

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
~~Paper 2 – PETITION FOR INTER PARTES REVIEW~~

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner,

v.

CELLULAR SOUTH, INC.,
Patent Owner.

Case IPR2025-00875
Patent 9,940,972 B2
Issue Date: April 10, 2018

Title: VIDEO TO DATA

DECLARATION OF DR. HENRY HOUH
~~PETITION FOR INTER PARTES REVIEW~~

* * *

CB. Ground 1: Claims 1-20 Are Obvious Over Fontana in view of Lau

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

68. For reference, claim 1 recites:

1. A method to generate video data from a video comprising:

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

[b] distributing the image files across a plurality of processors and processing the image files in parallel, wherein processing the image files comprises extracting one or more objects and identifying the one or more objects;

[c] processing the audio files;

[d] converting audio files associated with the video to text;

[e] converting the image files associated with the video to video data;

[f] generating a topical meta-data that describes content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio files;

[g] adding the topical meta-data to the video; and

[h] cross-referencing the text and the video data based on the generated

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

topical meta-data to determine topics;

[i] generating video text based on the cross-referencing, wherein the video text describes content of the video;

[j] generating a text, image, or animation based on the video text; and

[k] placing the text, image, or animation in the video.

(’972, 8:2-25 (Claim 1 (bracketed notation (e.g., “[a]”) added).)

69. The preamble of claim 1 recites “[a] method to generate video data from a video comprising.” Assuming the preamble provides a claim limitation, Fontana discloses it. Fontana discloses ~~a method that generates~~ “generat[ing] video data from a video,” such as meta-data describing the video generated from performing audio processing (for example speech-to-text processing), and/or object recognition of visual content of a video. ~~(Houh~~

~~, ¶¶69-71.)~~70. For example, Fontana teaches “constructing a set of text metadata describing an audio portion of the multimedia content, and generating a set of object metadata describing at least a portion of one or more objects appearing in the multimedia content.” (Fontana ¶0008.³¹) Fontana explains that the “multimedia content can include any type of content containing, for example, one or more of images, video, audio, or a combination thereof... In the context

³¹ Unless otherwise noted, all emphasis added.

of the present disclosure, a robust example of multimedia content is used in which video and audio information are included[.]” (Fontana ¶0042.) ~~Fontana can generate~~ Thus, Fontana’s “text metadata” and “object metadata” can be generated from a video, making them “video data from a video.”

71. Therefore, Fontana discloses “[a] **method to generate video data from a video.**” ~~As~~ And as explained below, the method as claimed is obvious over Fontana and Lau.

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

72. Fontana discloses ~~claim 1[a]~~ “generating audio files and image files from the video” by separately processing the audio and video content of the multimedia content, and further extracting thumbnail images (“**image files**”) from the video. ~~(Houh, ¶¶72-75.)~~

73. For example, Fontana teaches “processing of multimedia content at an audio processing module 606, a video processing module 608, and a video conversion module 610” where “[e]ach of these modules can be executed concurrently (e.g., in parallel), with jobs associated with each module operating on one or more computing systems as defined by a scheduler (e.g., scheduler **406** of Fig. 4).” (Fontana ¶0090.) “The audio processing module 606 is configured to process audio content associated with the multimedia content” while the “video

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

processing module 608 is configured to process the video portion(s) of multimedia content[.]” (Fontana ¶¶0091, 0094.) Fontana further discloses that ~~the audio and video may be separately processed by~~ “these two processes are separate: “the various distributed computing systems described in FIGS. 1-5, above, allow for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel pipelined processing of the data to ensure fast content processing[.]” (Fontana ¶0135.) Fontana thus teaches separately processing the audio and video from the multimedia content. As described below, Fontana teaches that these audio and video processing modules are run on audio and video content extracted ~~(, that is, “generated”)~~ from the multimedia content. A POSA would further appreciate that a video, separated from its audio, is a series of images, such that video files are a form of **“image files.”** ~~(Houh, ¶73.)~~

74. Fontana further teaches that the “audio processing module” can be run on **“audio files”** that are generated from the video. Specifically, Fontana teaches that “[i]f no closed captioning information is present, an audio separation operation 1128 strips, or extracts, the audio from the multimedia content.” (Fontana ¶0164.) A POSA would have understood that stripping or extracting the audio from the multimedia content generates an audio file. ~~(Houh, ¶74.)~~ Fontana additionally teaches that “it is recognized that a number of sources provide speech to text conversion program that approach the conversion differently.” (Fontana ¶0169.)

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

One such program disclosed in Fontana is “large vocabulary continuous speech recognition (LVCSR) engines” that “depend on a language model that includes a vocabulary/dictionary for speech-to-text conversion of audio files.” (Fontana ¶10169.) A POSA would thus have appreciated that Fontana’s disclosures encompass **“generating audio files ... from the video.”** ~~(Houh, ¶74.)~~

75. Fontana additionally teaches that the “video processing module” **generates ~~another~~ an additional** set of “**image files**” **as in the form of** thumbnail images derived from the video. Fontana teaches: “In the embodiment shown the video processing module 608 includes a thumbnail extraction module[.]” that “is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest In some embodiments, the thumbnail extraction module 618 generates a series of thumbnails representing scenes throughout the multimedia content.” (Fontana ¶10095.) A POSA would appreciate that a thumbnail is an image. ~~(Houh, ¶75.)~~ Such images were commonly stored as image files in 2013, such as in gif, tiff, jpg, or tiff formats, for example, which are common formats for graphical image files. ~~(Id.)~~ A POSA would thus have appreciated that by extracting thumbnail images from the video, Fontana’s disclosures satisfy **“generating ... image files from the video.”** ~~(Id.)~~

- (b) **“distributing the image files across a plurality of processors and processing the image files in parallel, wherein processing the image files comprises**

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

extracting one or more objects and identifying the one or more objects;” (Claim 1[b])

76. Fontana discloses or renders obvious ~~claim 1[b].~~ ~~(Houh, ¶¶76-86.)~~ “distributing the image files across a plurality of processors and processing the image files in parallel, wherein processing the image files comprises extracting one or more objects and identifying the one or more objects.”

77. Fontana ~~implements its method using~~ “discloses implementing its method in a distributed computing environment: “The multimedia processing system 104, although represented by a single computing system, is in preferred embodiments a plurality of distributed computing systems[.]” (Fontana ¶0045.)

A POSA would have understood that a distributed computer system is a form of “parallel” processing as required by the '972 patent—by distributing tasks across a plurality of processors, each processor ~~handles~~ is able to handle a portion of the tasks, resulting in parallel processing of the set of requests. ~~(Houh, ¶77.)~~ ~~Fontana’s~~ As Fontana explains, its distributed computing system allows tasks, including “video, image or audio algorithms[.]” to be “separated and performed across multiple computing systems in parallel.” (Fontana ¶0053; *see also id.* ¶0090.) As further detailed below, a POSA would have understood that Fontana’s distributed computing system includes both parallel processing of image files separate from audio files, and parallel processing for image-processing tasks, resulting in

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

processing multiple image files in parallel. (~~Houh, ¶77.~~)—Therefore, Fontana discloses that its distributed computing system distributes image processing requests, such that it “**distribut[es] the image files across a plurality of processors and process[es] the image files in parallel.**”

78. For example, Fontana teaches:

The network 200 can, in certain embodiments, correspond to an architecture underlying the multimedia processing system 104 of FIG. 1, for example in a cloud-based or other distributed computing environment. The network 200 includes, in the embodiment shown, a workflow server 202 interconnected to an integration framework 204 and a storage network 206. The integration framework 204 provides interconnectivity and data sharing among a plurality of computing systems, such that the computing systems can share workloads, messages, and other tasks. The integration framework 204 can be connected to any of a plurality of differing types of computing systems 208 capable of sharing workloads; in the embodiment shown, various shared computing systems are illustrated including workstations 208a, grid computing systems 208b, compute clusters 208c, data resources 208d, and one or more high performance computing systems 208e.

(Fontana ¶0048.)

79. Fontana explains that this workflow server “receives inbound data processing requests, for example from a content provider (as further discussed below) and distributes one or more portions of jobs associated with each data

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

processing request to the integration framework 204 and the storage network 206.”
(Fontana ¶0050.) The ~~result~~ “allow[s] resulting system “allows creation of pipelined data processing systems within a distributed computing environment, allowing computationally intensive jobs (e.g., video and audio content processing) to be distributed across a number of computing systems.” (*Id.*) ~~Thus~~In other words, the workflow server divides computationally intensive jobs, such as video processing (including image processing, as previously discussed), into portions, and distributes those portions across multiple computing systems. A POSA would have appreciated that the result of this system is that these processing tasks, such as image processing, would be divided up and performed in parallel. ~~(Houh, ¶¶78-79.)~~

80. Fontana further teaches that this distributed and parallel processing of image files can be carried out by a plurality of servers. Fontana teaches that its method is implemented using “a plurality of computing systems, illustrated as servers 302a-c. The servers 302a-c are communicatively interconnected, ... share a distributed memory cache 306, and are each capable of accessing a shared cache of memory[.]” (Fontana ¶0052.) These servers “are interfaced to inbound work ... for coordination and communication of data for processing.” (¶*Id.*) Fontana further teaches that “one or more of the servers 302a-c can include specific graphical processing units for processing lower level video, image or audio algorithms” ~~and are~~ “.... The servers 302a-c are configured to share processing jobs, such that tasks

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

can be performed by one or more of the computing systems, or separated and performed across multiple computing systems in parallel.” (Fontana ¶0053.) As a result, “any of those computing systems can perform all or a portion of a processing job as defined by a scheduling algorithm, allowing multimedia content to be processed efficiently when necessary.” (Fontana ¶0055.) A POSA would have appreciated that ~~the described~~this system of servers ~~that, which~~ share processing jobs ~~carry~~and workloads by, for example, separating and performing the jobs in parallel, carries out parallel processing. ~~(Houh, ¶80.), including for processing image files.~~ In the context of image processing, it would have been obvious to a POSA that image processing within Fontana’s distributed computing system could **process the image files in parallel** by, for example, distributing the image files for processing across multiple servers. ~~(Id.)~~

81. Indeed, Fontana further discloses processing images files to **“extract[] one or more objects and identify[] the one or more objects”** as part of **“processing the image files in parallel.”** Fontana explains that ~~its system~~ “includes a scheduler 406” that “in general receives tasks from the frontend 402 as defined by content providers, for example indicating that multimedia content should be processed to generate one or more objects of interest.” ~~(, to create a transcript of the multimedia content, or other typically computationally-intensive functions.”~~ Fontana ¶0061.) “The scheduler 406 receives and routes the content and processing

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

requests to the desired computing systems within the grid 408”~~and~~; the scheduler generally provides the ability to equally distribute resources to all jobs that are running at once[.]” (~~¶~~*Id.*) A POSA would have appreciated these disclosures as teaching that the scheduler is what distributes tasks among servers for parallel processing, and that Fontana’s scheduler distributes tasks to process multimedia content to generate one or more objects of interest. (~~Houh, ¶81.~~)

82. As discussed above for claim 1 [a], a POSA would have understood that Fontana’s image processing algorithm generates image files at least in the form of thumbnail images. Fontana further discloses “**extracting one or more objects and identifying the one or more objects**” as part of processing ~~these~~the image files. The objects to be identified by Fontana can be either provided to the system or automatically identified by the system~~.~~

83. Fontana explains:

In the embodiment shown the video processing module 608 includes a thumbnail extraction module 618 and an objects of interest module 620. The thumbnail extraction module 618 is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest --- (for example a first frame, a last frame, and immediately following major scene or sound changes in the content). In some embodiments, the thumbnail extraction module 618 generates a series of thumbnails representing scenes throughout the multimedia content. The objects of interest module 620 generates one

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

or more objects of interest as defined in metadata to be associated with the multimedia content. In various embodiments, the objects of interest module 620 can accommodate input from content providers to identify the objects of interest, or can at least partially automatically identify at least candidate objects of interest for confirmation by a user.

(Fontana ¶0095)

84. Fontana explains that the identification of “objects of interest” can occur automatically within the thumbnail images; specifically, “boundaries of a number of candidate objects of interest could be automatically detected within one or more thumbnails[.]”

(Fontana ¶0142.) Fontana further teaches “**extracting one or more objects**” by “generat[ing] a ‘filmstrip’ which is a strip of thumbnails containing ‘objects of interest’ from the video. These objects of interest can be items, people, or conditions in the video that the viewer may be interested in[.]”

(Fontana ¶0148.)

85. Fontana additionally teaches “**identifying the one or more objects.**” Fontana discloses, for example, that objects can be identified using OpenCV or other tools ~~that~~, all of which would have been familiar to a POSA:

After the content is received, a candidate object generation operation 906 generates candidate objects of interest from the multimedia content... The candidate object generation module can be performed by any of a number of object recognition programs, including computer vision programs. Example computer vision tools include OpenCV, which is a library of motion tracking, facial recognition, gesture

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

recognition, object identification, segmentation, and calibration tools. Other tools, such as MatLab or scale-invariant feature transform (SIFT) algorithms could be included in the object detection process as well.

