referencing the text and the video data with the video to determine the contextual topics” as claimed in 1[h]

This limitation requires cross-referencing the text (generated in 1[b]) with the video data's (1) object, (2) probability, (3) and contextual topic (generated in 1[g]) to determine contextual topics. Here, Petitioner's arguments are fatally deficient and reason alone for the Board to deny institution. The Petition and declaration

completely overlook and fail to argue how the Kritt teach cross-referencing of the text with the “probability” portion required in the video data (*see* 1[g], *supra*).

The '954 Patent explains that:

[A]udio transcript of a video can be used to derive certain context helpful in identifying and correcting or eliminating image false positives. In this way, an image anomaly or anomalies identified in a given video frame(s) are associated with time (time stamped) and correlated with a time range from the transcribed audio to establish certain probabilities of accuracy. Moreover, the aforementioned methodologies—establishing probabilities of accuracy of image identification from a set of frames and from the audio transcription—can be combined to improve the results.

Ex. 1001 at 5:53-63. The '954 Patent further explains that “topics generated from an image or a frame and the topics extracted from audio are cross-referenced and topics common to both the video data and audio text are given additional weights.” *Id.* at 7:41-48. For 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.” *Id.* at 5:8-12. 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. *Id.* In this way, the cross-

referencing of the video data and audio text ensure greater accuracy in determining the contextual topics.

Petitioner's mapping of Kritt's "emotions" fails to disclose or teach cross-referencing audio text with the objects, probability, and contextual topic determine from the image files. Instead, Petitioner contends that Kritt's cross-referencing text (key words) and video data (visual tags for the faces/human objects with the video (metadata, timestamps, and scenes) determines contextual topics (emotions). Pet. at 38-39. Absent from this analysis is any mention of the video data's probabilities. Nor does Petitioner or Dr. Houh explain how (or even if) Kritt's cross-referencing could or would result in more accurate contextual topics. Petitioner's mapping of "scenes" is just as flawed because it further ignores the "probabilities" required in the video data. Pet. at 40-41. At this point, Kritt's deficiencies have snowballed and there is no logical flow as to how Kritt's alleged "contextual topics" (i.e. Kritt's "shots/scenes") flow through the remaining claim limitations. Kritt does not teach nor suggest cross referencing text with video data (which by definition in 1[g] includes (1) the object, (2) the probability, and (3) the contextual topic) to more accurately determine the contextual topics for the video.

The Petitioner does not rely on Fontana for this limitation and Kritt fails to teach "cross-referencing the text and the video data with the video to determine the

contextual topics. Therefore, the combination of Kritt and Fontana fails to render obvious claim 1[h].

e. Kritt + Fontana Fails to Render Obvious “generating a contextual text, an image, or an animation based on the determined contextual topics” as claimed in 1[i]

The '954 patent provides multiple examples of generating a contextual text, an image or an animation based on the determined contextual topics. For example, the '954 Patent describes that:

At **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. For example, key words indicating topic and semantic that appear in both texts can be selected or emphasized. The output can also include metadata that can be time-stamped with frame references. The metadata can include the number of frames, the range of frames, and/or timestamp references.

Ex. 1001 at 7:44-53. The '954 patent further details that:

The server **220** then generates advertisement (text, images, or animation) based on the video text. In some embodiments, the server adds the advertisement to a specific symbol, image, frame, or a specific time in the video stream. The specific symbol, image, frame, or the specific time in the video stream can be selected based on the video text. The server **220** can add the audio text to the video stream in real time (i.e. real time close caption). The server **220** can generate video recommendation based on a database of the video text. In some

embodiments, the server **220** can search videos based on the video text (e.g., via a database of video text). In this fashion, video search can be optimized. Applications for the video search optimization may include search engine optimization (SEO), search engine marketing (SEM), censorship and removal materials of copyright violation.

Ex. 1001 at 8:58-9:8.

