

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner,

v.

CELLULAR SOUTH, INC.,
Patent Owner.

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

Title: VIDEO TO DATA

DECLARATION OF HENRY HOUH, PH.D.

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

TABLE OF CONTENTS

	Page
I. INTRODUCTION AND QUALIFICATIONS	6
A. Qualifications and Experience	6
B. Materials Considered.....	14
II. PERSON OF ORDINARY SKILL IN THE ART	15
III. STATEMENT OF LEGAL PRINCIPLES.....	18
A. Claim Construction	18
B. Obviousness (§ 103).....	20
IV. THE '954 PATENT.....	25
A. Overview of the Specification.....	25
B. The Challenged Claims	31
V. APPLICATION OF THE PRIOR ART TO CHALLENGED CLAIMS	32
A. Brief Summary and Overview of Prior Art.....	33
1. References Common to All Grounds.....	33
(a) Kritt (EX1003).....	33
2. Additional References.....	37
(a) Fontana (EX1004)	37
(b) Lau (EX1005)	41
B. Ground 1: Claims 1-8 and 11-12 Are Obvious Over Kritt in view of Fontana	44
1. Independent Claim 1: “A method to generate video data from a video comprising:” (Claim 1[pre]).....	44
(a) “generating audio files and image files from the video;” (Claim 1[a]).....	47

TABLE OF CONTENTS

(continued)

	Page
(b) “distributing the audio files and the image files across a plurality of processors and processing the audio files and the image files in parallel;” (Claim 1[b])	58
(c) “converting audio files associated with the video to text;” (Claim 1[c])	66
(d) “identifying an object in the image files;” (Claim 1[d])	67
(e) “determining a contextual topic from the image files;” (Claim 1[e]).....	69
(f) “assigning a probability of accuracy to the identified object based on the contextual topic;” (Claim 1[f])	75
(g) “converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic;” (Claim 1[g])	79
(h) “cross-referencing the text and the video data with the video to determine contextual topics;” (Claim 1[h])	88
(i) “generating a contextual text, an image, or an animation based on the determined contextual topics;” (Claim 1[i]).....	93
(j) “generating a content-rich video based on the generated text, image, or animation.” (Claim 1[j])	97
2. Claim 2: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”	100
3. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to extract a topic.”	104

TABLE OF CONTENTS
(continued)

	Page
4. Claim 4: “The method according to claim 1, further comprising: processing the image files to extract additional text.”	104
5. Claim 5: “The method according to claim 4, wherein the additional text is generated by segmenting the image files before processing the image files in parallel.”	105
6. Claim 6: “The method according to claim 1, further comprising: processing the image files to extract an object and determine a motion associated with the object.”	110
7. Claim 7: “The method according to claim 1, wherein the text is generated by segmenting the audio before processing the audio files in parallel.”	111
8. Claim 8: “The method according to claim 6, wherein the audio files and the image files are segmented at spectrum thresholds.”	116
9. Claim 11: “The method according to claim 1, further comprising: generating metadata.”	120
10. Claim 12: “The method according to claim 11, wherein the metadata comprises a number of frames, a range of said frames and/or timestamp references.”	121
C. Ground 2: Claims 2-3 and 9-10 Are Obvious Over Kritt in view of Fontana and Lau	122
1. Claim 2: “The method according to claim 1, further comprising: applying natural language processing to the text to determine context associated with the video.”	122
2. Claim 3: “The method according to claim 1, further comprising: applying natural language processing to the text to extract a topic.”	126
3. Claim 9: “The method according to claim 1, further comprising: generating an advertisement based on the text and the video data.”	126

TABLE OF CONTENTS
(continued)

	Page
4. Claim 10: “The method according to claim 9, further comprising: placing the advertisement in the video at a preferred time.”	130
D. Ground 3: Claim 13 Is Obvious Over Kritt.....	131
1. Independent Claim 13: “A system for extracting data from a video, comprising:” (Claim 13[pre]).....	131
(a) “a computer processor; and” (Claim 13[a]).....	131
(b) “a non-transitory computer readable medium containing instructions directing the system to execute the steps of:” (Claim 13[b]).....	132
(c) “converting audio associated with the video to text;” (Claim 13[c])	133
(d) “converting images associated with the video to video data, wherein the video data comprises identified objects;” (Claim 13[d])	133
(e) “extracting topics from the text and the video data;” (Claim 13[e]).....	134
(f) “assigning probabilities of accuracy for each of the identified objects, wherein the probabilities are based on the extracted topics;” (Claim 13[f]).....	135
(g) “cross-referencing the text, the video data, and the topics with the video to determine contextual topics;” (Claim 13[g]).....	136
(h) “generating a contextual text, an image, or an animation based on the determined contextual topics; and” (Claim 13[h]).....	136
(i) “generate a content-rich video based on the generated contextual text, image, or animation.” (Claim 13[i]).....	136
V. NO SECONDARY CONSIDERATIONS OF NONOBVIOUSNESS.....	137
VI. CONCLUSION.....	137