(Fontana ¶0139; ~~Houh ¶85.~~)

86. Fontana additionally teaches that objects could also be identified through “a neural network or other learning model to acquire knowledge of objects typically recognized or identified by users as objects of interest.” (Fontana ¶0140.) Fontana provides as examples of such neural networks those developed by Numenta, Inc.; Vidient Systems, Inc.; and Behavioral Recognition Systems, Inc. (Id.)

(c) “processing the audio files;” (Claim 1[c])

87. Fontana discloses “**processing the audio files,**” such as processing associated with performing speech-to-text recognition. (~~Houh ¶87.~~) Fontana discloses an “audio processing module 606” that “is configured to process audio content associated with the multimedia content. In certain embodiments, the audio processing module 606 is configured to generate a full text transcript of the audio included in the multimedia content, to allow content consumers to search and review transcripts for the appearance of desired items.” (Fontana ¶0091.) As discussed for claim 1[a] above, Fontana further discloses that this speech-to-text recognition may be performed on **audio files**.

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

88. As noted for claim 1[c], Fontana discloses “**converting audio files associated with the video to text,**” such as performing speech-to-text recognition to convert audio from the multimedia content (which may be a video, as discussed for the preamble) to text. ~~As~~And as discussed for claim 1[a], this speech-to-text recognition may be performed on **audio files.** ~~(Houh ¶88.)~~

(e) “**converting the image files associated with the video to video data;**” (Claim 1[e])

89. Fontana discloses “**converting the image files associated with the video to video data,**” such as performing facial and/or object recognition to generate data and meta-data about the video content from the extracted thumbnail images. ~~(Houh ¶¶89-90.)~~

90. As I discussed for claim 1[b] above, Fontana discloses “a thumbnail extraction module 618 and an objects of interest module 620.” ~~that~~ “(Fontana ¶0095.) The “objects of interest module 620 generates one or more objects of interest.” (Fontana ¶0095 as defined in metadata to be associated with the multimedia content.” (Id.) Fontana also discloses a “candidate object generation operation” that “can generate a number of candidate objects of interest defined by the content provider.” (Fontana ¶0141.) Additionally, “boundaries of a number of candidate objects of interest could be automatically detected within one or more thumbnails[.]” (Fontana ¶0142.) These modules are discussed ~~above~~in greater detail

for claim 1[b]. A POSA would appreciate that the data generated by the objects of interest module and the candidate object generation operation are “**video data**” as they are data about the video content. (*See, e.g.*, ’972, 1:10-11 (“The present invention relates to a method and a system for generating various and useful data from videos.”); *id.*, 1:27-29 (“The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.”); Houh ¶¶90.)

- (f) **“generating a topical meta-data that describes content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio files;”** (Claim 1[f])

91. Fontana discloses “**generating a topical meta-data that describes content of the video**” in the form of contextual information, an indexed transcript, keywords, and other meta-data derived from Fontana’s audio and video analyses. (Houh ¶¶91-98.)

92. The ’972 patent explains topical meta-data as follows:

Further, the engine applies the topical meta-data to the original full video. The image topics can be stored as topics for the entire video or each image segment. The topic generation process can be repeated for all identifiable symbols in a video in a distributed process. The outcome would be several topical descriptors of the content within a video. An example of the aggregate information that would be derived using the above example would be understanding that the video presented a dog,

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

which was jumping, on the beach, with people, by a resort.

(’972, 4:33-44.) Thus, the topical meta-data can include at least semantic understanding of identified objects (e.g., dog, beach, people, resort) and events (e.g., jumping). ~~(Houh ¶92.)~~

93. Fontana generates metadata describing both the objects recognized in the images and the audio processing. ~~For objects, Fontana discloses “generat[ing] object metadata~~As to object recognition, Fontana explains that a “plurality of processing operations occur to generate object metadata, text metadata, and format the received multimedia content, for example to generate and store the various types of content-specific metadata described above.” (Fontana ¶120.) ~~Likewise Thus,~~ “an object metadata operation 806 generates object metadata corresponding to” ~~inter alia,~~ “ information about the content overall, as well as objects appearing in or mentioned in the multimedia content. For example, the object metadata ~~can~~ can define the overall genre, title, producer, creation date, length or other characteristics of the multimedia content, but can also define people or objects appearing in the content as well.” (*Id.*) A POSA would understand the object metadata to correspond to topical meta-data, in the form of identifying people and objects appearing in the video. ~~(Houh ¶93.)~~

94. A POSA would further understand that Fontana discloses identifying topical meta-data in the form of what the ’972 patent specification refers to as

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

“events,” ~~namely~~that is, actions performed by the identified people or objects. ~~(Houh ¶94.)~~Specifically, Fontana teaches ~~implementing~~that “[t]he candidate object generation module”~~using, for example,~~ “can be performed by any of a number of object recognition programs, including computer vision programs. Example computer vision tools include OpenCV, which is a library of motion tracking, facial recognition, gesture recognition, object identification, segmentation, and calibration tools.” (Fontana ¶0139.) Motions and gestures are events, such as “jumping,” that the ’972 patent specification identifies as topical meta-data. (*See* ’972, 4:20-23, 4:41-44.)

95. For audio processing, “[a] text metadata operation 808 defines text metadata associated with the multimedia content,”~~which~~. The text metadata can take any of a number of forms, and can include a transcript of audio data included in the multimedia content, as well as additional textual information that a content presenter would like to display alongside the streamed multimedia content, such as additional contextual information, advertisements, or hyperlinks to other websites or content.” (Fontana ¶0121.) Notably, the text metadata can include both a transcript of the audio data, as well as additional text metadata for “additional contextual information.” (*Id.*) Furthermore, “[t]he transcript can be indexed ... to allow content consumers to search the spoken text transcript, as well as other descriptive information related to the multimedia content.” (*Id.*) As a result, the text

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

metadata can include additional text-based descriptive and contextual information about the video content, beyond the transcript of the audio. ~~(Houh ¶95.)~~

96. For example, Fontana discloses identifying keywords in the transcript, which enables searching of the multimedia content: “The text index information 630 can be used to provide a corresponding transcript ~~...~~alongside playback of multimedia content, or can be used to provide keyword searchability of the multimedia content.” (Fontana ¶10107.) Fontana describes example keyword data for facilitating searching the multimedia content as follows:

FIG. 7I illustrates example keyword data that can be used in association with particular content to facilitate searching of that content. In certain embodiments, the keyword data 706 can be used as a substitute for the text information 630, or can be used to reference a particular location within the text information to allow searching of content or metadata describing the content. ~~...~~In the embodiment shown, the keyword data 706 includes an identifier of the multimedia content as well as the keyword or keywords associated with that content. Other information can be included in the keyword data as well (e.g., links to a particular location within the multimedia content, or other associated keywords, etc.). In certain embodiments, the keyword data 706 can be made available to external search engines, to allow the content or portions of the content to be made available for search access by search engines that are remote from and unaffiliated with the systems and methods described herein.

[Ex. 1002 – DECLARATION OF DR. HENRY HOUH](#)
~~Paper 2 – PETITION FOR INTER PARTES REVIEW~~

(Fontana ¶0114.)

97. A POSA would have appreciated that Fontana’s above-described object metadata and text metadata (including both transcript and [additional descriptive or contextual information, such as](#) keywords) are all “**topical meta-data that describes content of the video.**” ~~(Houh ¶¶96-97.)~~

98. Furthermore, Fontana’s contextual information and meta-data, which correspond to the “**topical meta-data**” of claim 1, are generated “**by deriving semantic information from the identification of the one or more objects and semantic information from the audio files.**” ~~(Houh ¶98.)~~ A POSA would understand “**semantic information**” to refer to information conveying or associated with the meaning of content. ~~(Id.)~~ Fontana’s topical meta-data is generated by deriving semantic information from the object identification and audio processing because it includes words and concepts ascribing meaning to the audio and video content. For example, a POSA would recognize keywords identified from a text-to-speech transcript as semantic information because they convey meaning about the audio. (See ~~id.~~; ~~see also~~, e.g., Fontana ¶0107 (“The text index information 630 ~~...~~[can be used to provide a corresponding transcript alongside playback of multimedia content, or](#) can be used to provide keyword searchability of the multimedia content.”).) Likewise, a POSA would recognize identifications of people or objects as semantic information about the meaning of a video or image. ~~(Houh ¶98.)~~

(g) “adding the topical meta-data to the video; and”
(Claim 1[g])

99. Fontana discloses or renders obvious “**adding the topical meta-data to the video.**” (~~Houh ¶¶99-106.~~) For example, Fontana teaches that “specific start and end times can be defined, as associated with specific segments of the transcription text. In this way, the transcript could be linked, portion by portion, to the multimedia content based on the time at which the transcribed words are played in the content.” (Fontana ¶0107.) A POSA would have appreciated that this specific linking of transcript to video “**add[s] the topical meta-data to the video.**” (~~Houh ¶99.~~) The Indeed, the ’972 patent specification discloses a similar linking of topical meta-data to the video: “Data generated from an image and/or from audio transcription can be time stamped, for example, according to when it appeared, was heard, and/or according to the video frame from which it was pulled.” (’972, 4:28-31.)

100. Similarly, Fontana teaches that “keyword data 706 ~~...~~ can be used as a substitute for the text information 630, or can be used to reference a particular location within the text information~~...~~ to allow searching of content or metadata describing the content. In the embodiment shown, the keyword data 706 includes an identifier of the multimedia content as well as the keyword or keywords associated with that content. Other information can be included in the keyword data as well

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

(e.g., links to a particular location within the multimedia content, or other associated keywords, etc.).” (Fontana ¶0114.)

101. Furthermore, Fontana teaches that “text index information 630 can be used to provide a corresponding transcript alongside playback of multimedia content, or can be used to provide keyword searchability of the multimedia content.” (Fontana ¶0107.) A POSA would have appreciated that ~~this to~~ providing a transcript alongside playback of a video was a form of “**adding the topical meta-data to the video.**” ~~(Houh ¶101.)~~

102. Fontana additionally teaches that the topical meta-data can be presented concurrently with the video content, which is yet another form of “**adding the topical meta-data to the video.**” ~~Fontana’s~~ As Fontana explains, the “text metadata,” which is **topical meta-data**, “can take any of a number of forms, and can include ... additional textual information that a content presenter would like to display alongside the streamed multimedia content, such as additional contextual information, advertisements, or hyperlinks to other websites or content.” (Fontana ¶0121; ~~see also Fontana ¶0147 (example “.)~~ Additionally, “[a]n action definition operation 914 allows a user to define one or more actions associated with each object of interest ... identified in the multimedia content. Any of a number of different types of actions can be defined. Example actions include display of contextual information identifying the object”), as well as including click through actions such as a

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

hyperlink to related content, or other sections of the same piece of content.” (Fontana ¶0147.) A POSA would have appreciated that presenting text metadata alongside the video **“add[s] the topical meta-data to the video.”** ~~(Houh ¶102.)~~

103. Additionally, Fontana discloses a “container operation 810” that “applies a container to the received multimedia content.” (Fontana ¶0122; ~~see also~~ .) This “container operation 1304 can be performed by the multimedia processing systems of the present disclosure, with the container and associated metadata being stored either by the multimedia processing systems or managed by the content provider.” (Fontana ¶0176.) Such containers were well-known to a POSA at the time, as they were a common way to store the various components of multimedia content such as videos. ~~(Houh ¶103.)~~ Fontana discloses ~~several~~ some of these well-known container formats, such as “an Adobe Flash format” as well as “HTML5, Microsoft Silverlight, or other formats[.]” (Fontana ¶0122.)

104. A POSA would have known that container files such as those disclosed in Fontana routinely store metadata as well as multimedia content, and thus would have found it obvious to apply Fontana’s container operation to both the multimedia content and its associated metadata. ~~(Houh ¶104.)~~ For example, Adobe Flash format, which Fontana discloses, was known to include storage of metadata. (See, e.g., EX1008, Adobe Flash Video File Format Specification, https://rtmp.veriskope.com/pdf/video_file_format_spec_v10_1.pdf (describing

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

Adobe Flash Format metadata options.) Furthermore, MP3 was a well-known container file format for audio, and a POSA would have known that “MP3 files are capable of storing a certain amount of ‘meta-data’—extra information about each file—inside the file itself,” ~~which “... These tags will be inserted automatically by most tools as you rip and encode, or~~ can be added or edited later on, often directly through your MP3 player’s interface.” (EX1009, MP3: The Definitive Guide, p. 6; ~~see also id.~~, pp. 44-45, 105, ~~114, 116, 363-64, 369; see also, e.g., EX1008 (describing Adobe Flash Format metadata options); EX1010 ¶0040 (describing a method to add a metadata track to video data); Houh ¶104 (ID3 tags are “the extra space in MP3 files that let you store ‘meta data’ about a file”)~~, 114 (“the ID3v2 specification allows for a huge array of meta-data storage capabilities in MP3 files”), 116 (“Every MP3 file has the ability to store ‘meta-data’ related to the track in the file itself, in the form of what are known as ‘ID3 tags.’”), 363-64, 369.) As another example, U.S. Patent Application Publication No. 2011/0305394 (“Singer”) discloses a method that “adds a metadata track to the image data (e.g., video data)” where the “added track includes the generated face detection metadata.” (EX1010, Singer ¶0040.)