Because Petitioner fails to show that Krittr's cross-referencing of the text and video data (which requires the object, probability, and contextual topic) determined "contextual topics," Krittr also cannot generate a text, image, or animation based on the determined contextual topic. While Krittr discloses and depicts icons, emoticons, and keywords on its video summary screen, nothing in Krittr teaches using determined contextual topics (e.g. common contextual topics between the video data and text and verified by cross-referencing) to create these icons/keywords/emoticons. For example, Petitioner does not and cannot show that Krittr's emotional tag is a "determined contextual topic." Even assuming arguendo that the "emotion tag" is a contextual topic (1[e]) and ignoring the Petition's failure to address "probabilities", Krittr still does not teach comparing (cross-referencing) the video data (object = smiling face and contextual topic = happy scene) with the audio text to determine that the contextual topic (happy mood) is accurate and not an anomaly or false positive. Petitioner's argument regarding "visual tags" and "scenes" falls for the same reasons.

The Petitioner does not rely on Fontana for this limitation and Kritt fails to teach “generating a contextual text, an image, or an animation *based on the determined contextual topics*.” Therefore, the combination of Kritt and Fontana fails to render obvious claim 1[i].

f. Kritt + Fontana Fails to Render Obvious “generating a content-rich video based on the generated text, image, or animation” as claimed in 1[j]

Petitioner's mapping of 1[d] – 1[i] is fatally flawed for at least the reasons set forth above. And because each limitation in claim 1 builds on the prior limitations, those flaws trickle down through and to the last limitation of “generating a content-rich video based on the generated text, image, or animation.” Because Kritt does not teach generating a contextual text, an image, or an animation *based on the determined contextual topics*, it logically follows that Kritt does not and cannot generate “content-rich video based on the generated text, image, or animation.” Petitioner does not rely on Fontana here and Kritt fails to teach “generating a content-rich video based on the generated text, image, or animation” Therefore, for at least the reasons set forth in limitations 1[d]-1[i], the combination of Kritt and Fontana fails to render obvious limitation 1[j].

g. A Person of Ordinary Skill in the Art Would Not be Motivated to Combine Kritt and Fontana

Petitioner concedes that Kritt does not expressly disclose “*generating* audio and image files from the video” as required by claim 1[a]. Petition at 16. Petitioner therefore relies on Fontana to fill this gap. Petitioner points to Fontana’s teaching of “processing of multimedia content at an audio processing module 606 and a video 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 computers.” *Id.* Petitioner further argues that Fontana’s “thumbnail extraction module” generates thumbnails as image files and that a POSITA “would have appreciated that by extracting thumbnail images from the video, Fontana’s disclosure satisfies “generating ... image files from the video.” Pet. at 17.

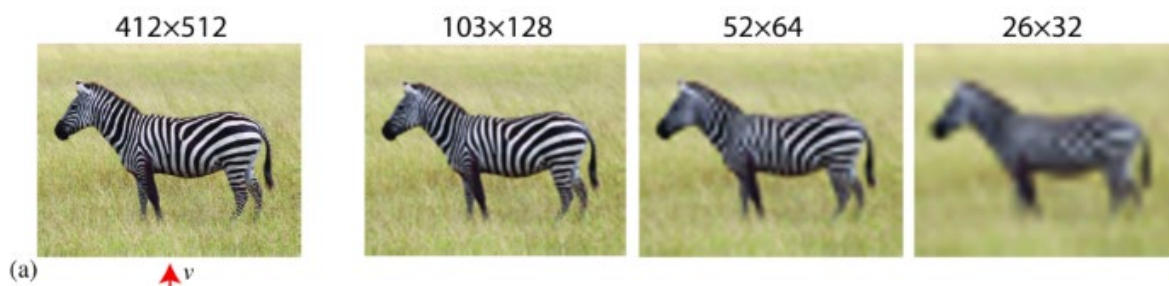
Petitioner’s motivation to combine is thin at best. Petitioner generally argues that the references are analogous and relate to generating multimedia content “including video, and . . . address how to search within such multimedia content.” Pet. at 18. Petitioner argues that Kritt’s key frame could be replaced by Fontana’s thumbnail and that Fontana provides the object recognition details missing in Kritt. Pet. at 19. Petitioner sums up its motivation to combine by concluding that “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. Pet. at 20.