I, Henry Houh, Ph.D., declare as follows:

I. INTRODUCTION AND QUALIFICATIONS

A. Qualifications and Experience

1. My complete qualifications and professional experience are described in my *Curriculum Vitae*, a copy of which can be found in Appendix A. The following is a brief summary of my relevant qualifications and professional experience.

2. I received a Ph.D. in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology (“MIT”) in 1998. Beforehand, I received a Master of Science degree in Electrical Engineering and Computer Science in 1991, a Bachelor of Science degree in Electrical Engineering and Computer Science in 1989, and a Bachelor of Science degree in Physics in 1990, all from MIT.

3. I am currently self-employed as an independent technical consultant. From its founding in 2012 until October 2022, I was also president of a company that provides supplemental science, technology, engineering, and mathematics (“STEM”) education to children of all ages.

4. I first entered telecommunications in 1987 when I worked as a summer intern at AT&T Bell Laboratories as part of a five-year dual degree program at MIT. I continued to work at AT&T Bell Laboratories as part of this MIT program. While at MIT, I was a teaching assistant (“TA”) in the Electrical Engineering and Computer

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

Science Department's core Computer Architectures course. I first was a TA as a senior for a role typically reserved for graduate students. I later became head TA. The course covered various topics in computer architectures. As a TA, I helped write homework assignments, lab assignments, and exams. I also taught in some of the recitation sections.

5. Later, as part of my doctoral research at MIT from 1991-1998, I was a research assistant in the Telemedia Network Systems ("TNS") group at the Laboratory for Computer Science. The TNS group built a high-speed gigabit network and created applications that ran over the network. Example applications included ones for remote video capture, processing, and display of video on computer terminals. In addition to working on the design of core network components, designing and building the high-speed links, and designing and writing the device drivers for the interface, I also set up the group's web server. Also, I helped to maintain, install, and upgrade the networking devices used within the group, along with other graduate students.

6. I also helped to build the web pages that initiated the above-mentioned video sessions via a web interface. Vice President Al Gore visited our group in 1996 and received a demonstration of—and remotely drove—a radio-controlled toy car with a wireless video camera mounted on it that was built by our group. This toy car

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

device received commands transmitted over a network from a remote computer, and video data from the toy car was transmitted wirelessly then over a computer network back to the user controller. On occasion, we allowed users visiting our web site to drive the toy car from their remote computer while they watched the video on their computer. The video stream was encoded by TNS-designed hardware, streamed over the TNS-designed network, and displayed using TNS-designed software.

7. Our video environment included an extensive amount of video processing. The closed captions were extracted from the analog video broadcasts, and our group built an extensive toolkit to perform video processing functions on the live video as it was being captured and optionally recorded. Our system included features such as motion detection, template matching, model matching, live green screening, face detection, histogram generation and matching, feature matching, edge detection, scene change detection, image transformation, filtering, image segmentation, object motion tracking, and other features.

8. I defended and submitted my Ph.D. thesis, titled “Designing Networks for Tomorrow’s Traffic,” in January 1998. As part of my thesis research, I analyzed local area and wide area flows to show a more efficient method for routing packets in a network, based on traffic patterns at the time. The traffic flow data included Ethernet, IP, TCP or UDP, and RTP header information, which I analyzed to come

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

to the conclusions in my thesis.

9. While I was in graduate school, I started a company that provided web site design services and targeted advertising opportunities. We also sold targeted advertising placements. As the web became popular first at colleges and universities, we focused initially on building college recruiting web sites and sold advertisements to companies wishing to target students to recruit at specific colleges. We developed technology that utilized the web user's college affiliation to customize advertisements targeted to students at that college.