105. Disclosures in Fontana would have confirmed to a POSA that the container could include both the video and its topical meta-data. As noted above, Fontana explains that the “container operation 1304 can be performed by the

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

multimedia processing systems of the present disclosure, with the container and associated metadata being stored ... by the multimedia processing systems[.]” (Fontana ¶0176.) As a further example, Fontana teaches “[a] storage operation 812 stores the content and associated metadata for use.” (Fontana ¶0123; ~~see also Fontana ¶0066 (describing storage of metadata).~~) A POSA would have appreciated that a container operation was one such storage operation. ~~(Houh ¶¶105-106.)~~

106. Likewise, Fontana teaches that “[t]he data storage 410 can be configured to store any of a number of different types of data, including the received multimedia content and data associated therewith. In certain embodiments, the data storage 410 includes a set of metadata associated with each piece of multimedia content processed by the computing grid 408, for example as generated by processing the multimedia content.” (Fontana ¶0066.)

(h) “cross-referencing the text and the video data based on the generated topical meta-data to determine topics;” (Claim 1[h])

~~Claim 1[h] is obvious over Fontana in view of Lau. (Houh ¶¶107-120.)~~

107. A POSA would have found it obvious to combine Fontana with Lau to “cross-referenc[e] the text and the video data based on the generated topical meta-data to determine topics.”

108. In the combination of Fontana with Lau, the topical meta-data generated by Fontana would be searched using the method of Lau to determine a

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

relevant context for the video and ultimately identify advertisements for display at specific times during a video. This search process **cross-references the text and the video data** by matching advertisements using keywords from **the text** (the results of the speech-to-text recognition) and content from the **video data** (metadata on the identified objects of interest) **based on the generated topical meta-data** (i.e., by incorporating context and other metadata, *see* claim 1 [f]). ~~(Houh ¶108.)~~

109. Lau discloses performing a search to determine advertisements whose content and/or topics match the content and/or topics for portions of video content. The search ~~counts~~ involves counting the number of keyword and/or concept matches for a search term near a particular time in the video content. A POSA would have found it obvious to combine Lau and Fontana by using the **topical meta-data** generated by Fontana (*see* claims 1[f] and 1[g]) as well as **the text and the video data** (*see* claims 1[d] and 1[e]) as keywords and concepts searched by Lau for purposes of identifying ads or ad units to match to the video based on context or subject matter. Each of the ads, ad units, and context or subject matter of the video is a “**topic**.” As discussed for claim 1[f], the **topical meta-data** (various contextual information and metadata) is derived from **the text and the video data**, such that searching the topical meta-data **cross-references the text and the video data**, and any such search is **based on the generated topical meta-data**. A POSA would have further found it obvious to use the text and the video data as part of Lau’s matching

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

search, which further **cross-references the text and the video data.** (Houh ¶109.)

As a result, this combination “cross-referenc[es] the text and the video data based on the generated topical meta-data to determine topics.”

110. Specifically, Lau teaches matching keywords and concepts in a video to ads based on context or subject matter: “Correlation engine 202 may relate a unique content ID with a time series of keywords and concepts (that advertisers have purchased), and in turn, relate the keywords and concepts to ads submitted by advertisers.” (Lau ¶0051.) This matching aggregates together the different metadata types and uses the resulting aggregated information to match the content to an ad: “for each ad, correlation engine 202 finds candidate content that may be relevant. This is done by searching for content in the index to match the keywords, categories, and concepts associated with the ad to information in the content.” (Lau ¶0056.) A POSA would have found it obvious to use the results of Fontana’s speech-to-text recognition (**the text**) to generate keywords and other “information in the content” for Lau’s search method. Similarly, a POSA would have found it obvious to use the results of Fontana’s object recognition module (**the video data**) as concepts and “other information” for Lau. It in the content for Lau’s search method. And because Lau’s correlation engine searches “to match the keywords, categories, and concepts,” it would further be obvious to perform Lau’s search with reference to the contextual information derived from Fontana, which is the **topical meta-data** and is

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

additionally “information in the content.” (See Lau ¶0056; Houh ¶110.) Lau teaches that the search to match advertisements to video content may ~~combine~~ be performed by combining the results of different types of analyses, such as text with images as well as conceptual ideas like context: “Advertisers may buy correlation information, such as keywords, phrases or concepts, either through a bidding process or some other means, and submit their ads and related information to correlation engine 202 through correlation assistant 214. ... The phrases may be any combination of words and other information, such as symbols, images, etc. The concepts may be a conceptual idea of something.” (Lau ¶0046.) Thus, a POSA would have found it obvious that in the combination of Fontana with Lau, correlating advertisements to the video could **cross-reference the text** (speech-to-text transcript and any derived keywords) **and the video data** (the identified objects of interest) **based on the generated topical meta-data** (additional contextual information). (Houh ¶110.)

111. This contextually aware keyword and topic searching aligns with the ’972 specification’s teaching regarding cross-referencing:

At 160, the topics generated from an image or a frame and the topics extracted from audio can be combined. The text can be cross-referenced, and topics common to both texts would be given additional weights. ~~...~~ At 170, the video-to-text engine generates video text, such as text describing the content of the video, using the result of the

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

combined texts and cross reference. For example, key words indicating topic and semantic that appear in both texts can be selected or emphasized.

(’972, 5:46-54.)

112. Returning to Lau, Lau discloses using its method to **determine topics**, that is, subject matter for the ads and ad units. For each ad, Lau identifies a set of candidate locations within the video, then scores each based on the strength of the match ~~based on the cross-referencing of keyword/concept matches:~~

For each piece of candidate content associated with an ad, correlation engine 202 determines candidate times where the content may be relevant to the ad. Correlation engine 202 locates the times where the keywords and concepts match. For each candidate time, correlation engine 202 creates an “ad anchor” holding the score for the match. The score may be a linear combination of the following weights:

...

1. Probability of the keyword/concept match pulled from the recognition lattice.
2. Concentration of the match—the more keywords/concepts for the ad matches near the time, the higher the score. One embodiment of this score may be a count of the number of matches within a certain window of the current time....

(Lau ¶¶0057–0059.)

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

113. As I discuss in greater detail for claim 1[i] below, Lau additionally discloses that “[a]n advertisement may be broken into ad units” that are each “a subset of a larger advertisement. ... Thus, advertiser system 106 is not restricted to just serving an entire advertisement. Rather, the most relevant pieces of the advertisement may be selected from the matrix of ad units.” (Lau ¶0023.) As a result, “[c]orrelation engine 202, when determining the advertisement, may determine one or more ad units that correlate to the subject matter,” ~~to combine and~~ “the 330 ad units may be combined and presented to the user” as the final advertisement. (Lau ¶0038.)

~~Searching~~114. Each of searching for candidate video location matches for ads or ad units based on the subject matter, and assigning weights to those ~~candidates,~~ ~~each~~candidate locations, involves **cross-referencing the text and the video data based on the generated topical meta-data** to determine ad or ad unit matches based on subject matter or context (“**topics**”). ~~(Houh ¶114.)~~

115. Rationale and Motivation to Combine (Fontana with Lau): A POSA would have been motivated to combine Fontana and Lau in order to use Lau’s ~~method for placing~~disclosures for how to place advertisements within a video to carry out Fontana’s disclosures ~~of linking that~~ advertisements can be linked to video content, and would have had a reasonable expectation of success in the combination. ~~(Houh ¶¶115-120.)~~ Fontana and Lau are analogous references to the ’972 patent: ~~all~~

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

~~are in.~~ Like the '972 patent, both Fontana and Lau are addressed to the field of processing and generating multimedia content, including video, and both specifically address ~~searching~~how to search within such multimedia content. ('972, 1:27-29 (“The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.”); ~~see also~~ '972, Abstract (“The text and the video content can be cross-referenced with the video.”); Fontana ¶0001 (“[T]he present disclosure relates to systems and methods for processing and delivery of multimedia content.”); Fontana ¶0041 (“In general, the present disclosure relates to methods and systems for receipt, processing, and delivery of multimedia content, as well as enrichment of multimedia content for enhanced search and delivery.”); Lau ¶0002 (“Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content.”); Lau ¶0006 (“In one embodiment, an advertisement is matched to subject matter in a portion of rich media content, such as digital video”); Lau ¶0053 (“The ads may be correlated to content in different ways. In one embodiment, keywords may be associated with each ad. Content may be searched to determine if the content includes the keywords.”).) Both further disclose linking advertisements to the multimedia content. ('972, 1:57-59 (“An advertisement can be placed at a specific time in the video based on the video content and/or section symbol of a video image.”); Fontana ¶0117 (“FIG. 7M illustrates example advertisement data 716

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

~~---~~that can be used in association with multimedia content, to link one or more advertisements with multimedia content during playback.”); Lau ¶0002 (“Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content.”.)

116. Fontana expressly discloses that advertisement data could be ~~used~~ “linked to the multimedia content: “FIG. 7M illustrates example advertisement data 716 that can be used in association with multimedia content, to link one or more advertisements with multimedia content during playback.” (Fontana ¶0117.) But Fontana ~~further teaches that~~does not teach how to identify advertisements to link with multimedia content; rather, “the matching of advertisements and content occurs based on a decision process separate from the content delivery system of the present disclosure.” (Fontana ¶0117.) A POSA thus would have been motivated to combine Fontana with an advertisement-matching method, such as that disclosed by Lau, in order to practice Fontana’s disclosures of linking ~~advertisements with~~advertisement information with the multimedia content. ~~—(Houh ¶116.)~~

117. A POSA would have had a reasonable expectation of success in combining Fontana and Lau. It would have been obvious to a POSA that the text and object meta-data generated by Fontana could be used as the keywords/content to search for matching advertisements to video content according to Lau. ~~(Houh ¶117.)~~ Fontana makes clear that its keywords are suitable for searching within the

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

video, including with search engines outside of Fontana’s disclosures. ~~(See, e.g.,~~
~~Fontana ¶0121~~ ~~(—a POSA would have found it obvious that Lau’s method of~~
~~searching for ad matches~~ within its search engine is one such application of keyword-
~~based searching within a search engine taught by Fontana. Fontana specifically~~
~~discloses that its metadata can include an indexed transcript to allow “enable~~
~~searching, and can be linked to particular start and end times of the video: “The~~
~~transcript can be indexed, as described below, to allow content consumers to~~
~~search[ing] the spoken text transcript”);~~, ~~as well as other descriptive information~~
~~related to the multimedia content” (Fontana ¶0121) and “specific start and end times~~
~~can be defined, as associated with specific segments of the transcription text. In this~~
~~way, the transcript could be linked, portion by portion, to the multimedia content~~
~~based on the time at which the transcribed words are played in the content” (Fontana~~
~~¶0107).~~

118. ~~Fontana explicitly contemplates that its “text index “information 630 ...~~
~~can be used~~ to provide keyword searchability of the multimedia content.”); ~~(Fontana~~
~~¶0114 (keywords “0107.)~~ Fontana teaches that the content and metadata it generates
~~can be searched through the use of keywords: “The enhanced multimedia content~~
~~described in the present disclosure generally relates to multimedia content with~~
~~associated interactive features, for example ... associated transcript information~~
~~linked to the multimedia content for keyword searching[.]” (Fontana ¶0042.) “FIG.~~

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

7I illustrates example keyword data that can be used in association with particular content to facilitate searching of that content. In certain embodiments, the keyword data 706 can be used as a substitute for the text information 630, or can be used to reference a particular location within the text information to allow searching of content or metadata describing the content.” including (Fontana ¶0114.)

119. Fontana then teaches a “content request operation” that “receives a request related to multimedia content. The specific type of request received in the content request operation 814 can take a number of forms, such as a search query related to keywords appearing in one or more fields of metadata associated with the content (e.g., titles, authors, producers, genre, etc.) or in the transcript or other text associated with one or more pieces of content.” (Fontana ¶0126.) Fontana responds to this request through a “provide metadata operation” that “provides metadata (and optionally the multimedia content) in response to the request. The provide metadata operation 816 selects at least a portion of the metadata associated with the content (e.g., including definitions of objects of interest, events, transcript information, position information, etc.) for inclusion with the content during playback.” (Fontana ¶0127.)

