

Petition for *Inter Partes* Review of
U.S. Patent No. 11,126,853 B2

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner,

v.

CELLULAR SOUTH, INC.,
Patent Owner.

Case IPR2025-00877
Patent 11,126,853 B2
Issue Date: September 21, 2021

Title: VIDEO TO DATA

**PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 11,126,853 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	9
VI. THE CHALLENGED CLAIMS ARE UNPATENTABLE.....	9
A. Overview of Grounds.....	9
B. Prior Art Status of Relied-Upon References.....	11
C. Ground 1: Claims 1-4 Are Obvious Over Zhao in View of Kritt, Steinberg, and Kouzani.....	11
1. Independent Claim 1: “A system for generating data from a video, comprising:” (Claim 1[pre]).....	11

Table of Contents (continued)

	Page
(a) “a coordinator communicatively coupled to a splitter and to a plurality of demultiplexer nodes, wherein the splitter is configured to segment the video, wherein the demultiplexer nodes are configured to extract audio files from the video and to extract still frame images from the video;” (Claim 1[a]) (b) “an image detector configured to detect an image of an object in the still frame images, wherein the image detector is adjustable to increase detection of non-primary images in the video; and” (Claim 1[b])..... (c) “an object recognizer configured to compare the image of the object to a fractal, wherein the fractal includes a representation of the object based on landmarks associated with the object, wherein the recognizer is further configured to update the fractal with the image.” (Claim 1[c]).....	13 30 36
2. Claim 2: “The system of claim 1, wherein the coordinator is configured to increase the plurality of demultiplexer nodes when a threshold of processing capacity is reached.”	45
3. Claim 3: “The system of claim 1, wherein the recognizer is configured to determine distinguishing geometric features of the object.”	47
4. Claim 4: “The system of claim 3, wherein the distinguishing geometric features comprise a contour of eye sockets, a nose, and a chin.”	48
D. Ground 2: Claims 5-7 Are Obvious Over Ground 1 Prior Art in Further View of Yang and Romdhani	49
1. Claim 5: “The system of claim 4, wherein the recognizer is configured to determine skin textures of the object.”.....	49

Table of Contents
(continued)

	Page
2. Claim 6: “The system of claim 5, wherein the distinguishing geometric features or the skin textures are determined based on a three-dimensional model.”	55
3. Claim 7: “The system of claim 6, wherein the three-dimensional model is a three-dimensional morphable model.”	56
E. Ground 3: Claims 8-9 Are Obvious Over Ground 1 Prior Art in Further View of Trivedi	57
1. Claim 8: “The system of claim 1, further comprising a processor and a camera for capturing the video.”	57
2. Claim 9: “The system of claim 8, wherein the processor is configured to transform the video into a rectilinear format.”	61
F. Ground 4: Claims 10-11 Are Obvious Over Ground 1 Prior Art in Further View of Singer	64
1. Claim 10: “The system of claim 1, wherein the coordinator is configured to embed metadata about the object into the video, and wherein the metadata comprises a timestamp and a coordinate location of the object in the still frame images.”	64
2. Claim 11: “The system of claim 1, wherein the coordinator is configured to generate a metadata stream corresponding to the image, wherein the metadata stream includes one or more timestamps corresponding to the image, and wherein the coordinator is configured to embed the metadata stream in the video.”	71
VII. CONCLUSION	73
CERTIFICATE OF SERVICE.....	75

List of Exhibits

Exhibit No.	Description of Document
1001	U.S. Patent No. 11,126,853 B2 to Bartlett Wade Smith, IV (filed February 8, 2019, issued September 21, 2021) (“ 853 ” or “ 853 patent ”)
1002	Declaration of Henry Houh, Ph.D. (“ Houh ”)
1003	U.S. Patent App. Pub. No. 2009/0141940 A1 to Liang Zhao et al. (filed December 3, 2008, published June 4, 2009) (“ Zhao ”)
1004	U.S. Patent App. Pub. No. 2014/0181668 A1 to Barry A. Kritt et al. (filed Dec. 20, 2012, published Jun. 26, 2014) (“ Kritt ”)
1005	U.S. Patent No. 7,574,016 B2 to Eran Steinberg et al. (filed June 26, 2003, issued August 11, 2009) (“ Steinberg ”)
1006	Kouzani, A.Z. et al., “Fractal Face Representation and Recognition,” in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp.1609-1613 (1997) (“ Kouzani ”)
1007	Yang, Ming-Hsuan et al., “Detecting Faces in Images: A Survey,” in IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 1 (2002) (“ Yang ”)
1008	Romdhani, S. et al., “Face Identification by Fitting a 3D Morphable Model using Linear Shape and Texture Error Functions,” European Conf. on Computer Vision (2002) (“ Romdhani ”)
1009	U.S. Patent App. Pub. No. 2006/0187305 A1 to Mohan M. Trivedi et al. (PCT filed July 1, 2003, published August 24, 2006) (“ Trivedi ”)
1010	U.S. Patent App. Pub. No. 2011/0305394 A1 to David William Singer et al. (filed June 15, 2010, published December 15, 2011) (“ Singer ”)
1011	Excerpts from Peter Lubbers et al., <i>Pro HTML5 Programming: Powerful APIs for Richer Internet Application Development</i> (2010)
1012	Excerpts from Microsoft Computer Dictionary (5th ed. 2002)
1013	Excerpts from Jura van Vliet et al., <i>Programming Amazon EC2</i> (2011)

Petition for *Inter Partes* Review of
U.S. Patent No. 11,126,853 B2

List of Exhibits

Exhibit No.	Description of Document
1014	Xiutao Tang et al., “Facial Image Recognition Based on Fractal Image Encoding,” in Bell Labs Technical Journal 15(1) (2010)
1015	Prosecution History of U.S. Patent No. 11,126,853 B2
1016	U.S. Patent No. 9,697,230 to Henry Houh et al. (filed Mar. 31, 2006, issued Jul. 4, 2017)
1017	Cellular South’s Opposition to Motion to Dismiss (Document 28) in No. 6:24-cv-00245-DAE (W.D. Tex.)
1018	Declaration of Ingrid Hsieh-Yee, Ph.D. (“ Hsieh-Yee ”)
1019	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 ’853 patent is 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. (EX1019, 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 '853 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 11 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 '853 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	1-4	Zhao in view of Kritt, Steinberg, and Kouzani
2	5-7	Ground 1 prior art in further view of Yang and Romdhani
3	8-9	Ground 1 prior art in further view of Trivedi
4	10-11	Ground 1 prior art in further view of Singer

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

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

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

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.

§314(a): The *General Plastic* factors are not relevant; this is the first and only IPR petition filed by Petitioner with respect to the '853 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 any of the references relied upon in **Ground 1** or **Ground 2** because none were presented during prosecution.

With respect to Trivedi and Singer, cited in **Ground 3** and **Ground 4**, respectively, Petitioner cites them for the same claim limitations in dependent claims 8-11 for which the Examiner cited them during prosecution, and which was not distinguished or disputed by the applicant. (See EX1015, pp.00145 (“Trivedi further teaches further comprising a processor and a camera for capturing the video”; “Trivedi further teaches wherein the processor is configured to transform the video into rectilinear format”), 00147-00148 (“Singer...teaches wherein the coordinator is configured to embed metadata about the object into the video, and wherein the

metadata comprises a timestamp and a coordinate location of the object in the still frame images”; “Singer further teaches wherein the coordinator is configured to generate a metadata stream corresponding to the image, wherein the metadata stream includes one or more timestamps corresponding to the image, and wherein the coordinator is configured to embed the metadata stream in the video”), 00167-00169 (Applicant Reply to Office Action of October 29, 2020).)

Petitioner reserves the right to address and respond to any assertions that Patent Owner may raise regarding discretionary factors.

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 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. (Houh, ¶¶18-23.)

B. Specification Overview

The ’853 patent describes its purpose at a very high level. It states that it

Petition for *Inter Partes* Review of
U.S. Patent No. 11,126,853 B2

“relates to a method and a system for generating various and useful data from source media, such as videos and other digital content.” (’853, 1:12-14.) 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.” (’853, 1:34-36.)

The ’853 patent purports to offer a solution to the common problem of how to describe visual information from audio-visual content. (Houh, ¶41.) It is in many ways similar to a museum curator providing text guides to an artist’s work, or a movie critic explaining the cinematography of a film. The curator or critic may distinguish images in the foreground from those in the background, and explain visual techniques such as object placement, shading, and texturing that make the artwork or scene visually effective. The curator or critic can provide commentary that tells the viewer what to look for, where to look, and (in a film) when to look.

Audio-visual data such as videos long predates the ’853 patent, and prior artists similarly had long been working on various ways to analyze and describe video content. (Houh, ¶41.) By the time of the alleged priority date (June 29, 2016), the fields of audio and image analysis for video content were well-developed. Indeed, the Background section of the ’853 patent acknowledges known image searching techniques to identify matching and similar images, as well as audio-to-text algorithms for transcribing text from audio. (’853, 1:21-29.)

The specification of the '853 patent provides primarily a series of high-level, functional explanations for how to implement the alleged invention, for which it relies on well-known techniques to implement. (*See, e.g.*, '853, 14:30-15:47 (describing various well-known “recognition models” for image detection).) For example, Figure 10 from the '853 patent, shown below, “illustrates exemplary system architecture with an exemplary process flow.”

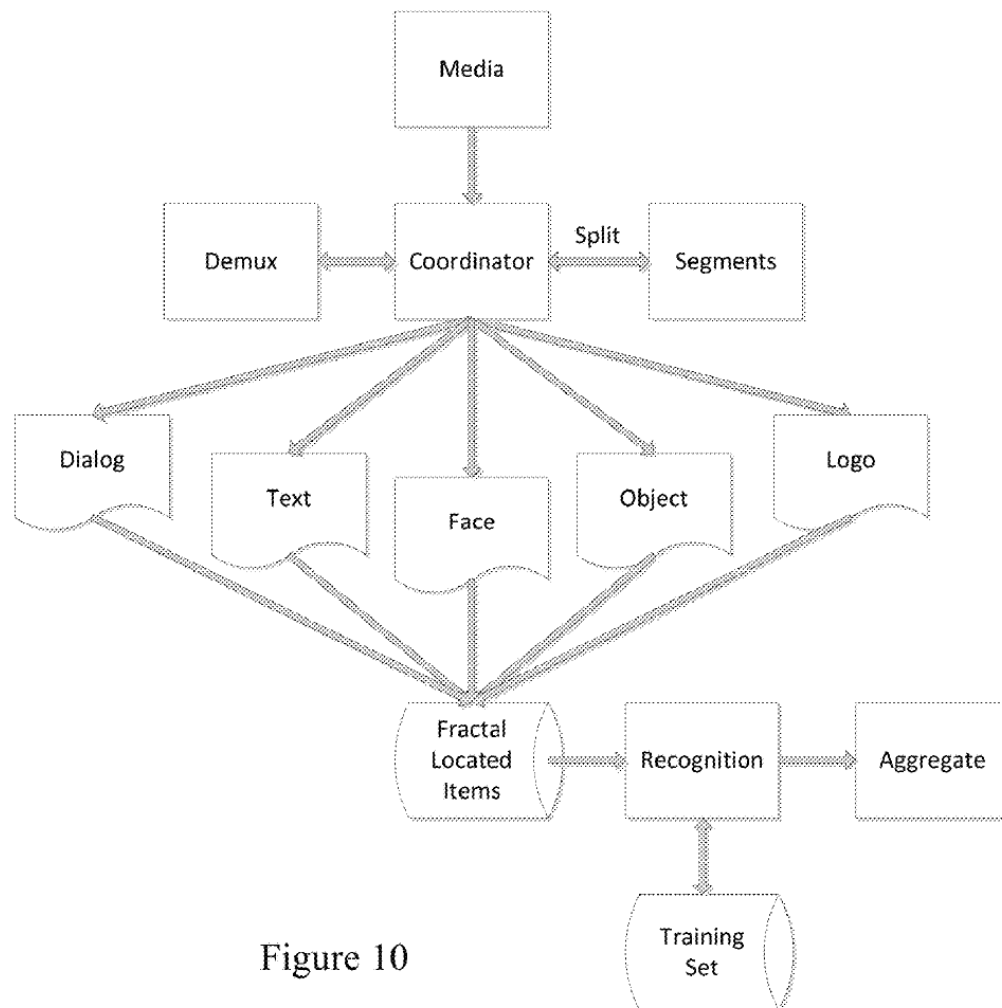


Figure 10

('853, Fig. 10, 3:50-51.) Referring to Figure 10, the '853 patent states that “[s]ource media can be provided to a coordinator.” ('853, 16:25-28.) The specification

Petition for *Inter Partes* Review of
U.S. Patent No. 11,126,853 B2

describes the coordinator functionally. The coordinator, among other things, “can be responsible for identifying new job requests and managing the routing of request input and result data output in the system.” (’853, 16:31-33.) “[T]he coordinator can send uploaded source media to a splitter and a demuxer/demultiplexer.” (’853, 16:60-63.) Like the coordinator, the splitter and demuxer/demultiplexer are described functionally. Then, data from the source media, such as faces, text, and dialog, can be extracted. (’853, 16:63-67.) “[T]he extracted data can be sent through a filter to compare fractal located items...for recognition.” (’853, 17:4-6.) The ’853 patent does not purport to have invented fractals; rather, the use of fractals for image recognition is one of the known recognition techniques identified in the ’853 specification. (’853, 14:30-15:47.) Based on the results of the fractal filtering, the fractal can be refined and the fractal training set updated. (’853, 17:6-10.)