Petitioner's combination fails for multiple reasons. First, a POSA would know that thumbnail images are not an appropriate choice for a key frame and nor for the object identification as described in Kritt. Thumbnails are typically utilized for quick parsing/viewing/browsing of video frames and are not utilized for object identification or segmentation given that most objects would appear at a much smaller scale within the thumbnail. As such, objects depicted in a thumbnail image appear at a much smaller scale than other images. For instance, a full resolution video frame showing a dog in the process of fetching a tennis ball would lose significant definition of the dog (potential object in the image) and likely most of the tennis ball (another potential object in the image) when converted to a thumbnail. This image degradation occurs because when a frame is subsampled from its original size to a thumbnail. The subsampling would also introduce other image quality deficiencies such aliasing artifacts (jagged edges, Moire patterns, color noise etc.), blurring, and loss of high frequency content when converting a full resolution video frame to a thumbnail size. Therefore, a POSA working with Kritt would utilize the video frames at their native resolution to perform object identification given the fact that the native resolution would provide maximum definition of the object details

rather than opt to subsample the video frame which would blur the image, introduce aliasing and incur loss of high frequency content in order to perform object identification using thumbnails.

Petitioner's motivation to substitute a low-resolution thumbnail for the key frame in Kritt makes no sense and would not result in improved object recognition within an image for at least the reasons discussed above. Additionally, Dr. Houh does attempt to explain why a thumbnail would provide improve object recognition, nor does he address how Fontana's alleged "improvements in object recognition" would cure the deficiencies of using Fontana's low-resolution image. *See* 1002, ¶¶ 55-58 and Ex. 2013, ¶¶ 55-58 (showing additions in blue are primarily cite parentheticals with the cited text itself and not additional support, explanation, or evidence from Dr. Houh.)



As shown above, images of a zebra depict how the image degrades when subsampled from 412x512 (native) down to 26x32 (thumbnail). A POSA interested in object identification would not use the 26x32 image but rather the 412x512 image to find and identify the zebra. As shown, the aliasing artifacts (the zebra's stripes are

distorted) in the zebra image 103x128 and even more pronounced in 52x64 image. This loss of image definition is significant and well below what is required of a key frame. Object recognition using a thumbnail would result in the very inaccuracies that the '954 Patent guards against. Dr. Houh makes no effort to address the well-known fact that thumbnails are lower resolution and are not generally used for object identification. Nor does he attempt to explain why doing so would be successful and not result in undue experimentation. In short, a POSA would not combine Kritt and Fontana with any hope of achieving the image files necessary to enable the contextual topic engine described and claimed in the '954 Patent.

h. Houh's Expert Testimony Is Entitled to Little or No Weight

Dr. Houh's declaration is fatally deficient and is a near carbon copy of the petition itself. Ex. 1002 at ¶¶71-139 In fact, a comparison of the Petition to the declaration shows that they are nearly identical. *See* Ex. 2013 at ¶¶71-199. Dr. Houh's expert testimony is conclusory and unsupported, adds little to the bare assertions for which it is offered to support, and is entitled to little or no weight. Upon a close review, the additions shown in blue in Ex. 2013 are not substantive at all. Instead, the majority simply replace citations with the quoted language itself. Because Dr. Houh's declaration merely repeats the conclusory arguments for which it is offered to support, his declaration is entitled to little or no weight. *See* 37 C.F.R.