~~by~~ 120. Fontana further expressly contemplates that search engines, other than those expressly disclosed by Fontana, could also be used to search the keyword data generated by Fontana: “In certain embodiments, the keyword data 706

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

can be made available to external search engines..., to allow the content or portions of the content to be made available for search access by search engines that are remote from and unaffiliated with the systems and methods described herein.”); (Fontana ¶¶0042; ~~Fontana ¶¶0126-0127 (keyword searching of metadata, including by~~ ¶0114; see also Fontana ¶0127 (“In certain embodiments, the preference information can be provided to a remote decision engine”).) ~~A POSA would have found it obvious that Lau’s method of searching for ad matches is one such keyword-based search engine. (Houh ¶¶117-120. that can then indicate a particular type, genre, or other grouping of enhancements to include with the multimedia content.”)~~.) Thus, a POSA would have appreciated that the **topical meta-data** generated by Fontana could be used as the keyword/content information for the video in Lau’s search method, and would have had a reasonable expectation of success in the combination, as Fontana indeed encourages such combinations. ~~(Houh ¶¶115-120.)~~

- (i) **“generating video text based on the cross-referencing, wherein the video text describes content of the video;” (Claim 1[i])**

121. The combination of Fontana with Lau ~~renders—obvious~~described above—using the topical metadata generated by Fontana to implement Lau’s method of matching advertisements to video content—“generat[es] video text based on the cross-referencing, wherein the video text describes content of the video.” The combination of Fontana and Lau is described for claim 1[h] above, including a

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

motivation to combine and reasonable expectation of success. As relates to claim 1[i]-, Lau discloses ~~matching~~that ads are matched to multimedia content through the cross-referencing by determining a context or subject matter (which I will refer to for simplicity, as a “context”) (“**video text**”) that “**describes content of the video**” and using that context to determine “ad units” from which the ads are generated. Additionally, Fontana discloses storing “advertisement data” including “topics, keywords, or content” that can be matched or related to the video content as additional “**video text.**” (~~Houh ¶¶121-127.~~)

122. As discussed for claim 1[h], the result of the cross-referencing are the ads, ad units, and context/subject matter of the video. Lau discloses “ad units” that are pieces of a larger advertisement, where each ad unit is associated with a concept. “An advertisement may be broken into ad units. An ad unit may be a subset of a larger advertisement. ... Each ad unit may be associated with a concept. The ad units may be selected individually to form an advertisement.” (Lau ¶0023.)

123. Lau explains that there are “[d]ifferent ways of creating an ad unit[.]” (Lau ¶0027.) As examples, an “ad unit may be created by taking a static ad and augmenting the unit with an advertiser-specified message dependent on **context** and **keywords.**” (Lau ¶0027.) Furthermore, A POSA would have appreciated that ~~this includes determining~~identifying an “advertiser-specified message dependent on context and keywords” additionally includes a determination of “context,” which

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

is a way of describing what a video is about beyond just the keywords. In other words, a POSA would have appreciated that “context” as used in Lau can be a form of **video text based on the cross-referencing ~~in claim 1[h] and~~ (derived from the text and video data, which are themselves the result of the audio and image processing)** that **describes content of the video.**—(Houh ¶123.)

124. Lau ~~provides~~ illustrates the concept of “ad units” with a BMW ~~ad~~ example that makes clear that Lau’s “context” can be video text based on the cross-referencing that describes content of the video:

Correlation engine 202, when determining the advertisement, may determine one or more ad units that correlate to the subject matter. For example, based on one or more keywords, ad units from the ad matrix are determined. ~~...~~ The ad units are then combined into an advertisement that is correlated to the subject matter. One example of this is BMW may provide a general ad unit for their logo and have a different ad unit for different models, such as the 330 model, 530 model, etc. ~~...~~ The logo unit and each of the model units can be combined at runtime based on the context of the content. If the content talks about the 330 model then the logo and the 330 ad units may be combined and presented to the user.

(Lau ¶0038.)

125. In this example, Lau determines that the “subject matter” of the video is the BMW 330 model. This subject matter (or context) is **video text** (BMW 330

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

model) that is **based on the cross-referencing** (identified “based on one or more keywords,” which can search text and image metadata (*see* Lau ¶0046), as the model appearing in the video content) and that **describes content of the video** (“the content talks about the 330 model”). (*See* Lau ¶0038.) A POSA would therefore have recognized that Lau teaches **generating video text based on the cross-referencing, wherein the video text describes content of the video.** ~~(Houh ¶¶122-126.)~~

126. In the combination of Fontana and Lau, information about the advertisements selected through Lau’s method are then incorporated into Fontana’s “advertisement data” to facilitate display to the viewer:

~~[T]he~~ FIG. 7M illustrates example advertisement data 716 that can be used in association with multimedia content, to link one or more advertisements with multimedia content during playback. In the embodiment shown, the advertisement data 716 can include an advertiser identifier, a definition of an advertisement, and associated topics, keywords, or content that can be linked to the advertisement. In certain embodiments, the advertisement data 716 is used to link the content to advertisements during playback; in alternative embodiments, the advertisement data 716 is managed to track advertisements appearing with content[.]”

(Fontana ¶0117.)

127. A POSA would have appreciated that Fontana’s “advertisement data” in combination with Lau is another form of **video text** (~~e.g.~~, “a definition of

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

an advertisement, and associated topics, keywords, or content”) that is **based on the cross-referencing** (because derived from Lau’s method of matching advertisements to video content) and that **describes content of the video** (such as data on “associated topics, keywords, or content” and other data “used to link the content to advertisements during playback”). ~~(Houh ¶¶126-127.)~~

(j) **“generating a text, image, or animation based on the video text; and” (Claim 1[j])**

128. The combination of Fontana with Lau ~~renders obvious~~described above—using the topical metadata generated by Fontana to implement Lau’s method of matching advertisements to video content—“generat[es] a text, image, or animation based on the video text.” The combination of Fontana and Lau is described for claim 1[h] above, including a motivation to combine and reasonable expectation of success. As relates to claim 1[j], Lau teaches that the “ad units” are selected based on context (which is **video text**, ~~see as explained for~~ claim 1[i]), and can then be combined together to generate an advertisement. Once the ad units are selected, ~~see~~through the process described for claim 1[i], “[t]he ad units are then combined into an advertisement that is correlated to the subject matter.” (Lau ¶0038.) This advertisement may be a **text** (such as “an advertiser-specified message”, Lau ¶0027), **image** (such as the BMW logo, Lau ¶0038), **or animation** (such as a “video that may serve as pre/mid/post-roll”, Lau ¶0027). The advertisement additionally is

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

based on the video text as it is the result of combining the ad units, which are selected by matching to context (**video text**), together into an advertisement. (*See* Lau ¶0038; Houh ¶128.)

(k) “placing the text, image, or animation in the video.”
(Claim 1[k])

129. The combination of Fontana and Lau plac[es] the text, image, or animation in the video. ~~Lau teaches placing~~ The combination of Fontana and Lau is described for claim 1[h] above, including a motivation to combine and reasonable expectation of success. As relates to claim 1[k], Lau discloses that the advertisement matched to the video content (the “text, image, or animation”) is placed in the video by turning

~~ad units~~ 130. Lau discloses that the advertisement is placed in the video as pre-roll, mid-roll, post-roll, (that is, video content that plays before, during, or after another video) or even “injected” into the video itself. For example, Lau teaches that the ad units may be turned into “video that may serve as pre/mid/post-roll.” (Lau ¶0027.) Lau further explains that “the advertisement may be displayed in serial, parallel, or be injected into the rich media content.” (Lau ¶0034; *see also* Lau ¶0084 (“rendering formatter 204 can determine that an advertisement should be rendered serially relative to the portion of rich media content, in parallel to the portion of rich media content, or injected into the rich media content”); Lau claim 5 (“the

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

advertisement is injected into or laid on top of portion of rich media content”).) A POSA would have appreciated that injecting the advertisement into the rich media content (i.e., video), **plac[es] the text, image, or animation in the video.**—(Houh ¶¶129-130.)

~~The combination of Fontana with Lau therefore renders obvious claim 1.~~

2. **Claim 2: “The method according to claim 1, further comprising: generating a content-rich video based on the video, the text, and the video data.”**

~~The additional limitations of~~131. A POSA would have found claim 2
are1 obvious over ~~Fontana in view of Lau.~~ The the combination of Fontana with Lau,
as I discuss above. Additionally, the combination of Fontana with Lau “generat[es]
a content-rich video based on the video, the text, and the video data.”

132. Specifically, a POSA would have understood that the end result of the combination of Fontana with Lau—a video indexed to searchable metadata and into which an advertisement has been added—is a **content-rich video.** ~~This~~As described for claim 1, above, this content-rich video is generated based on **the video** (the initial multimedia content, see claim 1[pre]), **the text** (converted from the audio files, see claim 1[d]), **and the video data** (the object metadata, see claim 1[e]).—(Houh ¶¶131-132.)

3. **Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”**

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

~~The additional limitations of~~ 133. A POSA would have found claim ~~3~~
~~are~~ 1 obvious over ~~Fontana in view of Lau. (Houh ¶¶133–136.)~~ the combination of
Fontana with Lau, as I discuss above. Additionally, Fontana renders obvious
“applying natural language processing to the text to determine context
associated with the video.”

134. As discussed for claim 1[d], Fontana teaches converting the audio files associated with the video to text ~~as~~ in the form of a transcript. Fontana then further teaches analyzing this transcript using **natural language processing**. ~~(Fontana ¶0184~~ (“The search performed within the content can in certain embodiments, be performed based on natural language processing of an existing transcript”) (closed captioning or subtitles file provided by the content provider) or from a new transcript created using speech to text technology and edited by the content provider.” ~~(Fontana ¶0184; see also Fontana ¶0100 (“In alternative embodiments, additional search arrangements can be included as well, such as a natural language search[.]”)).~~

135. Fontana then determines the topic of the multimedia content—the **context associated with the video**—based on using natural language processing. ~~Specifically,~~ Fontana additionally teaches that search queries, including those leveraging the above natural language processing of the text transcript, can be used **to determine context associated with the video**, specifically in the form of keywords or relevant portions of the multimedia content. Fontana explains, for

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

example: “A content request operation 814 receives a request related to multimedia content~~...~~. The specific type of request received in the content request operation 814 can take a number of forms, such as a search query related to keywords appearing in one or more fields of metadata associated with the content (e.g., titles, authors, producers, genre, etc.) or in the transcript or other text associated with one or more pieces of content.” (Fontana ¶0126.) Additionally, “[~~thea~~] provide metadata operation 816 provides metadata (and optionally the multimedia content) in response to the request. The provide metadata operation 816 selects at least a portion of the metadata associated with the content (e.g., including ~~...~~ definitions of objects of interest, events, transcript information, ~~...~~ position information, etc.) for inclusion with the content during playback.” (Fontana ¶0127.) A POSA would have appreciated that identifying metadata, portions of the transcript, and portions of the multimedia data related to keywords or other queries is **determin[ing] context associated with the video.** ~~(Houh ¶¶134–135.)~~

~~Lau also~~ 136. The combination of Fontana with Lau further renders claim 3 obvious, as Lau too discloses using **natural language processing** to determine context associated with a video for purposes of matching it to an advertisement. As discussed previously, Lau teaches that ~~determining ad units based on keywords can use~~ “based on one or more keywords, ad units from the ad matrix are determined.” (Lau ¶0038.) Lau additionally teaches that this process can

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

build off of natural language processing: “Recognition engine 212 receives rich media content that may be accessed by a user and uses correlation recognition detection techniques to recognize the content....In another embodiment, it could be [a] natural language processing engine.” (Lau, ¶10041.) A POSA would have appreciated that using a recognition detection techniques, such as natural language processing ~~engine,~~ to recognize the content is **applying natural language processing to the text to determine context associated with the video.** ~~(Houh ¶136.)~~

4. **Claim 4: “The method according to claim 2, further comprising: applying natural language processing to the text to extract the topical meta-data.”**

137. A POSA would have found claim 2 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, for the reasons I explain for claim 3, the combination of Fontana and Lau renders obvious “**applying natural language processing to the text to extract the topical meta-data.**” As discussed for claim 1[f], the **topical meta-data** in the combination of Fontana and Lau includes the text metadata and object metadata generated according to Fontana’s method.

~~The additional limitations of claim 4 are obvious over Fontana in view of Lau. (Houh ¶¶137-138.)~~ Fontana’s “138. Specifically, a POSA would have appreciated that Fontana’s disclosures of “a search query related to keywords appearing in one or more fields of metadata associated with the content ...(e.g., titles, authors,

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

producers, genre, etc.) or in the transcript or other text associated with one or more pieces of content” ~~are to be~~ a search within **the topical meta-data**. (Fontana ¶0126.) Fontana teaches that searching can **apply natural language processing**, which was also obvious in view of Lau. (See claim 3; Fontana ¶¶0100, 0184.) Fontana additionally discloses “[a] provide metadata operation” that “provides metadata (and optionally the multimedia content) in response to the request.” (Fontana ¶0127.) ~~It was obvious that~~ A POSA therefore would have appreciated that, in response to a keyword search query, the method could **apply natural language processing to the text** (transcript, which is part of the text metadata) **to extract the topical meta-data** (i.e., to provide metadata in response to the search request).

5. **Claim 5: “The method according to claim 1, further comprising: processing the image files to extract additional text.”**

~~The additional limitations of~~ 139. A POSA would have found claim ~~5~~ are 1 obvious over ~~Fontana in view of Lau~~. (~~Houh ¶¶139-141~~ the combination of Fontana with Lau, as I discuss above. Additionally, the combination of Fontana with Lau “process[es] the image files to extract additional text.”