The ’853 patent further states that, “[o]ften images are segmented only to locate and positively identify one or very few main images in the foreground of a given frame,” and “[t]he non-primary or background images are often treated as noise.” (’853, 9:8-12.) “Nevertheless,” according to the ’853 patent, “these can provide useful information, context, and/or branding for example,” and “it may become necessary or desirable to detect more detail from a frame or set of frames.” (’853, 9:12-23.) “In such circumstances, the computational thresholds for identification of an object, face, etc. can be altered according to a then stated need

or desire for non-primary, background, obstructed and/or grainy type images. Such image identification threshold adjustment capability can be implemented, for example, as a user-controlled interface, dial, slider, or button, which enables the user to make adjustments to suit specific needs or preferences.” (’853, 9:23-30.)

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 Zhao (**EX1003**), a strong reference that alone discloses or suggests the primary features of the ’853 patent. Like the ’853 patent, Zhao discloses a system that performs detection and recognition of objects (such as people and faces) in video images to generate data, including by creating and updating object models used for detection and recognition, and by generating data about the video that can be used for indexing and retrieval. (Houh, ¶52 (citing Zhao, ¶¶0051, 0053, 0055).)

Ground 1 also cites Kritt (**EX1004**), Steinberg (**EX1005**), and Kouzani (**EX1006**) to provide express confirmation of well-known features either already

suggested in Zhao or that would have been apparent and obvious. Kritt is cited primarily for its teachings related to the well-known concept of distributed/multi processing in connection with the analysis of claim 1[a] and the recited “**coordinator,**” “**splitter,**” and “**demultiplexer nodes.**” (Houh, ¶¶55-60.) Steinberg is cited primarily in connection with claim 1[b] and its recitation of an image detector that is “**adjustable to increase detection of non-primary images in the video.**” (Houh, ¶¶61-64.) And Kouzani, which confirms that “**fractal**”-based image recognition was well-known to a person of ordinary skill, is cited primarily for claim 1[c] and its recitation of comparing an image of the object to a “**fractal**” and updating “**the fractal**” with the image. (Houh, ¶¶65-70.)

Ground 2 further cites Yang (EX1007) and Romdhani (EX1008) in connection with dependent claims 5-7 and their limitations related to “**skin textures**” and a “**three-dimensional model.**” (Houh, ¶¶71-81.)

Ground 3 cites Trivedi (EX1009) in connection with dependent claims 8 and 9, and Ground 4 cites Singer (EX1010) in connection with dependent claims 10 and 11. (Houh, ¶¶82-92.) As noted in Part III.C above, Petitioner cites them for the same claim limitations for which the Examiner cited them during prosecution, and which was not distinguished or disputed by the applicant.

B. Prior Art Status of Relied-Upon References

Because the '853 patent claims priority to an earliest application filed June 29, 2016, AIA law applies to the challenged claims. All of the references cited in the Grounds qualify as prior art under §102(a)(1) because they were published before June 29, 2016. (EX1018 (Hsieh-Yee), ¶¶26-53 (Kouzani), 54-75 (Yang), 76-93 (Romdhani).)

C. Ground 1: Claims 1-4 Are Obvious Over Zhao in View of Kritt, Steinberg, and Kouzani

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

Assuming the preamble provides a claim limitation, Zhao discloses it. Zhao discloses a system that performs detection and recognition of objects, such as people and faces, in video images to generate data, including by creating and updating object models used for detection and recognition, and by generating data about the video that can be used for indexing and retrieval. (Houh, ¶¶93-94.) Figure 2 of Zhao, discussed in more detail in the analysis below, illustrates at a high level this overall process of generating data from a video performed by the system:

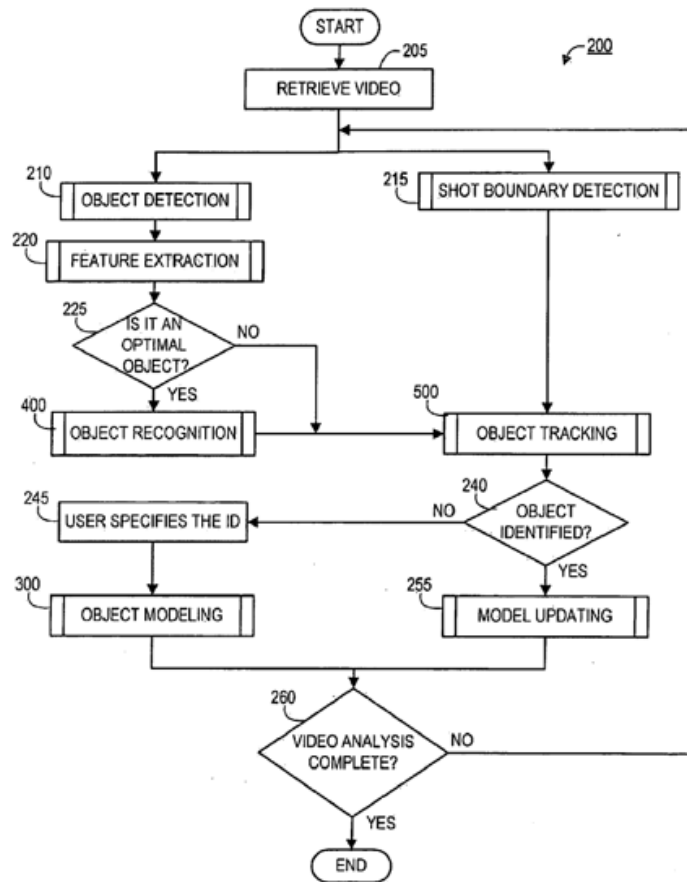


FIG. 2
OVERALL PROCESS

(Zhao, Fig. 2; *see also id.*, ¶0036 (“FIG. 2 illustrates a flowchart describing the steps involved, from a high-level, in one embodiment of the present system for detecting, modeling, recognizing, and tracking object images throughout one or more videos.”).)

Zhao explains, for example, that it “generally relate[s] to systems and methods for detection, modeling, recognition, and tracking of objects within video content” and to “indexing and retrieval systems for videos based on generated object models.” (Zhao, ¶0051.) “These objects include people, faces, articles of clothing, plants,

animals, machinery, electronic equipment, food, and virtually any other type of image that can be captured or presented in video.” (Zhao, ¶0051.) Zhao’s system “may be operated in a computer environment” and “any results or outputs relating to detection, modeling, recognition, indexing, and/or tracking of object images within videos may be stored in a database” or otherwise output. (Zhao, ¶0055.) Zhao states that its system is “useful for a wide variety of applications, including video indexing and retrieval, video surveillance and security, unknown person identification, advertising, and many other fields.” (Zhao, ¶0053.)

Therefore, Zhao discloses “**a system for generating data from a video.**” And as explained below, the system as claimed is obvious over Zhao, Kritt, Steinberg, and Kouzani.

- (a) **“a coordinator communicatively coupled to a splitter and to a plurality of demultiplexer nodes, wherein the splitter is configured to segment the video, wherein the demultiplexer nodes are configured to extract audio files from the video and to extract still frame images from the video;” (Claim 1[a])**

As explained below, claim 1[a] is disclosed by and obvious over Zhao in view of Kritt. (Houh, ¶¶96-113.)

For context, the ’853 patent describes a “**coordinator**,” “**splitter**,” and “**demultiplexer**” in terms of functions they may perform, and does not state any particular form or structure. (Houh, ¶97.) The ’853 patent states, for instance, that “[s]ource media can be provided to a coordinator, which can be a non-visual

component that can perform one or more of several functions.” (’853, 16:26-27.)

For example, “[t]he coordinator can direct distributed processing and aggregation.”

(’853, 18:16-17.) Referring to figure 10, reproduced below, the ’853 patent states

that “the coordinator can send uploaded source media to a splitter and a demuxer/demultiplexer”:

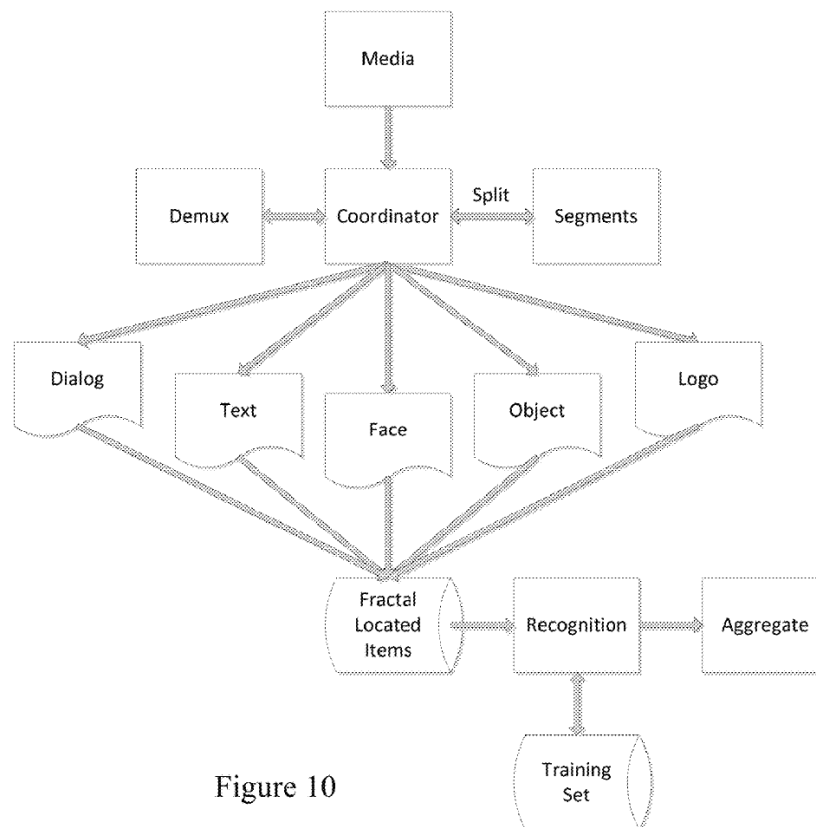


Figure 10

(’853, Fig. 10, 16:60-63; *see also id.*, 3:50-51.) A “**splitter**” can be invoked to

“‘slice’ [media] assets into media segments and/or multiple sub assets comprising

the entire stream.” (’853, 17:32-37.) A “**demultiplexer**” can “extract audio files

from the video and/or to extract still frame images from the video.” (’853, 1:45-47.)

For example, “a demultiplexer component can strip audio data from segments and store audio streams for future analysis.” (’853, 17:61-62.)

A person of ordinary skill would recognize that the arrangement of a “**coordinator**,” “**splitter**,” and “**plurality of demultiplexer nodes**” recited in claim 1[a] is merely an application of the well-known concept of distributed processing (or multiprocessing) to video processing, to obtain the obvious benefits that processing time can be reduced, and more processing overall can be performed, by sharing the workload and performing multiple operations in a parallel fashion. (Houh, ¶98.) The ’853 patent likewise explains that “the splitter can be configured to segment the video,” then later states, for example, that “[t]he video-to-data engine can segment the video into chunks for distributed, or parallel, processing....Distributed processing in this context can mean that the processing time for analyzing a video from beginning to end is a fraction of the play time of the video. This can be accomplished by breaking the processes into sections and processing them simultaneously.” (’853, 2:35-36, 6:15-21.) By 2016, distributed/multi processing was well-known and neither novel nor non-obvious. (Houh, ¶98 (citing EX1012, pp.005 (entry for “distributed processing”), 006 (entry for “multiprocessing”)).) And as explained below, claim 1[a] is obvious over Zhao in view of Kritt.

Turning to Zhao, Petitioner will first explain how Zhao teaches both “**splitter**” operations and “**demultiplexer**” operations. Petitioner will then explain how Zhao discloses and renders obvious a “**coordinator**” before turning to the combination with Kritt.

Zhao itself discloses a “**splitter**” “**configured to segment the video**” in the form of processes that perform “shot/scene boundary detection” to split a video into shots or scenes that are extracted for processing. (Houh, ¶100.) Zhao explains:

Once the video has been retrieved, the video undergoes a process of detection of objects and extraction of object features from video images (processes **210**, **220**). Concurrently, shots or scenes are detected in the video via a shot/scene boundary detection procedure (process **215**). The shots or scenes of the video are extracted to provide a sequence of images for analysis by the system, and each detected shot or scene is individually analyzed. Generally, the term “shot” or “scene” refers to a grouping of frames or images recorded by either a stationary or smoothly-moving camera, with little or no background change between the frames, corresponding to one continuous time period.