10. From 1997 to 1999, I was a Senior Scientist and Engineer at NBX Corporation, a start-up that made business telephone systems for streaming packetized audio over data networks instead of using traditional telephone lines. NBX was later acquired by 3Com Corporation and the phone system is still used today by numerous businesses.

11. As part of my work at NBX, I designed the core audio reconstruction algorithms for the telephones, as well as the packet transmission algorithms. I also designed and validated the core packet transport protocol used by the phone system. The protocol was used for all signaling in the phone system, including for the setup of conference calls. The NBX system also featured a computer interface for initiating phone calls, which could also initiate conference calls. The NBX system also

supported the Telephony Application Programming Interface (“TAPI”) that allowed other computer programs to integrate with our system telephony features. We obtained U.S. Patent No. 6,967,963, entitled “Telecommunication method for ensuring on-time delivery of packets containing time-sensitive data,” as part of this work. The first release of the phones had integrated Ethernet repeater hubs embedded in them, and I worked with the documentation and field support people to develop sets of guidelines so that customers would not build their networks to violate the Ethernet specifications, typically by chaining too many hub repeaters and NBX phones together. The first release of the phones also worked using Ethernet-layer packets (without the use of IP). I also programmed the first prototype of our phone which communicated using IP, and I demonstrated our IP phones working over the Internet when we attended a trade show in California. The phone connected over the Internet to our headquarters in Andover, MA.

12. From 1999-2004, I was employed by Empirix or its predecessor company, Teradyne. Teradyne had traditionally made products to test chips and hardware, but the founder and board chairman, Alex d’Arbeloff wanted to expand the company’s products to include software systems testing. Teradyne made a number of acquisitions to achieve this goal, and I was hired to work for Mr. d’Arbeloff to work with all the acquired software test divisions. These divisions

were later spun-out of Teradyne into a separate company, Empirix. Empirix was a leader in test tools for software systems including telecommunications protocols and systems, providing functional testing tools as well as load testing tools. From 2000-2001, I conceived and built a test platform for testing Voice-over-IP (VoIP). The first application on this new test platform was a cloud emulator for simulating the effects of transmitting VoIP over a busy network. The test platform was based on a network processor chip, which could be programmed to cut-through packets while processing packet data such as various protocol layer headers including addresses included therein and even packet data contents. I also designed a protocol analyzer built on the same platform. The application captured and performed protocol decoding at various layers in the protocol stacks of captured packets, including detailed Ethernet header decoding, IP header decoding, TCP header decoding, UDP header decoding, RTP header decoding, and many other specified protocols. The application was also designed to reconstruct entire conversations that spanned multiple packets.

13. In 2006, as part of my role at BBN Technologies, I helped found PodZinger Inc., now known as RAMP Inc. PodZinger utilized BBN's speech recognition algorithms to search through the spoken words in audio and video segments. Our speech recognition algorithms were also capable of extracting the

emotional prosody of the speech input. While I was Vice President of Operations and Technology, PodZinger followed its initial prototype with a full streaming audio and video search solution. The metadata generated from speech recognition process was used in the search result user interface so that the user could see highlights of their search term in the audio transcript, the context around their search term, and the user could select a word in the transcript to jump the video to play from exactly the point where the selected word was uttered.

14. I also created a social networking web site, which BBN sold to a venture-funded startup company. In the process of creating the web site, I designed and specified the authentication and authorization protocols.

15. I applied for and received a number of patents in the above areas. In particular, U.S. Patent No. 9,697,230 describes “an automated method and apparatus for generating metadata enhanced for audio/video search-driven applications. The apparatus includes a media indexer that obtains an media file/stream (e.g., audio/video podcasts), applies one or more automated media processing techniques to the media file/stream, combines the results of the media processing into metadata enhanced for audio/video search, and stores the enhanced metadata in a searchable index or other data repository.” EX1007, 4:31-39. A media indexer “applies a predefined set of audio or video processing routines to derive a portion of the

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

enhanced metadata from the media content.” *Id.* at 5:19-21. The indexer utilizes speech recognition, natural language processing, and video frame analyzers, that can recognize scenes, watermarks, objects, faces, and overlay text. *Id.* at Fig. 1B, 5:22-34, 6:43-57. Related advertising may then be displayed utilizing the results of such an analysis. (*Id.*, 15:52-16:25.) A partial list of my patents and patent applications include:

- U.S. Patent No. 9,697,231, H. Houh, J. Stern, R. Spina, M. Meteer [sic], “Methods and apparatus for providing virtual media channels based on media search,” issued July 4, 2017.
- U.S. Patent No. 9,697,230, H. Houh, J. Stern, “Methods and apparatus for dynamic presentation of advertising, factual, and informational content using enhanced metadata in search-driven media applications,” issued July 4, 2017. *See also* WO2007056485.
- U.S. Patent No. 7,801,910, H. Houh and J. N. Stern, “Method and apparatus for timed tagging of media content,” issued September 21, 2010. *See also* US2007/0112837, US2009/0222442, WO2007056535.
- U.S. Patent Application Publication No. 2007/0106660, “Method and apparatus for using confidence scores of enhanced metadata in search-driven media applications.”

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

- U.S. Patent Application Publication No. 2007/0118873, “Methods and apparatus for merging media content.”
- U.S. Patent Application Publication No. 2009/0222442, “User-directed navigation of multimedia search results.”
- U.S. Patent Application Publication No. 2007/0106646, “User-directed navigation of multimedia search results.”

16. I have been retained by counsel for Petitioner to provide my expert opinion in connection with the above-captioned proceeding as set forth herein.

B. Materials Considered

17. The analysis that I provide in this Declaration is based on my education, research, and experience, as well as the documents I have considered. In forming my opinions, I read and considered U.S. Patent No. 10,218,954 B2 (“954 patent”) (**EX1001**) and its prosecution history. I cite to the following documents in my analysis below:

Exhibit No.	Description of Document
1001	U.S. Patent No. 10,218,954 B2 to Naeem Lakhani (PCT filed February 7, 2015, issued February 26, 2019) (“954” or “954 patent”)
1003	U.S. Patent App. Pub. No. 2014/0181668 A1 to Barry A. Kritt et al. (filed Dec. 20, 2012, published Jun. 26, 2014) (“Kritt”)

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

Exhibit No.	Description of Document
1004	U.S. Patent App. Pub. No. 2012/0078712 A1 to James A. Fontana et al. (filed Sept. 27, 2010, published Mar. 29, 2012) (“ Fontana ”)
1005	U.S. Patent App. Pub. No. 2007/0112630 A1 to Wai Kit Lau et al. (filed Nov. 7, 2006, published May 17, 2007) (“ Lau ”)
1006	Prosecution History of U.S. Patent No. 10,218,954 B2
1007	U.S. Patent No. 9,697,230 to Henry Houh et al. (filed Mar. 31, 2006, issued Jul. 4, 2017)
1008	U.S. Patent No. 8,706,664 B2 to Oliver B. Down et al. (filed May 3, 2012, issued Apr. 22, 2014) (“ Downs ”)
1009	U.S. Patent No. 9,245,254 B2 to Richard T. Lord et al. (filed Feb. 15, 2012, issued Jan. 26, 2016) (“ Lord ”)

II. PERSON OF ORDINARY SKILL IN THE ART

18. The face page of the ’954 patent shows that the earliest patent application to which the ’954 patent claims priority was filed August 15, 2013 (Provisional Application No. 61/866,175).¹ Therefore, I understand that the claims of the ’954 patent are subject to the patent laws in effect under the America Invents

¹ I express no opinion on whether the ’954 patent is actually entitled to the benefit of either of the provisional applications to which it claims priority (61/866,175 and 62/021,666). This issue is not relevant here because all of the prior art references relied upon predate the provisional application filing dates. In the event entitlement to a provisional filing date may later become relevant, for example based on arguments later presented by Patent Owner, I reserve my right to provide opinions on that issue at that time.

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

Act (“AIA”), including Sections 102 and 103 regarding invalidity based on prior art, which apply to patent applications having an effective filing date on or after March 16, 2013. I understand that under the AIA, an assessment of claims of a patent should be undertaken from the perspective of a person of ordinary skill in the art is to be made as of the effective filing date of the patent (i.e., the earliest of the actual filing date of the patent or application, or the filing date of the earliest application with priority). I note that, for patents subject to pre-AIA patent laws, such an assessment should be made as of the earliest claimed priority date (i.e., the “time the invention was made”). For purposes of my analysis, whether the “effective filing date” or “earliest claimed priority date” applies does not affect my analysis. This is because I understand that the alleged “effective filing date” and “earliest claimed priority date” are the same in this case—August 15, 2013.