140.) For example, Lau discloses “extract[ing] additional text” from images, such as by performing optical character recognition. Lau explains ~~that~~ “[i]mage, “Image recognition can be used on visual portions of the rich media content. For example, optical character recognition (OCR), facial recognition, object

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

matching, etc.” (Lau ¶0040.) ~~OCR was a well-known image detection algorithm that A POSA would have appreciated that optical character recognition identifies text appearing in images and extracts that text.~~ ~~(Houh ¶¶140-141.)~~

141. It ~~was~~would have been obvious to ~~use Lau’s OCR on Fontana’s~~a POSA that the optical character recognition of Lau could be performed on the image files ~~(of Fontana—namely, the thumbnail images) to improve generating video metadata and matching ads.~~ ~~(Houh ¶141.)~~ Fontana within which Fontana performs its “objects of interest module” and “candidate object generation operation.” In addition to the motivations to combine explained for claim 1, a POSA would have been motivated to implement OCR, as taught by Lau, within the system of Fontana in order to improve the ability to generate metadata describing the video content and match ads to video content. A POSA implementing Fontana’s method would have understood that various algorithms could be used to process the image files, as Fontana expressly teaches that “[o]nce a user has selected one or more objects of interest, a number of optional detection algorithms can be applied to further define those or other objects of interest.” (Fontana ¶0144.) ~~Implementing A POSA would have been motivated to implement~~ OCR as ~~an~~one of those “optional detection ~~algorithm[.]” would~~algorithms” in order to capitalize on Lau’s teaching that OCR can improve the matching of advertisements to the video content. Lau specifically teaches that OCR can be performed as one of the “[c]orrelation

recognition detection techniques” that “may be used to determine that the advertisement is correlated to the portion of rich media content.” (Lau ¶¶0040; Houh ¶¶141-.) A POSA would have had a reasonable expectation of success because Fontana expressly discloses the use of additional detection algorithms. Additionally, OCR was well-known to a POSA in 2013 as one of many standard detection algorithms for image files.

6. **Claim 6: “The method according to claim 5, wherein the additional text is generated by segmenting the image files before processing the image files in parallel.”**

~~The additional limitations of claim 6 are disclosed by Fontana. (Houh ¶¶142-145.)~~

142. A POSA would have found claim 1 obvious over the combination of Fontana with Lau, as I discuss above. Furthermore, Fontana discloses “segmenting the image files before processing the image files in parallel.” Thus, in the combination of Fontana and Lau, a POSA would have understood that the image files (i.e., the thumbnail images) could be segmented before image processing, including the OCR processing described for claim 5 above.

143. Fontana teaches segmenting the image files in several ways. As a general matter, as described for claim 1[b], Fontana teaches a distributed computing system ~~that~~ “for video, audio, and image processing that segments all tasks into discrete portions that can be run in parallel: “As such, the various distributed

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

computing systems described in FIGS. 1-5, above, allow[s] for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing and resulting usability for content providers.” (Fontana ¶0135.) As described for claim 1[b], Fontana’s workflow server “received inbound data processing requests, for example from a content provider (as further discussed below) and distributes one or more portions of jobs associated with each data processing request to the integration framework 204 and the storage network 206.” (Fontana ¶0050.) The result “allow[s] resulting system “allows creation of pipelined data processing systems within a distributed computing environment, allowing computationally intensive jobs (e.g., video and audio content processing) to be distributed across a number of computing systems.” (Id.) ~~This~~ A POSA would have appreciated that this use of a distributed computing system segments tasks including image processing, prior to processing the images in parallel. ~~(Houh ¶143.)~~

144. Furthermore, Fontana also teaches segmenting image files into “even prior to distributing the image files to the workflow server for processing. Fontana discloses “generat[ing] a series of thumbnails representing scenes throughout the multimedia content” as part of “the thumbnail extraction module,” which is run prior to performing the object recognition. (Fontana ¶0095.) ~~In~~ This “series of thumbnails representing scenes” segments the image files into representative

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

thumbnail images for each scene. Similarly, Fontana explains that “[i]n certain embodiments, the candidate object generation operation 906 splits the multimedia content into a plurality of sections, and generates a thumbnail image associated with each of those sections for preview by the content provider.” (Fontana ¶0139.) This embodiment likewise segments the images by splitting the multimedia content into sections and generating a thumbnail image for each section.

145. Fontana also discloses ~~using~~ “that its image processing can use image processing algorithms to segment the images. For example, “[t]he candidate object generation module can be performed by any of a number of object recognition programs, including computer vision programs” such as “OpenCV,” which includes “segmentation” tools, as well as using “. Example computer vision tools include OpenCV, which is a library of motion tracking, facial recognition, gesture recognition, object identification, segmentation, and calibration tools.” (Fontana ¶0139.) Likewise, Fontana explains that “[t]hese operations associated with each content provider can be, for example, instructions provided to a video or other multimedia-editing web service” ~~to~~ “, for example to define specific elements of multimedia content, such as objects of interest appearing in the content, or to segment, edit, and reprocess the content.” (Fontana ¶0139; id. ¶0058.) Each Because Fontana implements image processing in a distributed computing system (see claim 1[b]), each of these ~~segments~~ algorithms results in the segmented image

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

files ~~before being processed in parallel processing.~~ (Houh ¶¶144-145.)

7. **Claim 7: “The method according to claim 1, further comprising: determining a motion associated with the one or more objects.”**

~~The additional limitations of claim 7 are disclosed or rendered obvious by Fontana. (Houh ¶¶146-148.) Fontana teaches using “OpenCV, which is a library of motion tracking, ... gesture recognition.” (Fontana ¶0139; see also Fontana ¶0144.) Motion tracking and gesture recognition determine a motion associated with the one or more objects. (Houh ¶¶147-148.)~~

146. A POSA would have found claim 1 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, Fontana discloses or renders obvious “**determining a motion associated with the one or more objects.**”

147. Specifically, Fontana teaches that “[t]he candidate object generation module can be performed by any of a number of object recognition programs, including computer vision programs. Example computer vision tools include OpenCV, which is a library of motion tracking, facial recognition, gesture recognition, object identification, segmentation, and calibration tools.” (Fontana ¶0139.) Both “motion tracking” and “gesture recognition” involve **determining a motion associated with the one or more objects.**

148. To the extent Fontana does not expressly teach the limitation of claim 7, it would have been obvious to implement the motion tracking and gesture recognition algorithms of OpenCV—as disclosed in Fontana—as among the “optional detection algorithms” that “can be applied to further define those or other

objects of interest.” (Fontana ¶0144.)

- 8. Claim 8: “The method according to claim 1, further comprising segmenting the audio files before processing the audio files in parallel.”**

149. A POSA would have found claim 1 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, a POSA would have found it obvious in light of Fontana’s parallel processing disclosures to “segment[] the audio files before processing the audio files in parallel” in order to take advantage of Fontana’s teaching that multiple servers can be used to efficiently carry out tasks like audio processing.

150. Specifically, Fontana teaches that multiple servers can be configured to perform audio processing, and that processing jobs can be separately performed across multiple of these servers in parallel:

~~The additional limitations of claim 8 are obvious over Fontana. (Houh ¶¶149-152.) Fontana teaches multiple servers with “In certain embodiments, the servers 302a-c are specifically designed according to the application the network 300 is intended to support; for example in the case where multimedia data is to be processed using the computing capabilities within network 300, one or more of the servers 302a-c can include specific graphical processing units for processing lower level video, image or audio algorithms” that“.~~ Other specific capabilities can be included into the servers 302a-c as well. The servers 302a-c are configured to share processing jobs, such that tasks can be performed

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

by one or more of the computing systems, or separated and performed
across multiple computing systems in parallel.”

(Fontana ¶0053.) ~~As a result,~~ “The “servers 302a-c can correspond to any of the
various computing systems 208 of FIG. 2, in that any of those computing
systems can perform all or a portion of a processing job ...as defined by a
scheduling algorithm, allowing multimedia content to be processed efficiently when
necessary.” (Fontana ¶0055.)

151. Fontana additionally teaches that segmentation may occur in order to
facilitate parallel processing. For instance, Fontana explains that “operations
associated with each content provider can be, for example, instructions provided to
a video or other multimedia-editing web service, for example ... to segment, edit,
and reprocess the content.” (Fontana ¶0058.) Similarly, Fontana teaches that ~~its~~
“the various distributed computing systems ...described in FIGS. 1-5, above,
allow for segmenting the processing into discrete portions (e.g., audio, video
processing separately, etc.) and parallel, pipelined processing of the data to ensure
fast content processing and resulting usability for content providers.” (Fontana
¶0135; ~~see also~~.)

152. Fontana ~~¶0058 (“instructions provided ... to segment, edit, and~~
~~reprocess the content”).) ~~Based on the foregoing, a~~ thus discloses parallel processing
of audio files across multiple servers, and segmenting data to facilitate parallel~~

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

processing. A POSA would thus have found it obvious that to implement audio processing in parallel, ~~Fontana would~~ the audio files would first be segmented. In other words, in order for the servers to “separate[]” an audio processing task to “perform[] across multiple computing systems in parallel,” the audio files could first be segmented in order to separate them across the servers. (See Fontana ¶0053.) It thus would have been obvious to a POSA based on Fontana to segment[] the audio files before processing the audio files in parallel. ~~(Houh ¶¶150-152.)~~

9. **Claim 9: “The method according to claim 7, wherein the audio files and the image files are segmented at spectrum thresholds.”**

153. A POSA would have found claim 7 obvious over the combination of Fontana with Lau, as I discuss above. Likewise, for the reasons discussed for claims 6 and 8, a POSA would have found it obvious over Fontana that “the audio files and the image files are segmented.”

~~The additional limitations of claim 9 are~~ 154. A POSA would further have found it obvious over Fontana. ~~(Houh ¶¶153-156.) Performing to perform~~ the segmentation of audio and image files **at spectrum thresholds** ~~could~~ in order to, for example, segment the audio and image files based on scene or content changes. Fontana teaches ~~using scene or content divisions for thumbnails,~~ “that “[t]he thumbnail extraction module 618 is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest (for example

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

...a first frame, a last frame, and immediately following major scene or sound changes in the content). In some embodiments, the thumbnail extraction module 618 generates a series of thumbnails representing scenes throughout the multimedia content.” (Fontana ¶0095.)

155. A POSA would have appreciated both that this division of the multimedia content by “major scene or sound changes in the content” corresponds to a segmentation of the multimedia content, and that identifying “major scene or sound changes in the content” could be performed by identifying **spectrum thresholds**. ~~(Houh ¶¶154-155.)~~ A POSA would have understood that a spectrum threshold is just a quantifiable valuation of some aspect of the media content at each moment in time – for example, the overall volume at a given time or frame of video, or its overall color saturation value. ~~(Houh ¶155.)~~ The ’972 patent, for example, gives ~~as an~~the example ~~of segmentation at spectrum thresholds~~ that “the audio data can be processed and converted into a spectrum. Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low audio activities in the audio data.” (’972, 5:16-20.) ~~One~~A POSA would have known that one common method ~~for~~of detecting major scene or sound changes wasis to look for changes above a **threshold**; ~~such approaches were known to a POSA before 2013.~~ ~~(Houh ¶155; see also EX1011, 11:44-54; EX1012~~given threshold in these quantifiable valuations of the multimedia content.

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

Such approaches were common in the literature and in practice well before 2013 and would have been well-known to a POSA. U.S. Patent No. 9,407,942 to Andrew Brenneman (“Brenneman”), which claims priority to October 3, 2008, for example, described the use of a “discontinuity profile” that “indicates the degree of discontinuity of each frame of the digital video content” and that “may be indicative of a change of scene ... or another point of segmentation of the digital video content.” (Brenneman, 11:44-54.) Likewise, U.S. Patent Publication No. 2004/0125877 A1 to Shin-Fu Chang et al. (“Chang”), which was filed on April 9, 2001, teaches “decomposition of video sequences into short shots by detecting a discontinuity in visual and/or audio features,” where scene changes may be detected for example “if the frame-to-frame color difference ratio is larger than a given threshold[.]” (Chang ¶¶0008, 0093.) More generally, a POSA would have known that performing a speech-to-text analysis, such as that taught by Fontana, commonly involved performing a Fourier analysis of the audio and segmenting the audio based on frequency information in order to divide the speech into individual phonemes. (Houh ¶155.) As This process of dividing speech by phonemes is described in Fontana describes: “Phonetic-based applications separate conversations into phonemes, the smallest components of spoken language; they then find segments within the long file of phonemes that match a phonetic index file representation of target words, phrases and concepts[.]” (Fontana ¶0169; see also Fontana ¶0171 (listing example

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

“[p]honetic-based applications useable as one or more of the speech to text conversion programs”).) ~~These~~ A POSA would have appreciated that these phonetic-based algorithms described in Fontana **segment audio files at spectrum thresholds.** (~~Houh ¶155.~~)

156. Additionally, a POSA would have known that periods of silence with low energy are commonly identified as gaps between words. (~~Houh ¶156.~~) ~~For~~ A POSA would have known that identifying the gaps between words was used, for example, in the “large vocabulary continuous speech recognition (LVCSR) engines” disclosed by Fontana ~~identify gaps between words.~~ (See Fontana ¶0169 (~~describing LVCSR searching text for~~ “LVCSR engines depends on a language model that includes a vocabulary/dictionary for speech-to-text conversion of audio files. The text file is then searched for target words, phrases and concepts.”); Fontana ¶0170 (listing ~~example LVCSR~~ “[e]xample sources of speech to text conversion programs performing LVCSR-based conversions”); ~~Houh ¶156.~~) This common understanding of the use of silence to segment audio aligns with the ’972 patent’s disclosure:

In some embodiments, the segmentation can be performed by a fixed period of time. In another example, quiet periods in the audio can be detected, and the segmentation can be defined by the quiet periods. For example the audio data can be processed and converted into a spectrum. Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

audio activities in the audio data. The quiet periods in the audio data can be ignored, and the processing requirements thereof can be reduced.