(Zhao, ¶0057.) Figure 2 of Zhao illustrates this process with shot boundary detection process **215**:

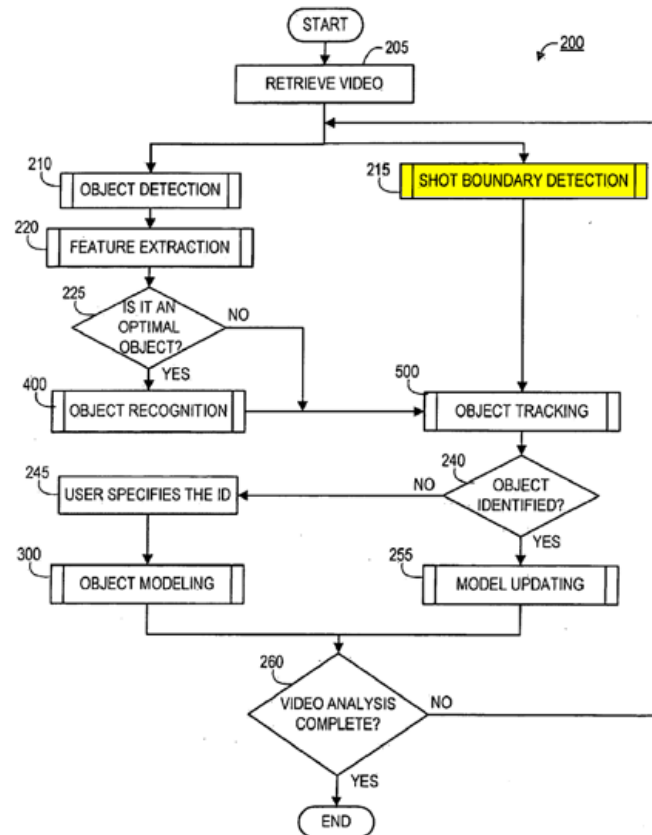


FIG. 2
OVERALL PROCESS

(Zhao, Fig. 2 (highlighted added).)

With respect to the “**demultiplexer**,” Zhao also discloses at least some of the claimed “**demultiplexer**” functionality—i.e., “**extract[ing] still frame images from the video.**” (Houh, ¶101.) Zhao describes “receiving a video file, wherein the video file comprises a plurality of video frames.” (Zhao, ¶0023.) Zhao teaches that the object detection and recognition processes involve extracting still frame images from the video. (Houh, ¶101.) For example, Zhao explains that the object

recognition process is performed on an “image or frame extracted from a video.”²

(Zhao, ¶0075 (“The process **400** shown in FIG. 4 is for one image or frame extracted from a video, but as will be appreciated, this process is typically repeated for each optimal object image in the video.”).) Zhao similarly explains that the object detection process involves “detection of objects and extraction of object features from video images (processes **210, 220**).” (Zhao, ¶0057; *see also id.*, ¶0059 (“Once an object has been detected within a video frame...”).)

With respect to “**demultiplexer**” functionality to “**extract audio files from the video**,” Petitioner notes that the claims of the ’853 patent do not recite any additional limitations related to “audio” or the extracted “audio files.” (Houh, ¶102.) Zhao likewise does not discuss audio information, which is not surprising because audio processing is not a focus of Zhao’s purported invention. (*Id.*) But it would have been obvious for Zhao’s system to further include processes for “**extracting audio files from the video**.” (*Id.*) For example, Zhao discloses performing its techniques on videos such as episodes of television shows such as the Gilmore Girls (Zhao, e.g., ¶¶0040-0041, 0093, Fig. 6), which a person of ordinary skill would have understood and found obvious to have included one or more audio tracks. (Houh, ¶102.) Zhao more generally teaches that its techniques are “useful for a wide variety

² All underlining added to quoted material unless noted.

of applications, including video indexing and retrieval, video surveillance and security, unknown person identification, advertising, and many other fields.” (Zhao, ¶0053.) Given Zhao’s broad applicability, a person of ordinary skill would have readily appreciated and found it obvious that videos processed by Zhao’s system would include audio tracks. (Houh, ¶102.) A person of ordinary skill would have found it obvious to extract audio tracks from the video into separate audio files—for example, to separately store or analyze—but again, those sorts of preparatory operations are not a focus of Zhao. (*Id.*) For example, a person of ordinary skill would have understood and found it obvious to perform operations such as speech recognition or speaker recognition on the audio data. (*Id.*) In a security and surveillance context (and numerous other contexts), for instance, these operations could be used to further confirm a person’s identity or to make a record of conversations associated with the captured video. (*Id.*) For example, **Kritt** describes using audio to “validate that the human object was correctly identified” in a video. (*Id.* (quoting Kritt, ¶0033).)

A step of extracting audio tracks from a video file in Zhao would have resulted in “**extract[ing] audio files**” in at least two separate ways. (Houh, ¶103.) First, it was well-known to a person of ordinary skill that video file formats took the form of “container” files (or “wrapper” files) that in turn stored video tracks, audio tracks, and other data as files within that container—much like how a single “.zip” archive

file can store multiple files. (*Id.*) For example, Kritt explains in its Background section:

A television show, movie, internet video, or other similar content may be stored on a disc or in other memory using a container or wrapper file format. The container format may be used to specify how multiple different data files are to be used. The container format for a video may identify different data types and describe how they are to be interleaved when the video is played. A container may contain video files, audio files, subtitle files, chapter-information files, metadata, and other files. A container also typically includes a file that specifies synchronization information needed for simultaneous playback of the various files.

(Kritt, ¶0002.) Similarly, the 2010 book *Pro HTML5 Programming* explains that “[a]n audio or video file is really just a *container* file, similar to a ZIP archive file that contains a number of files.” (EX1011, p.0013.) Figure 3-1 from the book, reproduced below, “shows how a video file (a video container) contains audio tracks, video tracks, and additional metadata. The audio and video tracks are combined at runtime to play the video.” (EX1011, p.0013.)

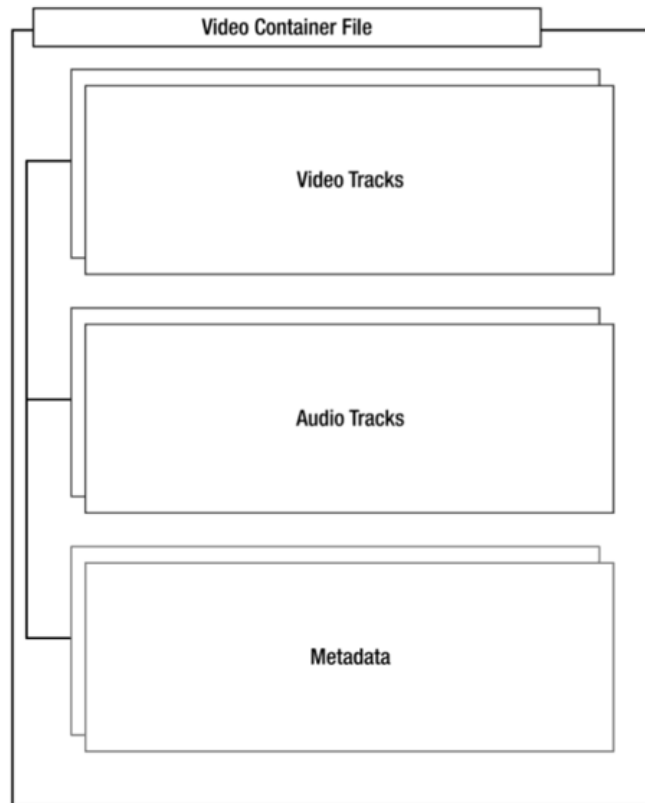


Figure 3-1. Overview of the video container

(EX1011, p.0014.) Therefore, a person of ordinary skill would have understood that, by extracting audio tracks from a video file, a step of “**extract[ing] audio files from the video**” would be performed. (Houh, ¶103.)

Second, by extracting audio from the video for separate analysis or storage, a person of ordinary skill would have understood and found it obvious that the resulting extracted audio would take the form of “files.” (Houh, ¶104.) This is because, for starters, file storage was one of a finite number of ways of storing data. (*Id.*) Moreover, it was routine to use files—known as “scratch files” or “temporary files”—for at least temporary storage while the associated data was processed. (*Id.*

(citing EX1012, pp.007 (entries for “scratch” and “scratch file”), 008 (entry for “temporary file”)).)

Turning back to the claimed “**coordinator**,” a person of ordinary skill would have further understood that, in order to manage the various video file retrieval, splitter, and demultiplexer operations, Zhao’s system would also include “**coordinator**” functionality. (Houh, ¶105.) For example, referring to figure 2, Zhao shows and explains how **splitter operations** (e.g., process **215**) and **demultiplexer operations** (e.g., processes **210**, **400**) take place concurrently:

Once the video has been retrieved, the video undergoes a process of detection of objects and extraction of object features from video images (processes **210**, **220**). Concurrently, shots or scenes are detected in the video via a shot/scene boundary detection procedure (process **215**).

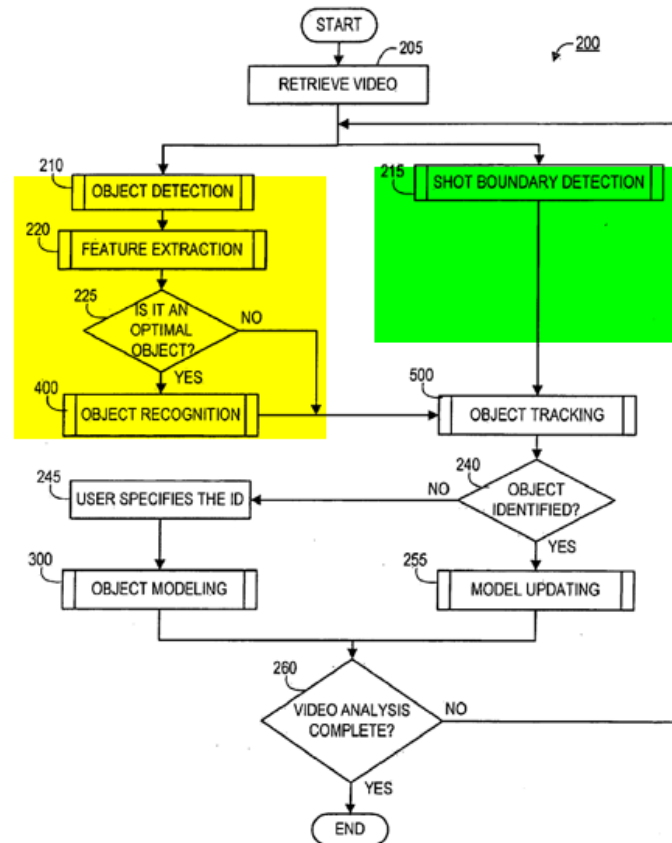


FIG. 2
OVERALL PROCESS

(Zhao, Fig. 2 (highlighting added), ¶0057.) A person of ordinary skill would have understood that various operations would need to be performed in order to retrieve a video file and then distribute the file for concurrent processing by processes **215** and **210**, for example. (Houh, ¶105.) Those operations, highlighted in figure 2 of Zhao below, would form at least part of a “**coordinator**.”

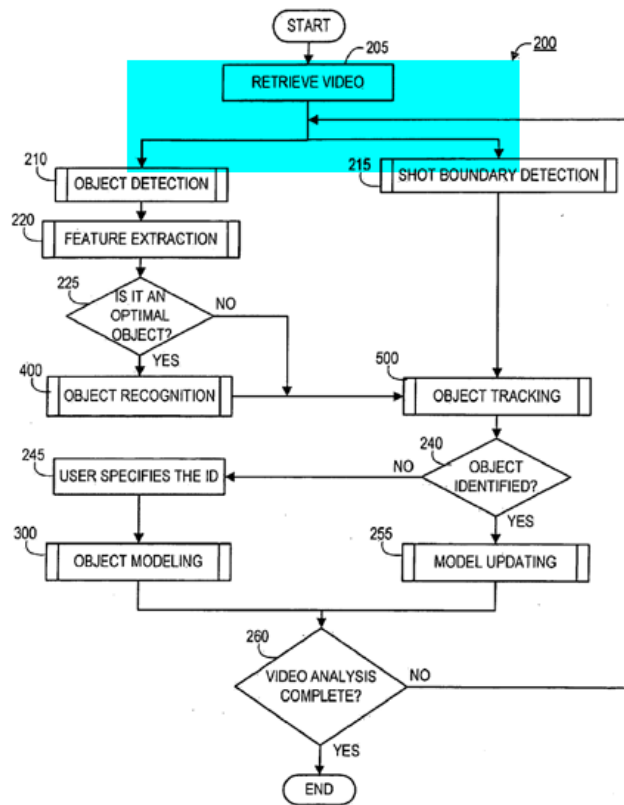


FIG. 2
OVERALL PROCESS

(Zhao, Fig. 2 (highlighting added); Houh, ¶105.) A person of ordinary skill would have further understood that, for Zhao’s “coordinator” functionality to distribute the video file that it retrieved, it would need to be “**communicatively coupled**” to the system’s “splitter” and “demultiplexer” functionality in order to notify those components in some manner. (Houh, ¶105.) In fact, Zhao’s disclosure of “concurrent” processing (Zhao, ¶0057) suggests that the video file is actually distributed to separate nodes (such as separate processors or computer systems) with which Zhao’s “coordinator” would be communicatively coupled. (Houh, ¶105.) This is just like “coordinator” functionality described in the ’853 patent, which

describes receiving a media file and distributing it to a splitter and demultiplexer. (*Id.* (citing '853, 16:60-63 (“the coordinator can send uploaded source media to a splitter and a demuxer/demultiplexer”))).)