§ 42.65(a) (“Expert testimony that does not disclose the underlying facts or data on which the opinion is based is entitled to little or no weight.”); *Upjohn Co. v. Mova Pharm. Corp.*, 225 F.3d 1306, 1311 (Fed. Cir. 2000) (“Lack of factual support for expert opinion to factual determinations, however, may render the testimony of little probative value in a validity determination.”); *Kinetic Techs., Inc., Petitioner, v. Skyworks Sols., Inc.*, IPR2014-00529, 2014 WL 4787238, at *10 (P.T.A.B. Sept. 23, 2014) (“Merely repeating an argument from the Petition in the declaration of a proposed expert does not give that argument enhanced probative value.”); *Xerox Corp. v. Bytemark, Inc.*, IPR2022-00624, 2022 WL 3648989, at *6 (P.T.A.B. Aug. 24, 2022) (“We have reviewed this excerpt from Dr. Jones’ declaration and note that it merely repeats, *verbatim*, the conclusory assertion for which it is offered to support. . . . Thus, the cited declaration testimony is conclusory and unsupported, adds little to the conclusory assertion for which it is offered to support, and is entitled to little weight.”)

Because claims 2-8 and 11-12 depend from claim 1, they fall for the same reasons as claim 1. In sum, Claims 1-8 and 11-12 are not rendered obvious by the combination of Kritt and Fontana.

2. Ground 2: The Combination of Kritt + Fontana + Lau Does Not Render Obvious Claims 2-3 and 9-10.

For Ground 2, Petitioner argues that certain dependent claims are obvious over Ground 1 in light of additional prior art to Lau. *See* Pet. at 56-64. Claims 2-3 and 9-10 depend from claim 1. Petitioners do argue that Lau fills any of the deficiencies in the combination of Kritt and Fontana with respect to independent claim 1. Thus, for at least the same reasons provided for Ground 1, the Petition should also be dismissed for Ground 2.

3. Ground 3: Petitioner Fails to Demonstrate that Claim 13 is Anticipated by Kritt

Here, Petitioner alleges that Kritt anticipates independent claim 13. Again, Petitioner is wrong.

Claim 13 recites:

A system for extracting data from a video, comprising:

13[a] a computer processor; and

13[b] a non-transitory computer readable medium containing

instructions directing the system to execute the steps of:

13[c] converting audio associated with the video to text;

13[d] converting images associated with the video to video data,

wherein the video data comprises identified objects;

13[e] extracting topics from the text and the video data;

13[f] assigning probabilities of accuracy for each of the identified objects, where the probabilities are based on the extracted topics;

13[g] cross-referencing the text, the video data, and the topics with the video to determine contextual topics;

13[h] generating a contextual text, an image, or an animation based on the determined contextual topics; and

13[i] generating content-rich video based on the generated contextual text, image, or animation.

Petitioner's mapping of Kritt to claim 13 largely points back to its mapping of Kritt to independent claim 1. Patent Owner incorporates below its reasons and arguments regarding claim 1 herein. The only limitation that Petitioner addresses with new arguments is claim 13[e]. Nevertheless, Petitioner's mapping of claim 13 fails for the reasons set forth below.

a. Kritt does not disclose “extracting topics from the text and video data” as required by limitation 13[e]

Nothing in Kritt teaches extracting a “topic” from the audio text and the video data.

First, Kritt's key words are just that, key words determined by “counting the frequency of occurrence” for those words. Pet. at 66-67. Just because a word appears

in an audio transcript multiple times does not mean that that key word provides a “topic” for the audio text, does not provide any indication of the context in which the key words occur, nor does it “extract topics” based on audio text and video data. Dr. Houh provides no additional support for this argument.