(’972, 5:13-22.) Therefore, a POSA would have found it obvious to use a spectrum threshold for segmenting audio and image files based on scene and/or sound changes, as taught by Fontana.

10. Claim 10: “The method according to claim 1, further comprising: generating an advertisement based on the text and the video data.”

~~The additional limitations of 157.~~ A POSA would have found claim 10
are obvious over ~~Fontana in view of Lau. (Houh ¶¶157-158.) Lau generates the~~
combination of Fontana with Lau, as I discuss above. Additionally, as explained for
claim 1, the combination of Fontana and Lau generat[es] an advertisement based
on the text and the video data. See claim 1[j].

158. Specifically, the advertisement generated by Lau is created by compiling ad units, each of which is selected by matching the ad units to the video based on the text (i.e., the audio transcript) and the video data (i.e., the metadata generated by Fontana). The details of this matching and advertisement generation are explained above for claim 1.

11. Claim 11: “The method according to claim 10, further comprising: placing the advertisement in the video at a preferred time.”

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

~~The additional limitations of claim 11 are obvious over Fontana in view of Lau. (Houh ¶¶159-162)159.~~ A POSA would have found claim 10 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, the combination of Fontana with Lau **plac[es] the advertisement in the video at a preferred time.**

~~160.) Lau’s~~ Specifically, Lau teaches that its “advertisements are time aligned to correlate to the subject matter” and turned into “video that may serve as pre/mid/post-roll.” (Lau ¶¶0027, 0045; see also to maximize the revenue.” (Lau ¶0045.) A POSA would have appreciated that time-aligned advertisements must be **placed in the video at a preferred time.** This understanding is consistent with Lau’s teachings on how to place advertisements in the video.

~~Lau ¶0034 (“161. As discussed for claim 1[k], Lau discloses that “the advertisement may be displayed in serial, parallel, or be injected into the rich media content.”); Lau ¶0084 (serial, parallel (Lau ¶0034; see also Lau ¶0084 (“rendering formatter 204 can determine that an advertisement should be rendered serially relative to the portion of rich media content, in parallel to the portion of rich media content, or injected rendering into the rich media content”); Lau claim 5.)~~ These time-aligned advertisements are **placed** (“the advertisement is injected into or laid on top of portion of rich media content”).) A POSA would have appreciated that injecting the advertisement into the video **places the advertisement in the video**

at a preferred time.

162. Lau likewise discloses that the ad units may be turned into “video that may serve as pre/mid/post-roll.” (Lau ¶0027.) Playing an advertisement as pre-roll, mid-roll, or post-roll each likewise place the advertisement in the video at a preferred time.

12. Claim 12: “The method according to claim 6 wherein the additional text includes information regarding context associated with the video.”

~~The additional limitations of claim 12 are obvious over Fontana in view of Lau. (Houh ¶¶163-164.) It would have been obvious to a POSA that text generated from the OCR of image files (the additional text of claim 6) is “information regarding context associated with the video.” Lau explains that “correlation-recognition detection techniques” including “optical character recognition (OCR)” “may be used to determine that the advertisement is correlated to the portion of rich media content.” (Lau ¶0040.) A POSA would understand that this corresponds to a “context associated with the video.” (Houh ¶164.)~~

163. A POSA would have found claim 6 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, it would have been obvious that in the combination of Fontana with Lau, the additional text—in this combination, the result of the OCR process on the image files—includes information regarding context associated with the video. It would have been obvious to a POSA that text generated from the OCR of image files is “information regarding context associated with the video.”

164. Lau confirms this understanding, as it explains that its OCR process is performed in order to determine the content of the rich media for purposes of

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

matching the content to an advertisement. A POSA would understand that this corresponds to a “context associated with the video.” As Lau teaches:

Correlation recognition detection techniques may be used to determine that the advertisement is correlated to the portion of rich media content. For example, keywords may be detected in the rich media using audio recognition. Audio recognition may include speech recognition, music detection on music portions, sound effect detection on sound effects, etc. Other techniques for keyword detection can include using preset word tags or indicators in the rich media content. Image recognition can be used on visual portions of the rich media content. For example, optical character recognition (OCR), facial recognition, object matching, etc. Other recognition techniques can be employed. For example, any suitable way of determining the content of rich media can be used to correlate a portion of the rich media content to an advertisement.

(Lau ¶10040.)

13. Claim 13: “The method according to claim 6, wherein the additional text relates to a symbol appearing in the video.”

~~The additional limitations of~~ 165. A POSA would have found claim ~~13~~ are 6 obvious over the combination of Fontana ~~in view of~~ with Lau ~~for~~, as I discuss above. For the same reasons as discussed for claims 5 and 6. ~~(Houh ¶165.), a POSA would have found claim 13 obvious over the combination of Fontana with Lau. The “additional text” of claims 5, 6, and 13 in the combination of Fontana with Lau is~~

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

the result of performing OCR on the images, as taught by Lau. A POSA would understand that OCR includes symbols such as letters, numbers, punctuation, and other typographic symbols (e.g., “+”). ~~(Id.)~~ Thus it would have been obvious to a POSA that the additional text relates to a symbol appearing in the video.

14. **Claim 14: “The method according to claim 13, wherein the symbol is a brand logo, and wherein the additional text includes information regarding placement and time of appearance of the brand logo.”**

~~The additional limitations of claim 14 are obvious over Fontana in view of Lau. (Houh ¶¶ 166-170.)~~

166. A POSA would have found claim 13 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, a POSA would have found it obvious that the symbol could be a brand logo, and that the additional text includes information regarding placement and time of appearance of the brand logo.

167. A POSA would have understood that the object recognition and OCR processing of the combination of Fontana with Lau could encompass recognizing brands and generating metadata and/or keywords corresponding to those brands. The ’972 patent explains that techniques such as object recognition and OCR can be used to identify brands:

For example, the normalized image frame files can be analyzed for text identification and/or by optical character recognition. ...The data can be improved through a dictionary verification step. Various maps can be created based on edge detection and/or image segmentation

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

techniques. Such techniques can be improved by focusing on regions of interest, for example based on brands, logos, objects, and/or features of interest.

(’972, 6:62-7:2.)

168. A POSA would have found ~~this~~such a use of object recognition and/or OCR obvious in view of Fontana and Lau. ~~(Houh ¶¶167-168.)~~In fact, Fontana expressly discloses that “[a]dditional objects of interest can be identified by a user,” thus permitting a user to specify, for example, a brand logo for the method to recognize. (Fontana ¶0089.) Similarly, Fontana explains that “additional objects and individuals appearing in the content” can be included in to recognize can be specified with a script provided with the multimedia content. ~~;~~

In some embodiments, a content provider can provide a script alongside the multimedia content 602 to the system 600. In such embodiments, the script can contain a number of descriptions of the content, such as dialog occurring in the content, objects and individuals appearing in the content, as well as mood, scene, and other information that can be used at least in part to assist in generating metadata describing the content for use in connection with the systems and methods of the present disclosure.

(Fontana ¶0088.) A POSA would have found it obvious that a brand logo could be one such object specified by a script as appearing within a video. ~~(Houh ¶168.)~~

169. Furthermore, Lau expressly teaches that brands may be a particularly

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

useful type of content to search for within the video in order to match ad units to specific locations within the video, ~~by giving an.~~ Lau gives the example of identifying ad units corresponding to the BMW model discussed in the video: “

Correlation engine 202, when determining the advertisement, may determine one or more ad units that correlate to the subject matter. For example, based on one or more keywords, ad units from the ad matrix are determined. The ad units are then combined into an advertisement that is correlated to the subject matter. One example of this is BMW may provide a general ad unit for their logo and have a different ad unit for different models, such as the 330 model, 530 model, etc. . . . The logo unit and each of the model units can be combined at runtime based on the context of the content. If the content talks about the 330 model then the logo and the 330 ad units may be combined and presented to the user.”

(Lau ¶0038.) A POSA combining Fontana with Lau therefore would have found it obvious to use brand logos as objects of interest for identification within the image files. A POSA would have been motivated to do so in order to achieve the benefits described by Lau of improved matching of ad units to the video content, and would have had a reasonable expectation of success in light of Fontana’s disclosures discussed above that objects of interest may be specified by the user and/or an accompanying script. ~~(Houh ¶169.)~~

170. As explained for claims 1[e], 1[f], and 5, the recognized objects

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
~~Paper 2 – PETITION FOR INTER PARTES REVIEW~~

(including the results of OCR and any brand logos) are converted into searchable metadata, i.e., **additional text**. And as I explained for claim 1[h], Fontana teaches that this metadata is linked to where in the video content it occurs, such that **the additional text includes information regarding placement and time of appearance of the brand logo.**

15. **Claim 15: “The method according to claim 1, wherein the one or more objects are letters appearing in the video.”**

~~The additional limitations of claim 15 are obvious over Fontana in view of Lau for the same reasons as discussed for claim 5. (Houh ¶171.) A POSA would have known that the **one or more objects** identified by OCR would include **letters appearing in the video.** (Id.)~~

171. A POSA would have found claim 1 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, for the same reason as discussed above for claim 5, it would have been obvious to a POSA to implement OCR as taught by Lau as part of the combination of Fontana with Lau. A POSA would have known that the **one or more objects** identified by OCR would include **letters appearing in the video.**

16. **Claim 16: “The method according to claim 6, wherein the additional text relates to faces appearing in the video.”**

~~The additional limitations of~~172. A POSA would have found claim ~~16~~are6 obvious over ~~Fontana in view of Lau as both references teach~~the combination of Fontana with Lau, as I discuss above. Additionally, a POSA would have found it obvious that in the combination of Fontana with Lau, the **additional text** could

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

relate to faces appearing in the video as both Fontana and Lau teach the use of facial recognition as part of the image processing. ~~(Houh ¶¶172-174~~

173.) Fontana teaches using OpenCV, whose tools include “explains that “[t]he candidate object generation module can be performed by any of a number of object recognition programs, including computer vision programs. Example computer vision tools include OpenCV, which is a library of motion tracking, facial recognition,” as well as gesture recognition, object identification, segmentation, and calibration tools.” (Fontana ¶0139.) Fontana also teaches the use of “additional detection algorithms” that “can include facial recognition[.]” (Fontana ¶¶0139, 0144.)

174. Similarly, Lau teaches “facial recognition” as a form of that, “[i]mage recognition can be used on visual portions of the rich media content. For example, optical character recognition (OCR), facial recognition, object matching, etc.” (Lau ¶0040.) ~~The~~ These image recognition; algorithms, including facial recognition, ~~is then~~are used to generate “meta-data ~~about the visual content” (“additional text”): “In the video or visual recognition embodiment, meta-data about the visual content is generated or culled from the content itself.” (Lau ¶0042; see also Lau ¶0040 (“Other techniques for keyword detection can include using preset word tags or indicators in the rich media content. Image recognition can be used on visual portions of the rich media content. For example, ... facial~~

recognition[.]”).)

17. Independent Claim 17: “A system for extracting data from a video, comprising:” (Claim 17[pre])

175. Assuming the preamble provides a claim limitation, Fontana discloses it. For the same reasons ~~as~~explained for claim 1, and in particular limitations 1[pre] and 1[a], Fontana discloses a system for extracting data from a video. To the extent the preamble of claim 1 is addressed to a “method” and the preamble of claim 17 is addressed to a “system,” Fontana teaches both a method and a system. (See Fontana ¶0008 (“In a first aspect, a method for providing multimedia content is disclosed.”); Fontana ¶0009 (“In a second aspect, a system for providing multimedia content is disclosed.”); ~~Houh ¶175.~~)

(a) “a computer processor having a plurality of processors for parallel processing; and” (Claim 17[a])

~~For the reasons~~176. I explained above for claim 1[b]; that Fontana discloses a distributed computing system for parallel processing. Fontana further discloses ~~implementing its~~that this system is implemented using a **computer processor having a plurality of processors for parallel processing:**

In addition, electronic computing device 500 comprises a processing unit 504. ~~---~~As mentioned above, a processing unit is a set of one or more physical electronic integrated circuits that are capable of executing instructions. In a first example, processing unit 504 may execute software instructions that cause electronic computing device

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

500 to provide specific functionality. In this first example, processing unit 504 may be implemented as one or more processing cores and/or as one or more separate microprocessors.

(Fontana ¶0074; ~~Houh ¶176.~~)

(b) “a non-transitory computer readable medium containing instructions directing the system to execute the steps of:” (Claim 17[b])

177. Fontana teaches that its system is implemented on a **non-transitory computer readable medium containing instructions directing the system to execute** certain steps. Fontana explains:

FIG. 5 is a block diagram illustrating example physical components of an electronic computing device 500, which can be used to execute the various operations described above, ~~...~~and provides an illustration of further details regarding any of the computing systems described above in FIGS. 1-4. A computing device, such as electronic computing device 500, typically includes at least some form of computer-readable media. ~~...[C]omputer-readable~~Computer readable media can be any available media that can be accessed by the electronic computing device 500. By way of example, and not limitation, computer-readable media might comprise computer storage media and communication media.