19. I have also been advised that to determine the appropriate level of a person having ordinary skill in the art, the following factors may be considered: (1) the types of problems encountered by those working in the field and prior art solutions thereto; (2) the sophistication of the technology in question, and the rapidity with which innovations occur in the field; (3) the educational level of active workers in the field; and (4) the educational level of the inventor.

20. The ’954 patent states that it “relates to a method and a system for

generating various and useful data from videos.” (’954, 1:18-19; *see also id.*, 1:36-38 (“The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.”).)

21. In my opinion, a person of ordinary skill in the art as of August 2013 would have possessed a bachelor’s degree in electrical engineering, computer science, or similar field, with two years of experience in developing and implementing computer software for processing and/or analyzing multimedia content, such as audio, video, or image data. A person could also have qualified as a person of ordinary skill in the art with some combination of (1) more formal education (such as a master’s of science degree) and less technical experience, or (2) less formal education and more technical or professional experience.

22. My opinions regarding the level of ordinary skill in the art are based on, among other things, my experience in the field of computer science, my understanding of the basic qualifications that would be relevant to an engineer or scientist tasked with investigating methods and systems in the relevant area, and my familiarity with the backgrounds of colleagues, co-workers, and employees, both past and present.

23. Although my qualifications and experience exceed those of the hypothetical person having ordinary skill in the art defined above, my analysis and

opinions regarding the '954 patent have been based on the perspective of a person of ordinary skill in the art as of August 2013.

III. STATEMENT OF LEGAL PRINCIPLES

A. Claim Construction

24. I understand that a purpose of claim construction is to determine what a person of ordinary skill in the art would have understood the claim terms to mean. Claim terms are generally given their ordinary and customary meaning, which is the meaning that the term would have to a person of ordinary skill in the art in question as of the effective filing date.

25. I understand that the person of ordinary skill in the art is deemed to read the claim term not only in the context of the particular claim in which the disputed term appears, but in the context of the entire patent, including the specification. I understand that the patent specification, under the legal principles, has been described as the single best guide to the meaning of a claim term, and is thus highly relevant to the interpretation of claim terms. And I understand for terms that do not have a customary meaning within the art, the specification usually supplies the best context of understanding the meaning of those terms.

26. I further understand that other claims of the patent in question, both asserted and unasserted, can be valuable sources of information as to the meaning of

Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

a claim term. Because the claim terms are normally used consistently throughout the patent, the usage of a term in one claim can often illuminate the meaning of the same term in other claims. Differences among claims can also be a useful guide in understanding the meaning of particular claim terms.

27. I understand that the prosecution history can further inform the meaning of the claim language by demonstrating how the inventors understood the invention and whether the inventors limited the invention in the course of prosecution, making the claim scope narrower than it otherwise would be. Extrinsic evidence, such as dictionaries, may also be consulted in construing the claim terms.

28. I understand that, in *Inter Partes* Review (IPR) proceedings, a claim of a patent shall be construed using the same claim construction standard that would be used to construe the claim in a civil action filed in a U.S. district court (which I understand is called the “*Phillips*” claim construction standard), including construing the claim in accordance with the ordinary and customary meaning of such claim as understood by one of ordinary skill in the art and the prosecution history pertaining to the patent.

29. I have been instructed by counsel to apply the “*Phillips*” claim construction standard for purposes of interpreting the claims in this proceeding, to the extent they require an explicit construction. The description of the legal

principles set forth above thus provides my understanding of the “*Phillips*” standard as provided to me by counsel.

30. For purposes of my analysis here, I do not believe express claim constructions are necessary because the prior art renders the claims obvious under any reasonable construction.

B. Obviousness (§ 103)

31. I understand that a patent claim is obvious if, as of the effective filing date, it would have been obvious to a person having ordinary skill in the field of the technology (the “art”) to which the claimed subject matter belongs.

32. I understand that the following factors should be considered in analyzing obviousness: (1) the scope and content of the prior art; (2) the differences between the prior art and the claims; and (3) the level of ordinary skill in the pertinent art. I also understand that certain other facts known as “secondary considerations” such as commercial success, unexplained results, long felt but unsolved need, industry acclaim, simultaneous invention, copying by others, skepticism by experts in the field, and failure of others may be utilized as indicia of nonobviousness. I understand, however, that secondary considerations should be connected, or have a “nexus,” with the invention claimed in the patent at issue.