Second, Petitioner contends that Kritt's identification of “scenes” is the same as “*extracting* topics from the text and video data.” Pet. at 66. However, Kritt's scenes—like Kritt's shot boundaries—concentrate on grouping together consecutive shots having the same visual and audio objects. Ex. 1003 at [0051]. Neither Kritt nor the Petitioner explains how Kritt's scenes (or sequence of frames with similar objects) provide any topics regarding the video data, the audio in the frame, or what is happening in the scene based on the video data **and** audio. Nor do Kritt's “scenes” determine an “extracted topic” from the video data and the audio text as required by claim 13[e]. Instead, without any further analysis or support from Dr. Houh, Petitioner simply concludes that Kritt's “determination of a scene” is an extracted topic because it is “accompanied by a common audio segment.” Pet. at 66. Like shots, Kritt's scenes are another method of parsing a video file and has nothing to do with extracting contextual information regarding the audio text and video data. Therefore, Kritt fails to anticipate 13[e].

- b. Kritt does not disclose “assigning probabilities of accuracy for each of the identified objects, where the probabilities are based on the extracted topics” as required by limitation 13[f]**

Patent Owner fully incorporates its arguments regarding claim 1[f] herein.

- c. Kritt does not disclose “cross-referencing the text, the video data, and the topics with the video to determine contextual topics” as required by limitation 13[g]**

Patent Owner fully incorporates its arguments regarding claim 1[g] herein.

- d. Kritt does not disclose “generating a contextual text, an image, or an animation based on the determined contextual topics” as required by limitation 13[h]**

Patent Owner fully incorporates its arguments regarding claim 1[h] herein.

- e. Kritt does not disclose “generating content-rich video based on the generated contextual text, image, or animation” as required by limitation 13[i]**

Patent Owner fully incorporates its arguments regarding claim 1[i] herein.

In sum, Kritt does not anticipate independent claim 13.

VII. CONCLUSION

For the foregoing reasons, the Petition for Inter Partes review of any claim of the '954 patent should be denied.

Dated: September 12, 2025

Respectfully submitted,

/s/ Amy Simpson
Amy Simpson (Reg. No. 54,688)
Amy.simpson@hklaw.com
HOLLAND & KNIGHT LLP
1901 Avenue of the Stars, Suite 1200
Los Angeles, CA 90067

Case No. IPR 2025-00876
U.S. Patent No. 10,218,954 B2

Patent Owner's Preliminary Response

Telephone: (310) 201-8925

Facsimile: (310) 201-8922

Counsel for Cellular South, Inc.

CERTIFICATE OF SERVICE

Pursuant to 37 C.F.R. 42.6(e), the undersigned hereby certifies that true copies of the foregoing PATENT OWNER PRELIMINARY RESPONSE have been served in their entirety on this 12th day of September, 2025, by electronic mail on the attorneys of record for Petitioners:

Heidi L. Keefe (Reg. No. 40,673)
hkeefe@cooley.com
Andrew C. Mace (Reg. No. 63,342)
amace@cooley.com
Mark R. Weinstein
mweinstein@cooley.com
Reuben Chen
rchen@cooley.com
Alexandra D. Leeper
aleeper@cooley.com
COOLEY LLP
ATTN: Patent Group
1299 Pennsylvania Ave. NW, Suite 700
Washington, DC 20004
Tel: (650) 843-5001
Fax: (650) 849-7400

Dated: September 12, 2025

/s / Amy E. Simpson
Amy Simpson
Reg. No. 54,688

CERTIFICATE OF COMPLIANCE WITH 37 C.F.R. § 42.24

Pursuant to 37 C.F.R. § 42.24 et seq., the undersigned certifies that this Response complies with the type-volume limitations excluding the cover page and the parts exempted by 37 C.F.R. § 42.24(a)(1). This document contains 10,454 words as calculated by the “Word Count” feature of Microsoft Word, the word processing program used to create it.

Dated: September 12, 2025

/s/ Amy Simpson
Amy E. Simpson (Reg. No. 54,688)
Holland & Knight LLP
1901 Avenue of the Stars, Suite 1200
Los Angeles, California 90067
Email: amy.simpson@hklaw.com
Telephone: (310) 201-8925
Facsimile: (310) 201-8922