(Fontana ¶0072.) Fontana further explains that this computer readable media can contain “computer readable **instructions,**”~~”~~ and can be implemented in forms of memory such as “RAM, ROM, EEPROM, and flash memory ~~or other memory~~”

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

~~technology,”~~ ~~tw~~hich~~that~~ a POSA would have recognized as **non-transitory computer readable media**.~~(;~~

In the context of the electronic computing device 500, computer storage media includes volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, various memory technologies listed above regarding memory unit 502, non-volatile storage device 510, or external storage device 516, as well as other RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium that can be used to store the desired information and that can be accessed by the electronic computing device 500.

(Fontana ¶0078;~~Houh ¶177.~~)

(c) **“converting audio associated with the video to text;”**
(Claim 17[c])

178. Fontana discloses **converting audio associated with the video to text** for the same reasons discussed above for claim 1[d].~~(Houh ¶178.)~~

(d) **“converting images associated with the video to video data;”** (Claim 17[d])

179. Fontana discloses **converting images associated with the video to video data** for the same reasons discussed above for claim 1[e].~~(Houh ¶179.)~~

- (e) “generating the video data by segmenting image files of the video before processing the image files in parallel;” (Claim 17[e])

180. Fontana discloses **generating video data** for the reasons discussed in claim 1[e], and **processing the image files in parallel**, for the reasons discussed for claim 1[b]. For the same reasons as discussed above for claim 6, Fontana further discloses **segmenting image files of the video before processing the image files in parallel.** ~~(Houh ¶180.)~~

- (f) “identifying one or more objects in the image files;” (Claim 17[f])

181. For the reasons discussed above for claim 1[b], Fontana discloses **identifying one or more objects in the image files.** ~~(Houh ¶181.)~~

- (g) “generating data topics, from the text and the video data, that describe content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio;” (Claim 17[g])

182. For the same reasons discussed above for claim 1[f], Fontana discloses ~~claim 17[g].~~ ~~As explained~~ generating data topics, from the text and the video data, that describe content of the video by deriving semantic information from the identification of the one or more objects and semantic information from the audio. My analysis for claim 1[f]; explains that a POSA would have appreciated that the various metadata generated by Fontana correspond to the “topical meta-data that describes content of the video” within claim 1 of the ’972 patent. A POSA would

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

~~also~~equally recognize that Fontana’s generated metadata also correspond to “**data topics ... that describe content of the video.**” For example, Fontana’s object metadata identifying objects are “data topics,” as are keywords or other terms derived from the text metadata. ~~(Houh ¶182.)~~

**(h) “adding the data topics to the video as meta-data; ”
(Claim 17[h])**

183. For the same reasons discussed above for claim 1[g], Fontana discloses **adding the data topics to the video as meta-data**, including by linking the data topics (i.e., Fontana’s text **metadata** and object **metadata**) to specific times in the video, synchronizing the metadata for simultaneous display to the user, and storing the video with its metadata. ~~(Houh ¶183)~~See claim 1[g].

(i) “cross-referencing the text, the video data, and the topics with the video based on the generated data topics; ” (Claim 17[i])

184. It would have been obvious to combine Fontana with Lau to **cross-referenc[e] the text, the video data, and the topics with the video based on the generated data topics** for substantially the same reasons as I explained above for claim 1[h]. The combination of Fontana and Lau ~~and motivation to combine are~~is the same as for claim 1. ~~(Houh ¶¶184-188.)~~

, a POSA would have been motivated to combine the references for the same reasons as discussed for claim 1, and a POSA would have had a reasonable

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

expectation of success for the same reasons discussed for claim 1. The ~~Minor~~ differences in wording between claim 1[h] and 17[i] do not materially alter the analysis. I address these below.

185. First, claim 1[h] requires cross-referencing “the text and the video data” whereas claim 17[i] requires cross-referencing “the text, the video data, and the topics with the video.” As I explained for claim 1[h]:

Lau discloses performing a search to determine advertisements whose content and/or topics match the content and/or topics for portions of video content. The search involves counting the number of keyword and/or concept matches for a search term near a particular time in the video content. A POSA would have found it obvious to combine Lau and Fontana by using the **topical meta-data** generated by Fontana (*see* claims 1[f] and 1[g]) as well as **the text and the video data** (*see* claims 1[d] and 1[e]) as keywords and concepts searched by Lau for purposes of identifying ads or ad units to match to the video based on context or subject matter. Each of the ads, ad units, and context or subject matter of the video is a “**topic.**” As discussed for claim 1[f], the **topical meta-data** (various contextual information and metadata) is derived from **the text and the video data**, such that searching the topical meta-data **cross-references the text and the video data**, and any such search is **based on the generated topical meta-data**. A POSA would have further found it obvious to use the text and the video data as part of Lau’s matching search, which further **cross-references the text and the video data**. As a result, this combination “**cross-referenc[es] the**

text and the video data based on the generated topical meta-data to determine topics.”

186. What claim 1[h] refers to as “topical meta-data” corresponds to the **topics** in claim 17[i], such that for the same reasons discussed for claim 1[h] above, the combination of Fontana with Lau **cross-referenc[e]s the text, the video data, and the topics.** ~~(Houh ¶¶185-186~~

187.) Furthermore, this occurs “**with the video.**” Specifically, as I explained for claim 1[h], Lau’s search to match ad units to the content identifies a set of candidate locations within the video, then scores each based on the strength of the match. ~~;~~

For each piece of candidate content associated with an ad, correlation engine 202 determines candidate times where the content may be relevant to the ad. Correlation engine 202 locates the times where the keywords and concepts match. For each candidate time, correlation engine 202 creates an “ad anchor” holding the score for the match. The score may be a linear combination of the following weights:

1. Probability of the keyword/concept match pulled from the recognition lattice.
2. Concentration of the match—the more keywords/concepts for the ad matches near the time, the higher the score. One embodiment of this score may be a count of the number of matches within a certain window of the current time....

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

(Lau ¶¶0057–0059.) Thus, the same combination described for claim 1[h] **cross-referenc[es] the text, the video data, and the topics with the video.**—(Houh ¶187

188.) Finally, claim 1[h]’s cross-referencing is “based on the generated topical meta-data to determine topics” whereas claim 17[i]’s cross-referencing is **“based on the generated data topics.”** Here, the topical meta-data for claim 1[h] corresponds to the data topics for claim 17[i], such that this element of claim 17[i] is obvious over Fontana with Lau for the same reasons as claim 1[h].—(Houh ¶188.)

(j) **“generating a text, image, or animation based on the data topics; and” (Claim 17[j])**

189. Claim 17[j] essentially combines the end of claim 1[h] (“to determine topics”), claim 1[i], and claim 1[j] into a single limitation – that the system **generat[e] a text, image, or animation based on the data topics.** For the same reasons discussed above for claims 1[h], 1[i], and 1[j], claim 17[j] would have been obvious over the combination of Fontana with Lau.—(Houh ¶189.)

(k) **“placing the text, image, or animation in the video.” (Claim 17[k])**

190. For the same reasons discussed above for claim 1[k], Fontana discloses **placing the text, image, or animation in the video.**—(Houh ¶190.)

~~The combination of Fontana and Lau therefore renders claim 17 obvious.~~

18. **Claim 18: “The system according to claim 17, wherein converting the audio comprises natural language processing.”**

~~The additional limitations of claim 18 are obvious over Fontana in view of~~

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
~~Paper 2 – PETITION FOR INTER PARTES REVIEW~~

~~Lau for the same reason as discussed above for claim 3. (Houh ¶191.)~~

191. A POSA would have found claim 17 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, for the same reason as discussed above for claim 3, the combination of Fontana with Lau renders obvious that converting the audio comprises natural language processing, as both Fontana and Lau disclose natural language processing as part of the audio processing and resulting search. (See claim 3, supra.)

19. **Claim 19: “The system according to claim 17, the computer directs the audio to be converted by at least one node of a cluster and the computer directs the images to be converted by at least one other node of the cluster in parallel.”**

~~The additional limitations of claim 19 are disclosed by Fontana. (Houh ¶¶192-195.)~~ For 192. A POSA would have found claim 17 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, for the same reason as discussed above for claim 1[b], Fontana discloses that **the computer directs the audio to be converted by at least one node of a cluster and the computer directs the images to be converted by at least one other node of the cluster in parallel.** (Houh ¶192.)

193. As also explained for claim 1[b], Fontana teaches a “distributed computing network” that performs computer system for parallel processing. (See, e.g.,

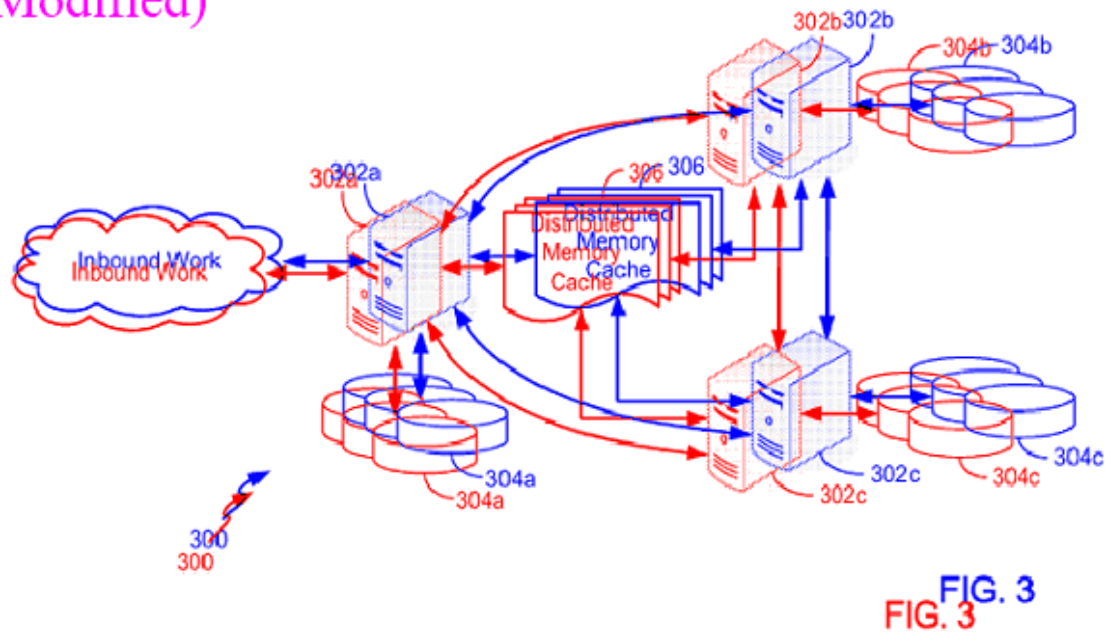
Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

In the embodiment shown, the distributed computing network 300 includes a plurality of computing systems, illustrated as servers 302a-c. The servers 302a-c are communicatively interconnected, and each includes a corresponding data storage system 304a-c. The servers 302a-c share a distributed memory cache 306, and are each capable of accessing a shared cache of memory that is not residing in any of data storage systems 304a-c. The servers 302a-c are interfaced to inbound work, such as from a scheduler system (as described in further detail in connection with FIG. 4, below) for coordination and communication of data for processing.

In certain embodiments, the servers 302a-c are specifically designed according to the application the network 300 is intended to support; for example in the case where multimedia data is to be processed using the computing capabilities within network 300, one or more of the servers 302a-c can include specific graphical processing units for processing lower level video, image or audio algorithms. Other specific capabilities can be included into the servers 302a-c as well. The servers 302a-c are configured to share processing jobs, such that tasks can be performed by one or more of the computing systems, or separated and performed across multiple computing systems in parallel.

(Fontana ¶¶0052, 0053; ~~Houh¶193~~.) Fontana illustrates this system in Figure 3:

(Modified)



194. A POSA would have appreciated that each of the servers, such as servers 302a-c in Figure 3, represents **at least one node of a cluster**, such that the distributed computing system of Fontana directs processing tasks to at least one node of a cluster. ~~(Houh ¶¶193-194.)~~

195. Fontana further explains that this distributed computing system and network of servers “allow for segmenting the processing into discrete portions (e.g., audio, video processing separately, etc.) and parallel, pipelined processing of the data to ensure fast content processing and resulting usability for content providers.” (Fontana ¶0135.) From this, a POSA would have appreciated that audio and images would be separated and processed in parallel, such that Fontana discloses that **the computer directs the audio to be converted by at least one node of a cluster and**

the computer directs the images to be converted by at least one other node of the cluster in parallel. ~~(Houh ¶195.)~~

20. Claim 20: “The server according to claim 17, wherein the audio and the images are segmented at spectrum thresholds.”

~~For the same reason as discussed above for claim 9, the additional limitations of claim 20 are obvious over Fontana. (Houh ¶196.)~~

196. A POSA would have found claim 17 obvious over the combination of Fontana with Lau, as I discuss above. Additionally, for the same reason as discussed above for claim 9, it would have been obvious to a POSA to segment the audio and the images at spectrum thresholds.

DC. Ground 2: Claims 8-9 and 20 Are Obvious Over Fontana in view of Lau and Arakawa

1. Claim 8: “The method according to claim 1, further comprising segmenting the audio files before processing the audio files in parallel.”