Lastly, before turning to the combination with Kritt, with respect to the claimed “**plurality of demultiplexer nodes**,” this too is suggested by Zhao. (Houh, ¶106.) Although Zhao focuses on describing its system with respect to a single video file, Zhao explains that “the process may be extrapolated to analyze a plurality of videos.” (Zhao, ¶0057.) Because as noted above, Zhao also describes “concurrent” processing, a person of ordinary skill would have therefore understood and found it obvious for Zhao’s system, in an implementation for processing a higher volume of video files, to comprise a “**plurality**” of nodes providing the claimed demultiplexer functionality. (Houh, ¶106.) As discussed above, the benefits of distributed computing were well known and therefore would have been obvious to try. (*Id.* (citing EX1012, pp.005 (entry for “distributed processing”), 006 (entry for “multiprocessing”))).)

Turning to the combination with **Kritt**, like Zhao, Kritt describes a system for processing both visual and audio data of videos to generate data about the videos. (Houh, ¶107 (citing Kritt, ¶¶0017 (“a viewer selects a video, and human, nonhuman, and audio objects of the video are identified”), 0043 (“visual objects in a shot may be identified by application of one or more known image recognition processes to

the shot”), 0059 (“Segments of the video for which the audio feature values are similar to those that are characteristic of speech may be classified as speech.”), 0065 (“In operation **612**, an audio transcript may be determined. An audio transcript may include all of the words spoken in the video....[S]poken words may be determined from audio features classified as speech using any known technique”)).)

With respect to claim 1[a] specifically, and concepts of distributed/multi processing, Kritt describes modules for processing video and audio of a video file, and that those modules can be distributed on different computer systems:

The memory **104** may store all or a portion of the following: an audio visual file container **150** (shown in FIG. 2 as container **202**), a video processing module **152**, an audio processing module **154**, and a control module **156**. These modules are illustrated as being included within the memory **104** in the computer system **100**, however, in other embodiments, some or all of them may be on different computer systems and may be accessed remotely, e.g., via a network.

(Kritt, ¶0022; *see also id.*, ¶0024 (“The video processing module **152** may include various processes that generate visual tags according to one embodiment. The audio processing module **154** may include various processes for generating audio and key word tags according to one embodiment.”); Houh, ¶108.)

Kritt further describes how its systems could be provided using “cloud-computing infrastructure.” (Kritt, ¶¶0078-0079.) Kritt explains, as was known to a

person of ordinary skill, that “[c]loud computing generally refers to the provision of scalable computing resources as a service over a network.” (Kritt, ¶0078; Houh, ¶109.) “More formally,” Kritt explains, “cloud computing may be defined as a computing capability that provides an abstraction between the computing resource and its underlying technical architecture (e.g., servers, storage, networks), enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.” (Kritt, ¶0078.) Kritt further explains that one benefit of cloud computing is that “[t]ypically, cloud-computing resources are provided to a user on a pay-per-use basis, where users are charged only for the computing resources actually used (e.g., an amount of storage space used by a user or a number of virtualized systems instantiated by the user).” (Kritt, ¶0079.) It also allows a user to “access any of the resources that reside in the cloud at any time, and from anywhere across the Internet.” (*Id.*) As discussed above for Zhao, a person of ordinary skill would have recognized that in such a distributed/multi processing system, “**coordinator**” functionality would be needed to manage the operations and workloads among the various video and audio processing nodes. (Houh, ¶109.)

Rationale and Motivation to Combine (Zhao with Kritt). It would have been obvious to a person of ordinary skill to incorporate Kritt’s teachings regarding using a distributed/multi processing system (such as a pool of nodes in a cloud-

computing system) for processing images and audio into Zhao’s video processing system. (Houh, ¶110.) The combination would have involved the conventional application of distributed processing techniques to Zhao’s system and would have predictably resulted in Zhao’s system in which multiple nodes were used for processing images and audio of a retrieved video file, along with functionality that managed that processing. (*Id.*) The combination therefore would have resulted in “**a coordinator**”—i.e., functionality for managing the overall distributed process of analyzing images and audio of a video file—“**communicatively coupled to a splitter**”—i.e., functionality to split a video into shots/segments—“**and to a plurality of demultiplexer nodes**”—e.g., multiple nodes in a cloud computing infrastructure for performing image and audio recognition. (*Id.*) The combination therefore would have resulted in a “**splitter [] configured to segment the video,**” and “**demultiplexer nodes [] configured to extract audio files from the video and to extract still frame images from the video.**” (*Id.*)

Zhao and Kritt are analogous references to the ’853 patent, being both in the same field of video processing and generating data from video content, and also reasonably pertinent to problems facing the ’853 inventors. (Houh, ¶111 (citing ’853, 1:34-36; Zhao, ¶0036, Fig. 2; Kritt, ¶0017).)

A person of ordinary skill would have been motivated to combine. (Houh, ¶112.) As noted above, although Zhao focuses on describing its system with respect

to a single video file, Zhao explains that “the process may be extrapolated to analyze a plurality of videos.” (Zhao, ¶0057.) Zhao also teaches that its techniques are “useful for a wide variety of applications, including video indexing and retrieval, video surveillance and security, unknown person identification, advertising, and many other fields.” (Zhao, ¶0053.) A person of ordinary skill would have therefore recognized that an application of Zhao’s system could involve high-volume video processing, along with potential time constraints—for example, a large-scale video indexing and retrieval website, or a commercial video surveillance and security service. (Houh, ¶112.) A person of ordinary skill would have therefore looked to an analogous reference such as Kritt in the same field as Zhao that provides further teachings regarding system architectures capable of providing such a larger-scale service. (*Id.*) Kritt also provides express motivation to combine, describing numerous advantages of a cloud-based system, including “convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction,” “resources [] provided to a user on a pay-per-use basis,” and “access any of the resources that reside in the cloud at any time, and from anywhere across the Internet.” (Kritt, ¶¶0078-0079; Houh, ¶112.) Moreover, with respect to audio processing, as explained above, Kritt also explains integrating audio and image processing would beneficially enable confirmation of recognized visual

objects in the video, as well as a transcript that could be used for further analysis of the video. (Kritt, e.g., ¶¶0033, 0065; Houh, ¶112.)

A person of ordinary skill would have had a reasonable expectation that the combination would have been successful. (Houh, ¶113.) It would have involved the straightforward application of conventional techniques for distributed processing (and audio processing). (*Id.*) As Kritt notes, cloud computing for example was already well-established, as were techniques for speech and speaker recognition.

(b) **“an image detector configured to detect an image of an object in the still frame images, wherein the image detector is adjustable to increase detection of non-primary images in the video; and” (Claim 1[b])**

As explained below, claim 1[b] is obvious over Zhao in view of Steinberg. (Houh, ¶¶114-124.)

The '853 patent describes image “detection” and “recognition” as a two-step process, first “detection,” then “recognition.” ('853, 13:39-46.) Zhao similarly discloses separate detection and recognition steps. (Houh, ¶¶115-116.)

With respect to image detection, Zhao discloses **“an image detector configured to detect an image of an object in the still frame images”**—such as processes to detect images of objects such as faces in extracted still frame images of a video. (Houh, ¶116.) Referring to Figure 2, reproduced below, Zhao explains that “[o]nce the video has been retrieved, the video undergoes a process of detection of objects and extraction of object features from video images (processes 210, 220)”:

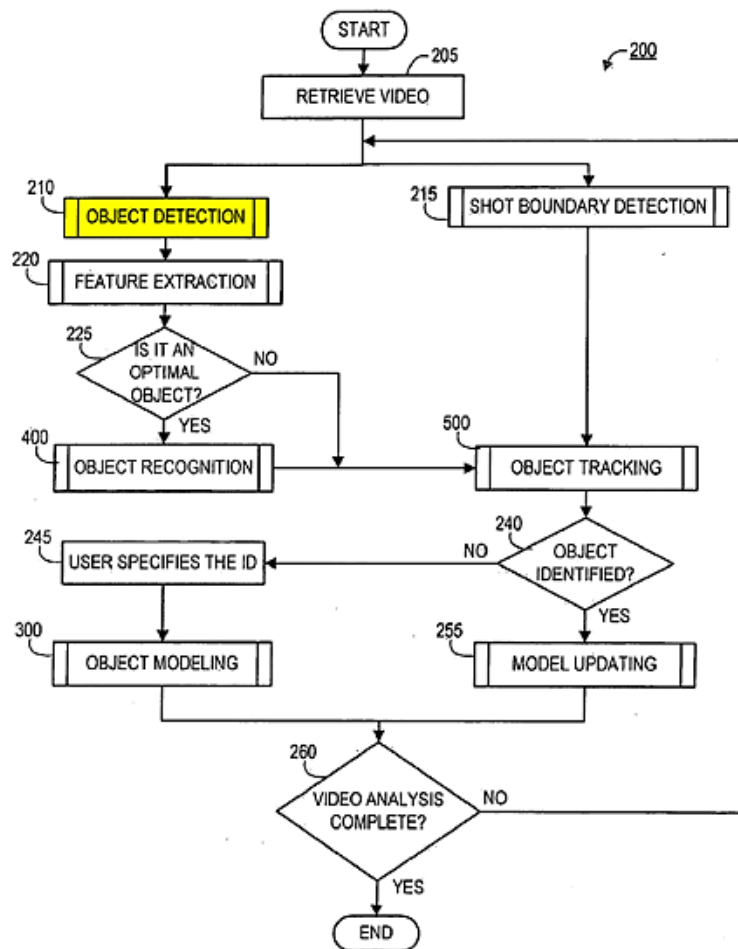


FIG. 2
OVERALL PROCESS

(Zhao, Fig. 2 (highlighting added).) Zhao describes in detail image detection process 210 to detect an object image from video images. (Zhao, ¶0058.)

Claim 1[b] further recites “**wherein the image detector is adjustable to increase detection of non-primary images in the video.**” Zhao does not disclose image detection functionality that is “**adjustable to increase detection of non-primary images in the video,**” but this would have been obvious in view of Steinberg. (Houh, ¶¶117-124.)

Zhao does describe techniques for a system operator to adjust settings based on desired accuracy and/or precision of recognition and/or object tracking, and also acknowledges the issue of false facial recognitions produced by extras or crowds in a movie. (Houh, ¶118.) For example, in connection with an object tracking/recognition process, Zhao explains how a system operator can adjust a similarity threshold. (Zhao, ¶0089 (“[T]he threshold value may be raised or lowered by a system operator depending on whether the operator would rather have higher accuracy, or more identifications.”).) Zhao also teaches that whether a high or low similarity threshold is desirable depends on the video content—for example, “in a facial recognition embodiment, extras or crowds in a movie often produce false recognitions.” (Zhao, ¶0091; Houh, ¶118.)

Like Zhao, **Steinberg** also relates to facial detection in digital images. (Houh, ¶119 (citing Steinberg, 1:49-53).)

With respect to claim 1**[b]**, Steinberg describes a “face detection process” to find faces in a captured image. (Steinberg, 18:25-31 (“An image-detection process, preferably a face detection process, is applied to the captured image to seek faces in the image, **1020**.”); *see generally id.*, 18:29-44, 19:18-36.) Steinberg explains that the face detection process can be implemented “by changing the sensitivity or threshold of the detection process.” (Steinberg, 18:35-37.) Steinberg describes how “each of the faces is given a weight based on size of the faces, location within the

frame, other parameters described herein, etc., which define the importance of the face in relation to other faces detected.” (Steinberg, 18:40-44.) Therefore, for example, “in response to false detection of one or more regions as one or more faces, or in response to a determination to concentrate on less image faces than faces identified in the identifying step,” “faces may be removed by increasing a sensitivity level of the face identifying step.” (Steinberg, 11:66-12:7; *see also id.*, 11:3-18 (“An automatically detected face may be removed in response to false detection of regions as faces, or in response to a determination to concentrate on less image faces or images faces that were manually determined to be of higher subjective significance, than faces identified in the identifying step. A face may be removed by increasing a sensitivity level of said face identifying step.”).) Although Steinberg focuses on increasing a sensitivity level to *decrease* detection of less important (i.e., non-primary) faces, a person of ordinary skill would have understood that decreasing the sensitivity level would have the opposite effect of *increasing* detection of non-primary faces in the image. (Houh, ¶120.)

Rationale and Motivation to Combine (Zhao and Kritt with Steinberg).

It would have been obvious to a person of ordinary skill to combine Steinberg’s teachings about an adjustable sensitivity threshold for detecting faces in images with Zhao’s video processing system (and under the proposed combination with Kritt). (Houh, ¶121.) The combination would have involved the straightforward

incorporation of an adjustable object detection setting as taught by Steinberg into Zhao's processes for object detection. (*Id.*) The combination would have predictably resulted in Zhao's video processing system in which the facial detection sensitivity could be adjusted, for example, to focus only on more important or prominent faces, or to additionally be able to detect less prominent faces (such as faces in the background or in a crowd). (*Id.*) The combination would have therefore resulted in **"an image detector configured to detect an image of an object in the still frame images, wherein the image detector is adjustable to increase detection of non-primary images in the video"** as claimed. (*Id.*)

Steinberg is an analogous reference to the '853 patent, being both related to object detection in images and reasonably pertinent to problems facing the '853 patent inventors. (Houh, ¶122.)

A person of ordinary skill would have been motivated to combine Zhao (and under the proposed combination with Kritt) with Steinberg. (Houh, ¶123.) Steinberg discloses that its techniques are applicable to video and provides motivation to incorporate an adjustable face detection step in Zhao, such as selective compression and sharpening. (Steinberg, 25:18-32 ("An image can be locally compressed so that specific regions will have a higher quality compression which equates to lower compression rate....By doing so, selected or determined regions of importance may maintain low compression or high resolution compared with regions

determined to have less importance or non-selected regions in the image.”); Houh, ¶123.) Other benefits would have been apparent to a person of ordinary skill such as processing time and efficiency improvements based on selectively increasing or decreasing face detection sensitivity based on needs—such as real-time processing—and subject matter—for example, decrease sensitivity for video involving a smaller group of individuals (thereby increasing detection of non-primary images) and increase sensitivity for video involving larger crowds such as sporting events. (Steinberg, 25:33-43 (“For example, such technology may be used for identification of faces in video sequences, particularly when the detection is to be performed in real-time.”); Houh, ¶123.)

A person of ordinary would have had a reasonable expectation that the combination would have been successful. (Houh, ¶124.) Because facial detection techniques were already well known, incorporating the ability to adjust the sensitivity of facial detection would have involved conventional programming techniques. (*Id.*)

- (c) **“an object recognizer configured to compare the image of the object to a fractal, wherein the fractal includes a representation of the object based on landmarks associated with the object, wherein the recognizer is further configured to update the fractal with the image.”**
(Claim 1[c])

As noted for claim 1[b], the '853 patent describes image “detection” and “recognition” as a two-step process. ('853, 13:39-46.) Zhao also describes a second, image “recognition,” step after performing image “detection.”

As explained below, claim 1[c] is obvious over Zhao in view of Kouzani. Zhao discloses processes to **“compare the image of the object”** to an object model that **“includes a representation of the object based on landmarks associated with the object,”** and to **“update”** an object model with the image. It would have been obvious, in view of Kouzani, for an object model to take the form of a **“fractal.”** Therefore, as explained below, the combination would have predictably resulted in an **“object recognizer”**—corresponding to Zhao’s processes under the proposed combination—configured as claimed. (Houh, ¶¶126-138.)

To start, Zhao discloses processes to **“compare the image of the object”** to an object model and **“update”** the object model with the image. (Houh, ¶127.) As shown in Figure 2, reproduced below, Zhao discloses an object recognition process **400** and a model updating process **255**:

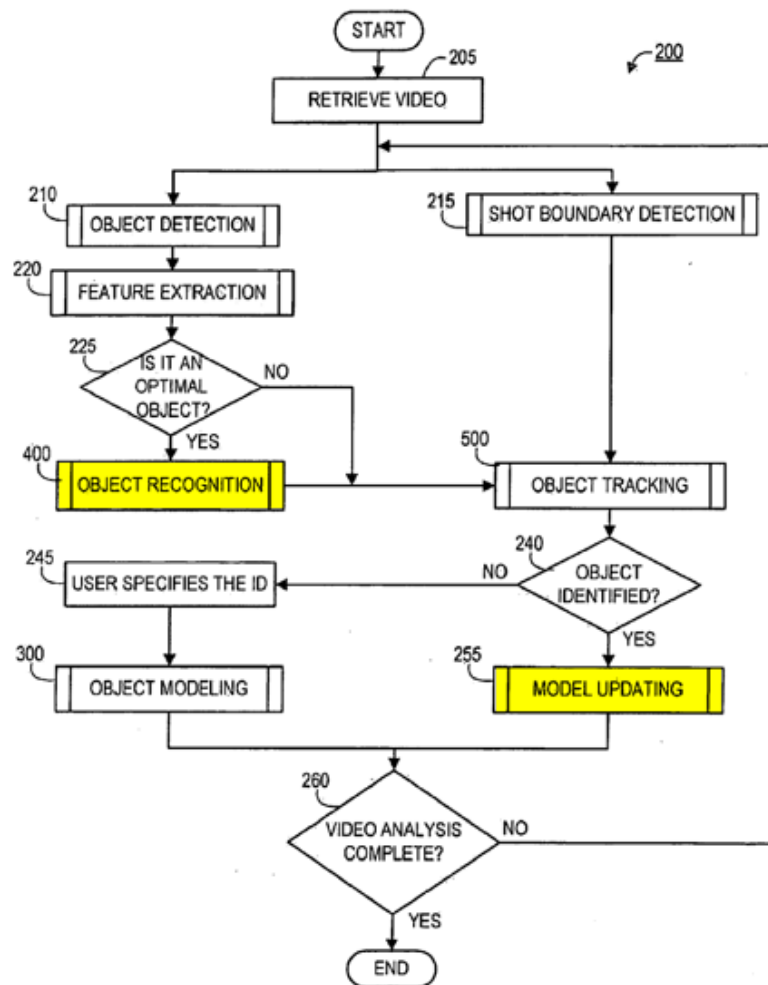


FIG. 2
OVERALL PROCESS

(Zhao, Fig. 2 (highlighting added).) Referring to Figure 2, Zhao explains that “the object recognition process compares an extracted object image, or set of object images, to a database of predefined object models in order to ‘recognize’ the extracted image as one of the models,” where an object model “comprises a set of data, wherein the data is further comprised of features associated with an object image.” (Zhao, ¶¶0061, 0065.) “[F]or facial recognition, features include eyes, noses, mouths, ears, chins, etc., of a face in an image, as well as the corners, curves,

color, overall shape, etc., associated with each feature.” (Zhao, ¶0058.) “Depending on whether the image is recognized or not, either an existing object model is updated or a new object model is created.” (Zhao, ¶0061.) “For example, the recognized object image could include an actor or actress, a political figure, an animal, or some type of inanimate object.” (Zhao, ¶0063.) If the image is recognized, an object model associated with the recognized object—such as a particular person—is updated with the image by using extracted features of the image:

[T]he system calculates the similarity between the extracted features and those already existing in the model (process 410), as described in greater detail below. For features that are determined to be similar to those already existing in the model, the relative weights of those features are updated, as it is assumed that these features are prominent in the object because they are reoccurring (step 325). Features that are not similar to those already existing are added to the model as new features (step 325), and thus the model becomes more robust as it develops additional features, that can be used to recognize object images in subsequent recognition procedures.

(Zhao, ¶0071.) “For example,” Zhao explains, “in a facial recognition embodiment, a given character’s or actor’s appearance often changes over time, due to weight gain, hair loss, aging, or other similar reasons.” (Zhao, ¶0073.) With Zhao’s system, “each model is updated over time as new episodes or videos are released, and thus each model adapts and changes in response to changes in a given actor’s

appearance.” (Zhao, ¶0073.)

Zhao further discloses that an object model **“includes a representation of the object based on landmarks associated with the object.”** (Houh, ¶128.) As just noted, Zhao explains how an object model includes weighted features of the object, and in a facial recognition application, those features include facial **“landmarks”** such as “eyes, noses, mouths, ears, chins, etc., of a face in an image, as well as the corners, curves, color, overall shape, etc., associated with each feature.” (Zhao, ¶¶0058 (facial features), 0065 (object model “comprises a set of data, wherein the data is further comprised of features associated with an object image”), 0071 (object model updating).) Zhao’s example facial **“landmarks”** are just like those identified in the ’853 patent. (Houh, ¶128 (citing ’853, 13:53-57).)

Zhao therefore discloses every aspect of claim 1 [c], except possibly **“fractal”**:

an object recognizer configured to compare the image of the object to a[n object model], wherein the [object model] includes a representation of the object based on landmarks associated with the object, wherein the recognizer is further configured to update the [object model] with the image.

(’853 claim 1 [c] (replacing “fractal” with “object model”); Houh, ¶129.)

Turning to **“fractal,”** the ’853 patent states, for example, that “[r]ecognition processing can include creating a fractal representation of the face or object that was detected” and that “[f]or detection and/or recognition requests, fractal datasets can

be compared to a reference dataset and/or scored with a probability of a match based on comparing individual fractal values against one another.” (’853, 13:47-49, 19:13-16.)

Zhao does not expressly describe comparing an image of the object to a “**fractal**” or “**updat[ing] the fractal**” with the image. However, in Patent Owner’s opposition to Petitioner’s motion to dismiss in the related district court litigation, Patent Owner appears to argue that a “fractal” is nothing more than a shape or polygon, or potentially a shape or polygon used to identify objects or faces. (EX1017, p.007.) Under Patent Owner’s apparent interpretation, Zhao’s object model would disclose a “**fractal**,” since the object model comprises one or more shapes. (Houh, ¶131 (citing Zhao, ¶0058).) And updating Zhao’s object model would accordingly disclose “**updat[ing] the fractal**.” (*Id.* (citing Zhao, ¶¶0071, 0073)).) Zhao therefore discloses every aspect of claim 1[c] under Patent Owner’s apparent interpretation of “fractal.” (*Id.*)

Putting aside Patent Owner’s apparent interpretation, it would have been obvious to compare an image of the object to a “**fractal**” or “**updat[ing] the fractal**” with the image. (Houh, ¶132.) The use of fractals for facial recognition was already known before the ’853 patent. (*Id.*) In fact, the ’853 patent identifies the use of fractals among a number of known recognition techniques, including Principal Component Analysis (PCA) and Elastic Bunch Graph Matching (EBGM), among

others. (Houh, ¶132 (citing '853, 14:30-15:47; Zhao, ¶0072 (Elastic Bunch Graph Matching); Kouzani, p.009 (Principal Component Analysis)).) It was known, for instance, that “[a] human face is a surface lying in the fractal space intrinsically. Since many of the parameters are self-similar, the fractal model can better represent faces, especially in handling facial variations, such as pose, illumination, etc.” ('853, 15:40-44; Houh, ¶132.) For example, a 2010 paper entitled “Facial Image Recognition Based on Fractal Image Encoding” explains:

Fractal image encoding uses self-similarity to encode an image. Utilizing this feature, a new method of facial image recognition based on fractal image encoding is examined. Fractal image encoding can approximate any given image by capturing intrinsic self-similarity within the image. This method has some invariance to rotation, scaling, translation, and luminance.

(EX1014, p.005.) Even as of 2010, the use of fractals for facial recognition was not at all new—Kouzani, for example, dates from 1997. (EX1006, p.007; Houh, ¶132.)

It would have been obvious in view of **Kouzani** for Zhao’s “object model” for a face to take the form of a “**fractal**.” (Houh, ¶133.) Kouzani, entitled “Fractal Face Representation and Recognition,” “presents a face representation and recognition scheme based on the theory of fractals.” (Kouzani, p.007.) Kouzani explains that “[a] fractal is a fragmented geometric shape that can be subdivided in parts, each of which is a reduced-size copy of the whole,” that “[f]ractals model

many real-world objects that do not correspond to simple geometric shapes, e.g. clouds, mountains, turbulence, and coastlines,” and that “[f]ractals are generally self-similar and independent of scale.” (*Id.*) In Kouzani’s system, “[e]ach face image is represented by its fractal model which is a small collection of transformation parameters.” (*Id.*) “For recognition, the input face image is transformed and its fractal model is then compared against the database of fractal models of known faces.” (*Id.*) Kouzani explains a comparison process for recognizing a face performed with a neural network:

The second network has also a feedforward architecture and has been trained to recognise faces. It receives two group of inputs, the fractal model of an input image and the fractal model of a known face from the fractal database. It calculates the degree of similarity between the two fractal face models. If the output is below a threshold, the input face is considered as a known face whose fractal is currently under examination. Similar to the compression stage, more networks with the same architecture can be used to speed up the recognition process.

(Kouzani, p.009.)

Kouzani also touts advantages of using fractal models for facial recognition, including being less computationally expensive than prior art principal component analysis (PCA) techniques and providing high data compression:

In the recognition stage, two groups of experiments were performed. In the first experiment, the known face images were presented to the

system and the recognition process was carried out. 100% recognition was achieved in this step. The maximum compression ratio obtained for successful recognition of all known faces was 89:1. Figure 6 demonstrates recognition of the known face number 100. In the second experiment, 50 unknown face images were directed to the system and the recognition process was carried out. None of the input images were recognised by the system. Figure 7 shows the recognition result for an unknown input face image.

The proposed system is less computationally expensive than that for PCA. Moreover, when a new face is added to the database, recalculation of eigenfaces and re-compression of known faces are not necessary.

(Kouzani, p.009; *see also id.* (“Simulation results show that the proposed system overcomes the shortcomings of PCA.”); Houh, ¶134.) Kouzani further notes that “real-time recognition” was achieved. (Kouzani, pp.008-009 (“The main reason for employing neural networks is to take advantage of the highly parallel architecture of neural networks in order to achieve real-time recognition.”).)