~~Claim 197. A POSA would have found claim 1 is~~ obvious over the combination of Fontana ~~in view of~~ with Lau ~~for the same reasons as in,~~ as I discuss above for Ground 1. ~~The additional limitations of claim 8 are obvious over Arakawa. Specifically~~ Additionally, a POSA would have found it obvious to combine Fontana with Arakawa to “segment[] the audio files before processing the audio files in parallel” in order to take advantage of Fontana’s teaching that multiple servers can be used to efficiently carry out tasks like audio processing while also increasing

efficiency by processing voice clips separately from background noise.—(Houh ¶¶197-205.)

198. Arakawa teaches **segmenting audio** by frame, and further segmenting audio files into voice sections (corresponding to speech) versus non-voice sections corresponding to other noise, prior to performing speech recognition on the voice sections. The result is “a voice recognition system capable of, while suppressing negative influences from sound not to be recognized, correctly estimating utterance sections that are to be recognized.” (Arakawa, Abstract.) Arakawa’s segmentation into voice versus non-voice sections is performed by comparing each frame to a threshold value, such as Arakawa explains:

The voice segmentation unit 103 calculates a voice segmentation feature value which indicates possibility of being voice for each frame input sound data. Then, the voice segmentation unit 103 classifies each frame into a voice frame or a non-voice frame by comparing a threshold value (hereinafter, it is referred to as threshold θ) and a voice segmentation feature value for each frame. If a calculated voice feature value for a frame is larger than the threshold value θ , the frame is classified as a voice frame. If a calculated voice feature value is less than the threshold value θ , the frame is classified as a non-voice frame. Then, the voice segmentation unit 103 merges connected voice frames classified above into a voice section (hereinafter, referred to as a first voice section). As a voice segmentation feature value, amplitude power for each frame can be used, for example. However, the voice

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

segmentation feature value which indicates possibility of being voice is not limited to amplitude power.

(Arakawa ¶0043.)

~~(Arakawa ¶0043.)~~ 199. Arakawa next teaches calculating “a feature value ~~used~~ for voice recognition” for each audio frame:

~~, such as, for example,~~ “The voice recognition feature value calculating unit 105 calculates a feature value used for voice recognition (hereinafter, it is described as a voice recognition feature value) for each frame input sound data. As a voice recognition feature value, cepstrum feature or its derivative feature.” ~~(Arakawa ¶0044.)~~ can be used, value for example.

(Arakawa ¶0044.)

200. Arakawa further teaches using the feature value for each frame to identify words and/or phonemes for speech recognition, as well as to update the threshold for distinguishing voice from non-voice:

The searching unit 108 calculates, based on the voice recognition feature value, a likelihood of voice and a likelihood of non-voice for each frame, and searches for a word sequence using these likelihoods and the above-mentioned models. ~~...~~ The searching unit 108 may search for a maximum-likelihood word sequence among calculated the likelihoods of voice, for example.

Also, the searching unit 108 segments a section to be the target of the

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

voice recognition (hereinafter, it is referred to as a second voice section) based on the likelihood of voice and the likelihood of non-voice that have been calculated. ~~...~~Specifically, the searching unit 108 segments a section during which the likelihood of voice that has been calculated based on the voice recognition feature value is higher than the likelihood of non-voice that has been calculated based on a voice recognition feature value as the second voice section.

Thus, the searching unit 108 obtains the word sequence corresponding to the input sound (a recognition result) using the feature value for each frame, the vocabulary/phoneme model and the non-voice model, and, in addition to that, obtains the second voice section....

According to a difference between length of the first voice section and length of the second voice section, the parameter updating unit 109 updates the threshold value θ . That is, the parameter updating unit 109 compares the first voice section and the second voice section, and updates the threshold value θ to be used by the voice segmentation unit 103.

(Arakawa ¶¶0047–0050.)

201. Arakawa thus teaches at least three segmentations of the audio files: the first segmentation of voice versus non-voice sections, the second segmentation of voice segments based on phoneme recognition, and the frame-by-frame segmentation for purposes of voice segmentation and voice recognition.—(Houh

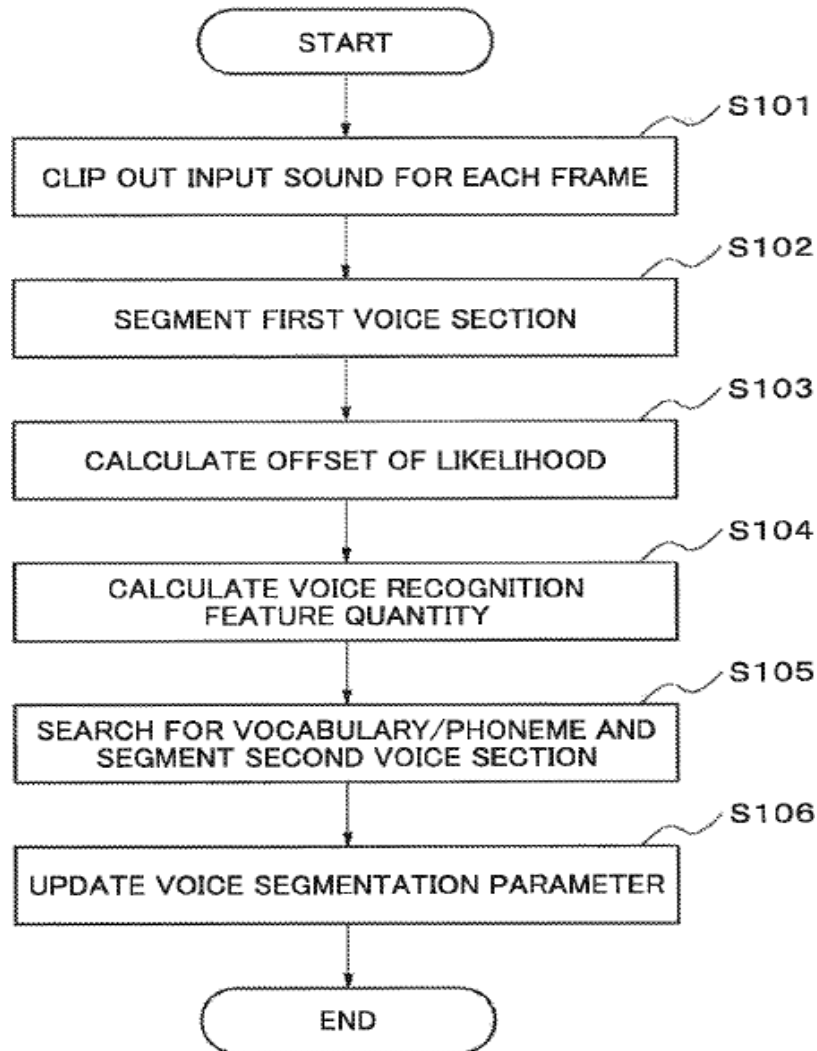
~~¶¶198-201.)~~

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

202. It would have been obvious to a POSA to combine Arakawa’s method of performing voice recognition using segmentation with Fontana in order to take advantage of Arakawa’s ~~improved~~benefits of improving speech recognition in the presence of background noise. (See, e.g., Arakawa Abstract, ¶0001 (“The present invention relates to a voice recognition system, a voice recognition method and a voice recognition program which recognize voices in an environment where background noise exists.”); ~~see also Arakawa ¶0012; Houh ¶202.~~), ¶0012 (“When voice recognition is performed, there is a case where the background noise, line noise and sudden noise such as a sound of hitting a microphone and the like exists. In such case, by using the voice recognition apparatus disclosed in patent document 1 and patent document 2, an error of voice recognition can be suppressed.”).)

203. Furthermore, a POSA combining Fontana with Lau would have found it obvious to segment the audio files by frame and into voice sections according to Arakawa’s “voice segmentation unit” **before processing the audio files in parallel.** (~~Houh ¶203.~~) Arakawa explains that dividing the audio into frames, and segmenting voice versus non-voice sections using the voice segmentation unit, each occur prior to performing the speech recognition functions described in Arakawa. This is illustrated in Figure 2, which shows that “clip out input sound for each frame” and “segment first voice section” each occur before the speech recognition functions (e.g., “search for vocabulary/phoneme”):

Fig.2



(See Arakawa Fig. 2.)

204. A POSA would have found it obvious that Fontana’s distributed computing system could separate the resulting audio segments across multiple servers to process in parallel for speech recognition—either by having each frame

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

processed in parallel by the “voice recognition feature value calculating unit” and “searching unit” of Arakawa, or by having each voice versus non-voice section of Arakawa processed in parallel. ~~(Houh ¶204.) As also discussed for Ground 1 Claim 8~~Indeed, Fontana expressly discloses that processes such as audio processing could be separated and performed in parallel in this fashion—~~its~~:

~~servers, which include~~ “In certain embodiments, the servers 302a-c are specifically designed according to the application the network 300 is intended to support; for example in the case where multimedia data is to be processed using the computing capabilities within network 300, one or more of the servers 302a-c can include specific graphical processing units for processing lower level video, image or audio algorithms,” “Other specific capabilities can be included into the servers 302a-c as well. The servers 302a-c are configured to share processing jobs, such that tasks can be performed by one or more of the computing systems, or separated and performed across multiple computing systems in parallel.” ~~(Fontana ¶0053.)~~

(Fontana ¶0053.) I also discuss this for Ground 1 Claim 8.

205. Rationale and motivation to combine (Fontana and Lau with Arakawa): A POSA would have been motivated to combine Fontana and Lau with Arakawa to take advantage of Arakawa’s benefits of improving speech recognition in the presence of background noise, as ~~discussed~~I discuss above. ~~(Houh ¶205.)~~ Additionally, Fontana explains that “a plurality of different speech to text algorithms

Ex. 1002 – DECLARATION OF DR. HENRY HOUH
Paper 2 – PETITION FOR INTER PARTES REVIEW

can be applied.” (Fontana ¶0093.) As such, a POSA would have been motivated to identify audio processing algorithms, such as that disclosed by Arakawa, that provided additional benefits. ~~(Houh ¶205.)~~ Arakawa is an analogous reference to the ’972 patent. Arakawa, like the ’972 patent, addresses audio processing using segmentation. (’972, 1:49-53 (“The text from the audio can be generated by first segmenting images of the audio, and then converting the segments of images to text in parallel. The audio can be segmented at spectrum thresholds.”); Arakawa, Abstract (“Provided is a voice recognition system capable of, while suppressing negative influences from sound not to be recognized, correctly estimating utterance sections that are to be recognized. A voice segmenting means calculates voice feature values, and segments voice sections or non-voice sections by comparing the voice feature values with a threshold value.”).) A POSA would have had a reasonable expectation of success in deploying Arakawa’s audio processing algorithms within the context of Fontana’s distributed computing for multimedia processing. ~~(Houh ¶205.)~~ Indeed, as noted, Fontana expressly teaches that “a plurality of different speech to text algorithms can be applied” (Fontana ¶0093) and that audio processing can be done in parallel using a plurality of servers (Fontana ¶0053).

2. **Claim 9: “The method according to claim 7, wherein the audio files and the image files are segmented at spectrum**

thresholds.”

206. A POSA would have found claim 7 obvious over the combination of Fontana with Lau, as ~~discussed~~discuss above for Ground 1. Additionally, as discussed above for Ground 2 Claim 8, Arakawa discloses segmenting audio files at **spectrum thresholds** to distinguish voice from non-voice sections of an audio file. It would have been obvious to combine Fontana with Arakawa to **segment[] the audio files and the image files at spectrum thresholds.** ~~(Houh ¶¶206-208.)~~

207. The general motivation to combine Fontana with Arakawa is discussed above for Ground 2 Claim 8. A POSA would additionally have been motivated to segment the image files in the same locations as the audio files based on Arakawa’s thresholds for the voice segmentation unit. ~~(Houh ¶¶207-208.)~~ Specifically, Fontana explains that it may be desirable to generate thumbnails (i.e., segment the image files) based on sound changes in the multimedia content: “The thumbnail extraction module 618 is arranged to generate thumbnails at possible locations the content provider would like to create an object of interest (for example a first frame, a last frame, and immediately following major scene or sound changes in the content). In some embodiments, the thumbnail extraction module 618 generates a series of thumbnails representing scenes throughout the multimedia content.” (Fontana ¶0095.)

208. A POSA would have found it obvious that Arakawa’s voice

segmentation unit identifies “sound changes” based on its use of a **spectrum threshold**. (See Arakawa ¶0043 (“If a calculated voice feature value for a frame is larger than the threshold value θ , the frame is classified as a voice frame. If a calculated voice feature value is less than the threshold value θ , the frame is classified as a non-voice frame. Then, the voice segmentation unit 103 merges connected voice frames classified above into a voice section (hereinafter, referred to as a first voice section).”).) Thus, a POSA would have appreciated that the same segmentation into voice sections versus non-voice sections performed for audio files in Arakawa could be used for generating the thumbnail images to segment the image files in Fontana.—(Houh ¶208.)

3. Claim 20: “The server according to claim 17, wherein the audio and the images are segmented at spectrum thresholds.”

209. A POSA would have found claim 17 obvious over the combination of Fontana with Lau, as ~~discussed~~discuss above for Ground 1. Additionally, for the same reason as discussed above for claim 9, it would have been obvious to a POSA to combine with Arakawa to **segment the audio and the images at spectrum thresholds**.—(Houh ¶209.)