Rationale and Motivation to Combine (Zhao, Kritt, and Steinberg with Kouzani). It would have been obvious to a person of ordinary skill to combine Kouzani’s techniques for facial recognition using fractals with Zhao’s system (and under the proposed combination with Kritt and Steinberg). (Houh, ¶135.) The combination would have involved the use of Kouzani’s known fractal-based facial recognition techniques to improve Zhao’s facial recognition system in a similar way

(by for example, providing a less computationally expensive system). (*Id.*) The combination would have predictably resulted in Zhao's system for facial detection and recognition in which recognition was performed by processes that compared a fractal of a detected facial image against a fractal model, and that updated the fractal model with the facial image. (*Id.*) The combination would have therefore resulted in **“an object recognizer configured to compare the image of the object to a fractal, wherein the fractal includes a representation of the object based on landmarks associated with the object, wherein the recognizer is further configured to update the fractal with the image”** as claimed. (*Id.*)

Kouzani is an analogous reference to the '853 patent, being in the same field of image recognition and reasonably pertinent to problems facing the '853 patent inventors. (Houh, ¶136.)

A person of ordinary skill would have been motivated to combine. (Houh, ¶137.) As discussed above, Kouzani touts the value of fractal-based facial recognition, including recognition results, real-time performance, high data compression, and being less computationally expensive compared to other techniques. (*Id.* (citing Kouzani, p.009).) A person of ordinary skill would have therefore been motivated to implement Kouzani's techniques in Zhao's system to obtain the same benefits and advantages. (*Id.*) Moreover, as one of a number of

known facial recognition techniques, the use of fractals would have been obvious to try. (*Id.*)

A person of ordinary skill would have had a reasonable expectation that the combination would be successful. (Houh, ¶138.) As discussed above, the use of fractals for facial recognition and image processing generally was well known in the art. The combination would have involved the straightforward incorporation of known fractal coding techniques to Zhao’s system. (*Id.*)

2. Claim 2: “The system of claim 1, wherein the coordinator is configured to increase the plurality of demultiplexer nodes when a threshold of processing capacity is reached.”

The additional limitations of claim 2 are disclosed by and obvious over Zhao in view of Kritt. (Houh, ¶¶139-142.)

As discussed for claim 1, it would have been obvious to implement Zhao’s video processing system in a cloud-based environment involving a number of nodes configured to perform demultiplexing operations. (Houh, ¶¶106-113, 140.)

It would have further been apparent and obvious to a person of ordinary skill that Zhao’s system, in a cloud-based implementation, would include “**coordinator**” functionality “**to increase the plurality of demultiplexer nodes when a threshold of processing capacity is reached**” as claimed. (Houh, ¶141.) The ability to dynamically scale capacity based on workload was a key objective, and selling point, of cloud-based infrastructure. (*Id.*) As Kritt explains, cloud computing “generally

refers to the provision of scalable computing resources as a service over a network” and “enable[es] convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.” (Kritt, ¶0078.) One example of a commercial service providing such functionality that would have been well-known to a person of ordinary skill is Amazon’s AWS and Elastic Compute Cloud (EC2) service and its Auto Scaling feature. (Houh, ¶141.) For example, the 2011 book *Programming Amazon EC2* explains that “[t]he cloud’s biggest promise is elasticity: running with just the amount of computing resources your workload demands at any given time.” (EX1013, p.0042.) “[F]or some application, you need to be able to dynamically scale EC2 capacity, or instances. And that is what Auto Scaling does. It allows you to define conditions for increasing or decreasing the number of instances.” (EX1013, p.0042.) The authors further explain that “[a]ccording to AWS, ‘Auto Scaling is particularly well suited for applications that experience hourly, daily, or weekly variability in usage.’ ...At peak hours, we want to increase our capacity to handle the load, and when it’s not so busy, we want to run idle.” (EX1013, p.0042.)

Therefore, as discussed for claim 1, in an application of Zhao involving a large-scale system (such as a commercial video indexing and retrieval website or security/surveillance service), it would have been obvious to a person of ordinary

skill for the system to include functionality to scale computing resources based on whether a processing capacity threshold had been reached. (Houh, ¶142.) At times when that threshold is reached, more computing resources—such as demultiplexer nodes—could be provisioned in order to continue to provide a responsive service. And at times with less demand, demultiplexer nodes could be released. (*Id.*) Therefore, **“the coordinator is configured to increase the plurality of demultiplexer nodes when a threshold of processing capacity is reached”** as claimed would have been obvious. (*Id.*)

3. Claim 3: “The system of claim 1, wherein the recognizer is configured to determine distinguishing geometric features of the object.”

The additional limitations of claim 3 are disclosed by and obvious over Zhao and Kouzani. (Houh, ¶¶143-146.)

The **“recognizer”** recited in claim 3 refers back to the **“object recognizer configured to compare the image of the object to a fractal...”** recited in claim 1. As discussed for claim 1[c], the claimed **“object recognizer”** is obvious over Zhao and Kouzani.

With respect to **“determining distinguishing geometric features of the object,”** claim 4 of the ’853 patent identifies some examples: “a contour of eye sockets, a nose, and a chin.” As discussed for claim 1[c], Zhao determines these sorts of geometric features of a face in its recognition process, by both identifying

them in the object image and comparing them against those features in an object model. (Zhao, e.g., ¶¶0071 (“[T]he system calculates the similarity between the extracted features and those already existing in the model (process **410**)”), 0058 (“[F]or facial recognition, features include eyes, noses, mouths, ears, chins, etc., of a face in an image, as well as the corners, curves, color, overall shape, etc., associated with each feature.”).) Under the proposed combination with Kouzani, these geometric features would have been embodied in fractal representations. (Houh, ¶145.)

Therefore, the proposed combination of Zhao and Kouzani would have predictably resulted in the “**object recognizer**” “**configured to determine distinguishing geometric features of the object.**” (Houh, ¶146.)

4. Claim 4: “The system of claim 3, wherein the distinguishing geometric features comprise a contour of eye sockets, a nose, and a chin.”

Petitioner already largely addressed claim 4 in the discussion of claim 3 above. As explained, Zhao discloses distinguishing geometric features including “**a contour of eye sockets**”—e.g., the overall shape associated with eyes—“**a nose, and a chin.**” (Zhao, ¶0058 (“For example, for facial recognition, features include eyes, noses, mouths, ears, chins, etc., of a face in an image, as well as the corners, curves, color, overall shape, etc., associated with each feature.”); Houh, ¶147.)

D. Ground 2: Claims 5-7 Are Obvious Over Ground 1 Prior Art in Further View of Yang and Romdhani

1. Claim 5: “The system of claim 4, wherein the recognizer is configured to determine skin textures of the object.”

As explained below, the additional limitations of claim 5 are disclosed by and obvious in view of Yang and Romdhani. (Houh, ¶¶148-159.)

Zhao mentions “skin color” as a feature used in facial/object detection and recognition. (Zhao, ¶0066.) Zhao provides only exemplary lists of facial features used for object detection and recognition, and does not describe determining or recognizing skin textures. (Zhao, ¶¶0058 (“for facial recognition, features include eyes, noses, mouths, ears, chins, etc., of a face in an image, as well as the corners, curves, color, overall shape, etc., associated with each”), 0066 (“such as a nose, mouth, eyes, ears, chin, and other similar features” for face detection).) But using skin textures for recognition (and detection) would have been obvious. (Houh, ¶49.)

Kritt, for example, describes using “skin texture” for recognizing a facial object in a video image. (Kritt, ¶0044 (“[A] method may be of a type that employs rules that identify so-called ‘invariant features’ that are present in a frame regardless of the position or pose of the object, the lighting, or camera viewpoint. Methods of this type, especially when employed to identify a human facial object, may employ an image recognition processes that identifies: (i) facial features using edge detectors

(e.g., a Sobel filter) and templates; (ii) skin or hair texture using a neural network; and (iii) skin color using a pixel chrominance classifier.”); Houh, ¶150.)

As another example, Steinberg discloses that known face detection techniques include identification of facial features, texture and skin color. (Steinberg, 1:65-2:5 (“...methods of face detection which may be divided into four main categories: (i) knowledge-based methods; (ii) feature-invariant approaches, including the identification of facial features, texture and skin color; (iii) template matching methods, both fixed and deformable and (iv) appearance based methods, including eigenface techniques, statistical distribution based methods and neural network approaches.”); Houh, ¶151.)

Yang, which Steinberg both discusses and expressly incorporates by reference (Steinberg, 1:62-65, 7:1-3; *see generally id.*, 6:40-7:6 (“Incorporation by Reference” section)), provides more detail about facial recognition involving “**skin textures.**” (Houh, ¶152.) Yang, entitled “Detecting Faces in Images: A Survey,” reviews various techniques for detecting faces in images. Yang explains that “[h]uman faces have a distinct texture that can be used to separate them from different objects.” (Yang, p.009.) Yang describes how “Augusteijn and Skufca developed a method that infers the presence of a face through the identification of face-like textures” and “[t]o infer the presence of a face from the texture labels, they suggest using votes of the occurrence of hair and skin textures.” (Yang, p.009.)

Yang also describes another technique by Dai and Nakano that uses a “face texture model” that incorporates color information. “Using the face texture model, they design a scanning scheme for face detection in color scenes in which the orange-like parts including the face areas are enhanced.” (Yang, p.009.)

Romdhani, also cited in Steinberg (Steinberg, references cited, p.005), describes facial recognition using texture that produced “state of the art results.” (Romdhani, 0039; Houh, ¶153.) Entitled “Face Identification by Fitting a 3D Morphable Model using Linear Shape and Texture Error Functions,” Romdhani describes techniques for face recognition that involve matching face texture to an image of a known individual. (Romdhani, p.0024 (“Abstract”)) (“This paper presents a novel algorithm aiming at analysis and identification of faces viewed from different poses and illumination conditions. Face analysis from a single image is performed by recovering the shape and textures parameters of a 3D Morphable Model in an analysis-by-synthesis fashion.”).) Romdhani explains the face recognition techniques in great detail, describing how “[t]he fitting procedure provides shape and texture coefficient which ... depends of [sic] the identity only. Identification of a novel face image is performed by matching the coefficients recovered from the image with the coefficients recovered from the image of a known individual.” (Romdhani, p.0036.) Romdhani explains that “[s]tate of the art results

were obtained” (Romdhani, p.0039), outperforming “one of the most successful commercial face recognition system[s]” (Romdhani, p.0037).

Rationale and Motivation to Combine (Ground 1 Prior Art with Yang and Romdhani). It would have been obvious to a person of ordinary skill to combine Yang’s and Romdhani’s teachings regarding using skin and face textures for facial detection and recognition with Zhao’s video processing system (and under the proposed combination). (Houh, ¶154.) The combination would have involved the incorporation of skin texture feature(s) as one of several features used for face recognition, and it would have predictably resulted in Zhao’s video processing system in which Zhao’s face recognition processes used skin texture as a feature when comparing a detected face image to an object model (or fractal, under the combination with Kouzani). (*Id.*) The combination therefore discloses and renders obvious “**the [object] recognizer**” “**configured to determine skin textures of the object**” as claimed. (*Id.*)

Both Yang and Romdhani are analogous references to the ’853 patent, being both directed to the field of image processing and object detection/recognition, including facial detection and recognition, and reasonably pertinent to problems facing the ’853 patent inventors. (Houh, ¶155 (citing Yang, e.g., p.004 (Abstract); Romdhani, p.0024 (Abstract)).)

A person of ordinary skill would have been motivated to combine. Houh, ¶156.) For starters, as noted above, both Yang and Romdhani were identified in Steinberg. (See Steinberg, 1:62-65 (Yang), references cited, p.5 (Romdhani).) Steinberg specifically identifies Yang in connection with face detection techniques using “identification of facial features, texture and skin color” (Steinberg, 1:62-65) and moreover, expressly incorporates Yang (Steinberg, 7:1-3; see generally *id.*, 6:40-7:6 (“Incorporation by Reference” section)). Yang therefore is effectively part of Steinberg as if Yang were explicitly contained in Steinberg. See *Paice, LLC v. Ford Motor Co.*, 881 F.3d 894, 906-907 (Fed. Cir. 2018) (incorporated material “effectively part of the host document as if it were explicitly contained therein”). Based on at least their express identification by title (and incorporation, in Yang’s case), a person of ordinary skill would have appreciated that both references were directly relevant to techniques for face detection and recognition. (Houh, ¶156.) Because, as discussed, Zhao provides only exemplary lists of facial features (which largely overlap) to use for detection and recognition, a person of ordinary skill would have been motivated to investigate additional features that could provide improved recognition results. (*Id.* (citing Zhao, ¶¶0058, 0066).)

In fact, Zhao teaches that, for face recognition, “more discriminative and detailed features than those used for general model detection [e.g., face detection] should be extracted from the given image.” (Zhao, ¶0069.) A person of ordinary

skill would have therefore been motivated to investigate additional “more discriminative and detailed features,” and would have appreciated that the incorporation of face textures, such as skin textures, would provide a further way to discriminate among potential recognition matches. (Houh, ¶157.) At a minimum, a person of ordinary skill would have appreciated that, as one of a number of identified features for facial detection and recognition, using skin textures would have at least been obvious to try. (*Id.*)

Moreover, as noted Romdhani explains that “[s]tate of the art results were obtained” (Romdhani, p.0039), outperforming “one of the most successful commercial face recognition system[s]” (Romdhani, p.0037). A person of ordinary skill would have therefore been motivated to incorporate the use of face textures, such as skin textures, to try to achieve similar state of the art results. (Houh, ¶158.)

A person of ordinary skill would have had a reasonable expectation that the combination would have been successful. (Houh, ¶159.) The combination would have involved the straightforward incorporation of another face feature, skin textures, to use for facial recognition. (*Id.*) As Yang and Romdhani both make clear systems making use of face textures, including skin textures, was known in the art.

2. Claim 6: “The system of claim 5, wherein the distinguishing geometric features or the skin textures are determined based on a three-dimensional model.”

Zhao does not describe determining distinguishing geometric features or skin textures “**based on a three-dimensional model**,” but this would have been obvious. (Houh, ¶¶160-163.)

Romdhani discloses using a “**three-dimensional model**” for face recognition. (Houh, ¶161.) Romdhani presents an algorithm, called “LiST” (for “Linear Shape and Texture,” that involves “fitting a 3D Morphable Model to face images.” (Romdhani, p.0039.) Romdhani explains that the algorithm “exploits the linearity of the model in the recovery of shape and texture parameters leading to a robust and efficient algorithm” and “[s]tate of the art results were obtained” when “LiST was evaluated on the difficult problem of combined pose-illumination variation face identification.” (Romdhani, p.0039.) Romdhani notes that their face identification results outperformed “one of the most successful commercial face recognition system[s]” and that the results “clearly show that, for pose-invariant face identification, a 3D representation is valuable.” (Romdhani, p.0037.)

It would have been obvious to a person of ordinary skill to combine Romdhani’s teachings regarding using a three-dimensional facial model for facial recognition with Zhao’s video processing system. (Houh, ¶162.) The combination would have predictably resulted in Zhao’s object recognition processes using a

three-dimensional model for facial recognition, and accordingly, in determining distinguishing geometric features—such as eyes, nose, and chin—or skin textures based on a three-dimensional model. (*Id.*) The combination with Romdhani therefore discloses and renders obvious “**wherein the distinguishing geometric features or the skin textures are determined based on a three-dimensional model**” as claimed. (*Id.*)

Petitioner discussed the rationale and motivation to further combine with Romdhani in the analysis of claim 5 above, which applies here as well. (Houh, ¶¶156-159, 163.) With respect to claim 6, Romdhani provides express motivation to incorporate a three-dimensional model for facial recognition, explaining that state of the art results were obtained, and that “a 3D representation is valuable.” (Romdhani, pp.0037-0039; Houh, ¶163.)

3. Claim 7: “The system of claim 6, wherein the three-dimensional model is a three-dimensional morphable model.”

As discussed for claim 6, Romdhani expressly discloses a “**three-dimensional model**” that is a “**three-dimensional morphable model**.” (Romdhani, e.g., pp.0024, (Title: “Face Identification by Fitting a 3D Morphable Model using Linear Shape and Texture Error Functions”), 0039 (“fitting a 3D Morphable Model to face images”); *see also id.*, 0025-0026 (describing “3D Morphable Face Model”); Houh, ¶164.)

The rationale and motivation to further combine with Romdhani was discussed above in connection with claims 5 and 6, which applies here as well. (Houh, ¶¶156-159, 163, 165.)

E. Ground 3: Claims 8-9 Are Obvious Over Ground 1 Prior Art in Further View of Trivedi

1. Claim 8: “The system of claim 1, further comprising a processor and a camera for capturing the video.”

Claim 8 is obvious in view of Trivedi. (Houh, ¶¶166-173.)

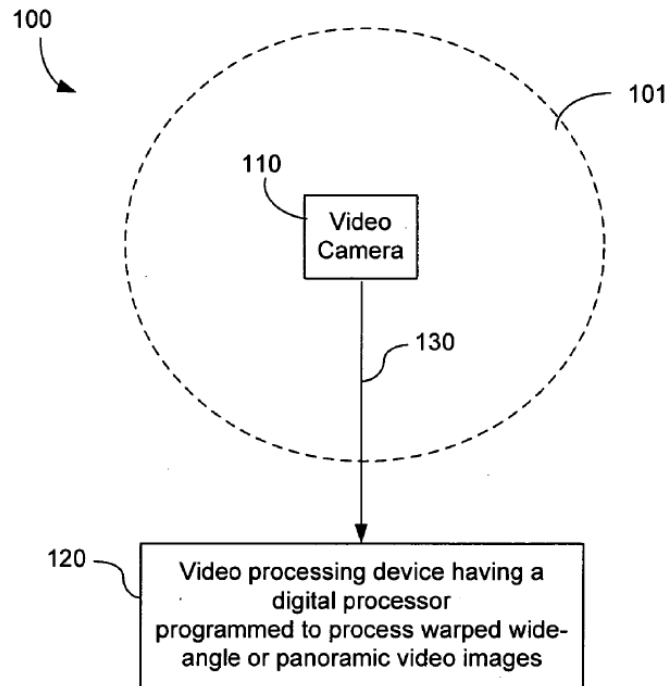
Zhao discloses “**a processor.**” (Zhao, e.g., ¶¶0029 (“The system comprises a processor and a computer program product...”), 0055 (“[E]mbodiments of the present system may be operated in a computer environment including databases and other storage apparatuses, servers, processors...”); Houh, ¶167.)

With respect to “**a camera for capturing the video,**” Zhao also teaches using its techniques in systems that a person of ordinary skill would have understood and found obvious would include a camera, such as video surveillance systems. (Houh, ¶168.) For example, Zhao explains that “[e]mbodiments of the present system are useful for a wide variety of applications, including video indexing and retrieval, video surveillance and security, unknown person identification, advertising, and many other fields.” (Zhao, ¶0053; *see also id.* ¶¶0003 (“security and surveillance”), 0007 (“security systems can benefit from accurate and consistent identification of perpetrators or victims within surveillance videos”).)

To the extent there is any question that Zhao alone discloses and renders obvious “**a camera for capturing the video,**” **Trivedi** confirms that video cameras are used in security/surveillance systems and expressly discloses a security system that includes a camera. (Houh, ¶169.) Trivedi, entitled “Digital Processing of Video Images,” explains:

Video cameras have been used to capture video images for various applications such as surveillance, machine vision, security monitoring, inspection, sensing and detection. In these and other applications, the captured video images may be sent to a nearby or remote image processing device to be processed and displayed. Digital image processing techniques may be used to process the video images to extract information from the video images.

(Trivedi, ¶0003.) As shown in Figure 1, reproduced below, Trivedi discloses a video system **100** that includes a video camera **110** and video processing device **120**:



(Trivedi, Fig. 1; *see also id.*, ¶0040 (“FIG. 1 illustrates one example of a video system 100 where a video camera 110 is installed at a location 101 to capture a video of the location 101....A video processing device 120 is located at a different location and is linked to the video camera 110 by a communication link 130 to receive the video signal.”).)

Petitioner notes that, as discussed in Part III.C above, the Petition cites Trivedi for the same claim limitations in dependent claims 8-9 for which the Examiner cited them during prosecution, and which was not distinguished or disputed by the applicant.

Rationale and Motivation to Combine (Ground 1 Prior Art with Trivedi).

It would have been obvious to a person of ordinary skill to combine Trivedi’s

teachings regarding a camera for capturing video. (Houh, ¶170.) The combination would have involved the straightforward incorporation of a video camera into Zhao’s system (to the extent a camera was not already part of Zhao’s system), and would have predictably resulted in Zhao’s system having a video camera for capturing video. (*Id.*) The combination therefore discloses a system having “**a camera for capturing the video**” as claimed. (*Id.*)

Trivedi is an analogous reference to the ’853 patent, both by describing a similar video processing system that involves “real-time tracking of persons and objects” and “face recognition of persons,” and being reasonably pertinent to problems facing the ’853 patent inventors. (Houh, ¶171 (citing Trivedi, ¶0005).)

A person of ordinary skill would have been motivated to further combine with Trivedi. (Houh, ¶172.) As discussed, Zhao teaches that its system has a wide range of applications, including video surveillance and security. (Zhao, ¶0053; *see also id.* ¶¶0003, 0007.) Therefore, for starters, Trivedi simply provides express confirmation of what was well-known to laypersons and persons of ordinary skill alike—video surveillance and security systems include one or more video cameras for capturing video. (Houh, ¶172.) Moreover, a person of ordinary skill in the art, in view of Zhao’s teachings, a person of ordinary skill when implementing Zhao’s system would have looked to a reference such as Trivedi for techniques for capturing video—including by incorporating a video camera. (*Id.*)

A person of ordinary skill would have had a reasonable expectation that the combination would have been successful, as it would have involved the straightforward incorporation of a video camera into Zhao’s video processing system. (Houh, ¶173.)

2. Claim 9: “The system of claim 8, wherein the processor is configured to transform the video into a rectilinear format.”

Claim 9 is obvious in view of Trivedi. (Houh, ¶¶174-178.)

Trivedi describes capturing video in a warped, non-rectilinear format—for example, video captured with a wide angle lens—and “**transforming the video to a rectilinear format.**” (Houh, ¶175.) Trivedi explains:

Certain video cameras are designed with optical imaging modules to capture wide-angle or panoramic video images. Such video images are distorted due to the designs of various optical imaging modules. Digital imaging processing may be used to transform the distorted images into images that a viewer would normally perceive when directly look at the scene being captured.

(Trivedi, ¶0004.) Trivedi’s system includes digital processing techniques to “efficiently transform warped video images into rectilinear video images.” (Trivedi, ¶0005.) Trivedi explains that the transform of video into a rectilinear format is performed by the digital processor of the system’s video processing device. (Trivedi, ¶0042 (“[T]he digital processor may be programmed, among other features and functions, to transform the warped video images into rectilinear video

images[.]”); *see also id.*, ¶0041 (“The video processing device **120** includes a digital processor that is programmed to process the warped wide-angle or panoramic video images to extract desired information.”), Fig. 2 (showing device **120** showing device **120**, stating “Video processing device having a digital processor programmed to process warped wide-angle or panoramic video images”).)

As noted in the analysis of claim 8 (and Part III.C above), the Examiner cited Trivedi during prosecution for disclosing the limitations recited in claims 8 and 9, and was not distinguished by the applicant.

It would have been obvious to further incorporate Trivedi’s teachings regarding transforming the video, such as video captured by a wide-angle camera of a surveillance system, into a rectilinear format. (Houh, ¶177.) The combination would have predictably resulted in Zhao’s video processing system having a processor “**configured transform the video into a rectilinear format**” as claimed. (*Id.*)

Petitioner discussed the rationale and motivation to combine Ground 1 Prior Art with Trivedi in the analysis of claim 8 above, which applies here as well. (Houh, ¶¶170-173, 178.) With respect to claim 9, incorporating Trivedi’s teachings regarding transforming the video, such as video captured by a wide-angle camera of a surveillance system, into a rectilinear format would serve two purposes. (Houh, ¶178.) First, it would advantageously enable a viewer of the video to see the scene

in an undistorted way that the viewer would normally perceive when directly looking at the scene. (Trivedi, ¶0004 (“Digital imaging processing may be used to transform the distorted images into images that a viewer would normally perceive when directly look at the scene being captured.”); Houh, ¶178.) Second, like Zhao, Trivedi describes using its system for face recognition. (Trivedi, e.g., ¶¶0005, 0092-0094.) A person of ordinary skill would have readily appreciated that, in order to best perform face recognition, video should be un-warped into a rectilinear format to better recognize distinguishing facial features. (Houh, ¶178.) Trivedi explains, for example, how the face video stream used for face recognition “may be obtained from the unwarped rectilinear video stream produced by the DTV module of the digital processor.” (Trivedi, ¶0095.) Moreover, because video input into Trivedi’s system for face recognition could comprise both rectilinear and non-rectilinear formatted video (Trivedi, ¶0095), a person of ordinary skill would have also understood that it would be desirable for all video input to be converted to a common, rectilinear format to help ensure better, more consistent face recognition. (Houh, ¶178.)

F. Ground 4: Claims 10-11 Are Obvious Over Ground 1 Prior Art in Further View of Singer

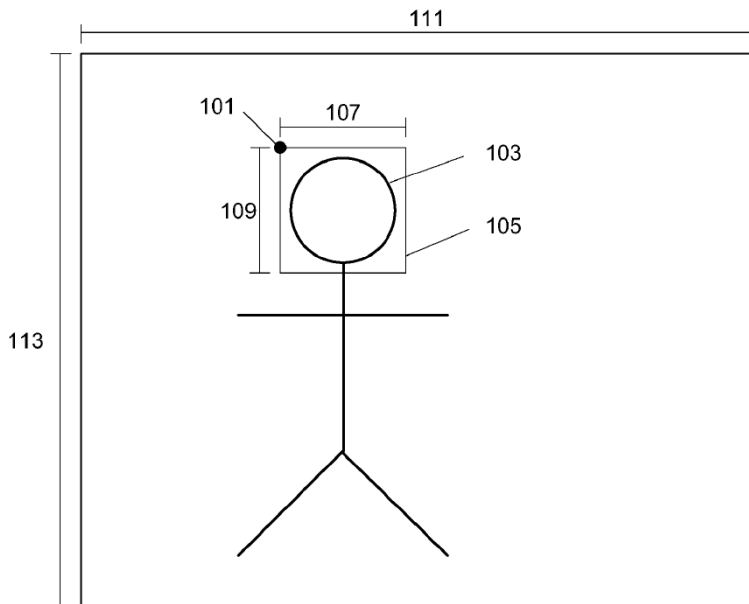
- 1. Claim 10: “The system of claim 1, wherein the coordinator is configured to embed metadata about the object into the video, and wherein the metadata comprises a timestamp and a coordinate location of the object in the still frame images.”**

Zhao describes various uses for “**metadata about the object**” generated as a result of its image processing techniques, including indexing and retrieval in order to locate and retrieve videos or portions of videos with the associated recognized images. (Houh, ¶179 (quoting Zhao, ¶0092 (“[I]mages in videos that have been identified and labeled as known objects (i.e. corresponding to model) are encoded with metadata attributes identifying those images. These encoded metadata attributes are generally stored in index files and are configured to be searched in order to locate and retrieve videos or portions of videos with the associated images.”); citing Zhao, ¶¶0053 (“Embodiments of the present system are useful for a wide variety of applications....”), 0007 (“The ability to effectively and efficiently index, store, and retrieve videos, or portions of videos, based on objects in those videos is important for a variety of fields.”).)

Zhao describes “metadata attributes identifying those images” that can be “searched in order to locate and retrieve videos or portions of videos with the associated images.” (Zhao, ¶0092.) Zhao does not further detail what exactly the “metadata attributes” are, but because they are used to locate portions of videos with

the associated object images, a person of ordinary skill would have understood and found it obvious that the metadata attributes would “**comprise[] a timestamp and a coordinate location of the object in the still frame images.**” (Houh, ¶180.) It would have also been obvious to “**embed metadata about the object into the video.**” (*Id.*) As explained below, both of these aspects of claim 10 are disclosed in Singer.

Singer, entitled “Object Detection Metadata,” describes techniques for generating and storing metadata about objects detected in video image data, including face detection metadata. (Singer, ¶¶0023-0024; Houh, ¶181.) For example, Figure 1 of Singer illustrates detection of a face in an image of a video:



(Singer, Fig. 1.) Singer explains that “the location of each face in a frame of image data may be described with a simple rectangle.” (Singer, ¶0025.) For example, in

Figure 1, “[f]ace detection rectangle **105**,” which bounds face **103**, “has dimensions of height **109** and width **107** and an initial point **101**.” (Singer, ¶0026.) Singer in figure 9 illustrates an example of metadata for a face detection rectangle:

Detected-face entry 903	
X =	15
Y =	20
Width =	32
Height =	26
ID =	3
Layer =	2
DetectWeight =	243
MatchWeight =	173

(Singer, Fig. 9.) Singer explains that “Detected-object entry **903** provides one example of a detected-face entry in a metadata track of image data.” (Singer, ¶0053.) The X and Y values, and Width and Height values, specify the location and dimensions, respectively, of the detected face in the image. (Singer, ¶0053.)

Petitioner notes that Singer’s depiction and description of detection rectangles matches Zhao’s depiction of using rectangles to identify detected objects in a still frame image of a video:

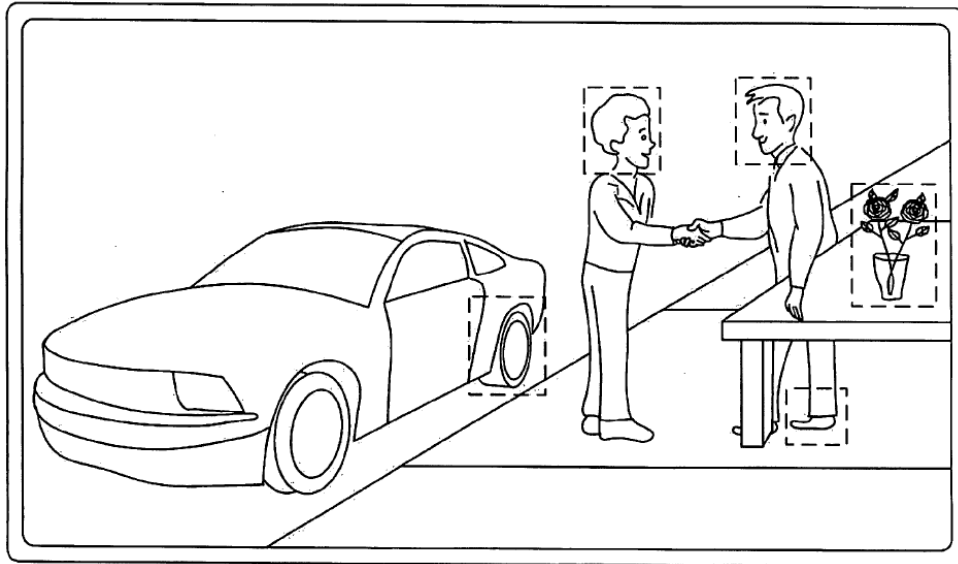


FIG. 1

(Zhao, Fig. 1, ¶0003 (“For example, FIG. 1 illustrates a sample frame of a video scene. Exemplary objects that are capable of being recognized within the illustrated video include characters’ faces, a plant in a vase, a shoe, and an automobile tire, each of which is shown within a dashed box to indicate its detection and recognition within the frame.”); Houh, ¶182.) This suggests that Zhao collects similar coordinate location metadata for detected (and recognized) objects. (Houh, ¶182.)

Singer further explains that the object detection metadata may be stored “as timed metadata associated with a video track,” such as a track of a video in QuickTime File Format. (Singer, ¶0023.) “In time-based metadata, each detected-object entry applies to one or more temporally aligned frames of video data.” (Singer, ¶0056.) Figure 13 of Singer, reproduced below, illustrates an example of temporally-aligned, timed metadata for a detected face embedded in a video:

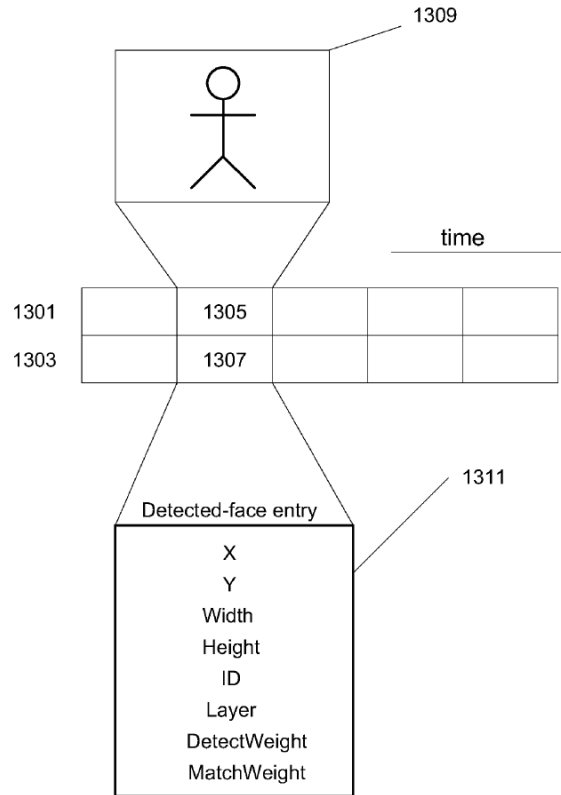


Fig. 13

(Singer, Fig. 13.) Referring to figure 13, Singer explains that “[t]rack **1301** includes frames of image data over time which may be replayed as a video” and “track **1303** includes detected object metadata entries.” (Singer, ¶0068.) “Frame **1305** in track **1301** corresponds to image frame **1309**,” while “metadata entry **1307** corresponds to detected face entry **1311**.” (Singer, ¶0068.) Detected face entry **1311** “includes metadata describing a detected object in frame **1305**, such as the face of the person captured in image data **1309**.” (Singer, ¶0068.)

Singer therefore discloses both “a **coordinate location of the object in the still frame images**”—i.e., X and Y values in a metadata entry for a detected object—

as well as “**a timestamp**”—i.e., information that “temporally aligns” the metadata entry with the associated video frame. (Houh, ¶¶183-184.) Thus, Singer discloses “**embed[ding] metadata about [an] object into [a] video, and wherein the metadata comprises a timestamp and a coordinate location of the object in the still frame images.**” (Houh, ¶184.)

Petitioner notes that, as discussed in Part III.C above, the Petition cites Singer for the same claim limitations in dependent claims 10-11 for which the Examiner cited them during prosecution, and which was not distinguished or disputed by the applicant.

Rationale and Motivation to Combine (Ground 1 Prior Art with Singer).

It would have been obvious to a person of ordinary skill to combine Singer’s teachings regarding embedding metadata about detected objects in video, including “**coordinate location**” and “**timestamp**,” with Zhao’s video processing system (and under the proposed combination). (Houh, ¶186.) The combination would have involved the straightforward incorporation of Singer’s teachings for embedding metadata, and would have predictably resulted in Zhao’s video processing system in which detected metadata was embedded into videos processed by Zhao’s system. (*Id.*) In the proposed combination, the functionality to embed metadata would form part of the “**coordinator**” functionality—for example, as a function that received metadata from demultiplexer nodes and embedded that metadata into the retrieved

video file. (*Id.*) This is consistent with Zhao’s description of storing any “results or outputs” after performing video analysis. (Zhao, ¶0055.) The combination therefore discloses and renders obvious claim 10. (Houh, ¶186.)

Singer is an analogous reference to the ’853 patent, by both relating to video processing and object detection in video images, including face detection, and being reasonably pertinent to problems facing the ’853 patent inventors. (Houh, ¶187.)

A person of ordinary skill would have been motivated to combine. (Houh, ¶188.) A person of ordinary skill would have appreciated the embedding metadata in a video, as opposed to only storing metadata separate from the video, would beneficially enable the video to be transferred and shared without losing the associated metadata. (*Id.*) The recipient of the video could therefore use the metadata without having to perform the recipient’s own processing on the video. (*Id.*) This sharing of videos is suggested in Zhao, which notes “online video viewing” (Zhao, ¶0007), and discussed in Singer, which describes both streaming and downloading a video with the embedded metadata:

At block **607**, the method transmits the image data, including the added metadata track. For example, the image data may be streamed to a remote client. Alternatively, the image data may be downloaded to a remote client and then played locally on the remote client.

(Singer, ¶0040.) Zhao states how “encoded metadata attributes are generally stored in index files” (Zhao, ¶0092), but a person of ordinary skill would have recognized

that this would not diminish the motivation to embed metadata into the video, as doing so serves different purposes. (Houh, ¶188.) In fact, Singer itself states that the metadata is both embedded into video as well as stored in a separate search index. (Singer, ¶0065 (“At block **1205**, the method stores the object detection metadata and the object identification metadata in a searchable index. The searchable index may be an index database.”); Houh, ¶188.)

A person of ordinary skill would have had a reasonable expectation of success. (Houh, ¶189.) Zhao already discloses object detection and metadata creation, and Singer acknowledges the existence of commercially-available video formats (such as Apple Inc.’s QuickTime File Format) capable of storing embedded metadata. (Singer, ¶0023.) The combination therefore would have involved at most the routine application of conventional programming techniques and technologies. (Houh, ¶189.)

2. **Claim 11: “The system of claim 1, wherein the coordinator is configured to generate a metadata stream corresponding to the image, wherein the metadata stream includes one or more timestamps corresponding to the image, and wherein the coordinator is configured to embed the metadata stream in the video.”**

Petitioner largely addressed the limitations of claim 11 in the analysis of claim 10 above. Claim 11 differs from claim 10 mostly in its recitation of a “**metadata stream**” that is embedded in the video. (Houh, ¶190.) The metadata stream, like the metadata in claim 10, includes one or more “**timestamps**.”

Claim 11 would also have been obvious in further view of Singer. Singer discloses embedding metadata into a video that is “streamed to a remote client.” (Singer, ¶0040.) Because the video itself is streamed, a person of ordinary skill would have understood that the embedded metadata track would similarly comprise a “**metadata stream**.” (Houh, ¶191.)

As noted for claim 10 (and in Part III.C above), the Examiner cited Singer during prosecution for disclosing the limitations recited in claims 10 and 11, and was not distinguished or disputed by the applicant.

In the proposed combination, just as with claim 10, the functionality to generate a metadata stream and embed the metadata stream would form part of the “**coordinator**” functionality—for example, as a function that received metadata from demultiplexer nodes and embedded that metadata into the retrieved video file for streaming. (Houh, ¶193.) This is consistent with Zhao’s description of storing any “results or outputs” after performing video analysis. (Zhao, ¶0055.)

The combination with Singer would have predictably resulted in “**the coordinator is configured to generate a metadata stream corresponding to the image, wherein the metadata stream includes one or more timestamps corresponding to the image, and wherein the coordinator is configured to embed the metadata stream in the video**” as claimed. (Houh, ¶194.)

Petition for *Inter Partes* Review of
U.S. Patent No. 11,126,853 B2

VII. CONCLUSION

Petitioner respectfully requests IPR institution.

Dated: May 10, 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 13,818 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 10, 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. 11,126,853 B2**, including all exhibits (**Nos. 1001-1019**) and related documents, are being served via Federal Express on the 10th 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 10, 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