33. I understand that a reference qualifies as prior art for obviousness

purposes when it is analogous to the claimed invention. The test for determining what art is analogous is: (1) whether the art is from the same field of endeavor, regardless of the problem addressed, and (2) if the reference is not within the field of the inventor's endeavor, whether the reference still is reasonably pertinent to the particular problem with which the inventor is involved.

34. I understand that a person of ordinary skill in the art is assumed to have knowledge of all prior art. I understand that one skilled in the art can combine various prior art references based on the teachings of those prior art references, the general knowledge present in the art, or common sense. I understand that a motivation to combine references may be implicit in the prior art, and there is no requirement that there be an actual or explicit teaching to combine two references. Thus, one may take into account the inferences and creative steps that a person of ordinary skill in the art would employ to combine the known elements in the prior art in the manner claimed by the patent at issue. I understand that one should avoid "hindsight bias" and *ex post* reasoning in performing an obviousness analysis. But this does not mean that a person of ordinary skill in the art for purposes of the obviousness inquiry does not have recourse to common sense.

35. I understand that when determining whether a patent claim is obvious in light of the prior art, neither the particular motivation for the patent nor the stated

purpose of the patentee is controlling. The primary inquiry has to do with the objective reach of the claims, and that if those claims extend to something that is obvious, then the entire patent claim is invalid.

36. I understand one way that a patent can be found obvious is if there existed at the time of the invention a known problem for which there was an obvious solution encompassed by the patent's claims. I understand that a motivation to combine various prior art references to solve a particular problem may come from a variety of sources, including market demand or scientific literature. I understand that a need or problem known in the field at the time of the invention can also provide a reason to combine prior art references and render a patent claim invalid for obviousness. I understand that familiar items may have obvious uses beyond their primary purpose, and that a person of ordinary skill in the art will be able to fit the teachings of multiple prior art references together like the pieces of a puzzle. I understand that a person of ordinary skill is also a person of at least ordinary creativity. I understand when there is a design need or market pressure to solve a problem and there are a finite number of identified, predictable solutions, a person of ordinary skill has good reason to pursue the known options within his or her technical grasp. If this finite number of predictable solutions leads to the anticipated success, I understand that the invention is likely the product of ordinary skill and

common sense, and not of any sort of innovation. I understand that the fact that a combination was obvious to try might also show that it was obvious, and hence invalid, under the patent laws. I understand that if a patent claims a combination of familiar elements according to known methods, the combination is likely to be obvious when it does no more than yield predictable results. Thus, if a person of ordinary skill in the art can implement a predictable variation, an invention is likely obvious. I understand that combining embodiments disclosed near each other in a prior art reference would not ordinarily require a leap of inventiveness.

37. I understand that obviousness may be shown by demonstrating that it would have been obvious to modify what is taught in a single piece of prior art to create the patented invention. Obviousness may also be shown by demonstrating that it would have been obvious to combine the teachings of more than one item of prior art. I understand that a claimed invention may be obvious if some teaching, suggestion, or motivation exists that would have led a person of ordinary skill in the art to combine the invalidating references. I also understand that this suggestion or motivation may come from the knowledge of a person having ordinary skill in the art, or from sources such as explicit statements in the prior art. I understand that when there is a design need or market pressure, and there are a finite number of predictable solutions, a person of ordinary skill may be motivated to apply common

sense and his skill to combine the known options in order to solve the problem.

38. I understand the following are examples of approaches and rationales that may be considered in determining whether a piece of prior art could have been combined with other prior art or with other information within the knowledge of a person having ordinary skill in the art:

- (1) Some teaching, motivation, or suggestion in the prior art that would have led a person of ordinary skill to modify the prior art reference or to combine prior art reference teachings to arrive at the claimed invention;
- (2) Known work in one field of endeavor may prompt variations of it for use in the same field or a different field based on design incentives or other market forces if the variations would have been predictable to a person of ordinary skill in the art;
- (3) Combining prior art elements according to known methods to yield predictable results;
- (4) Applying a known technique to a known device, method, or product ready for improvement to yield predictable results;
- (5) Applying a technique or approach that would have been “obvious to try” (choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success);

(6) Simple substitution of one known element for another to obtain predictable results; or

(7) Use of a known technique to improve similar products, devices, or methods in the same way.

39. I understand that, when determining whether a claimed combination is obvious, the correct analysis is not whether one of ordinary skill in the art, writing on a blank slate, would have chosen the particular combination of elements described in the claim. Instead, I understand the correct analysis considers whether one of ordinary skill, facing the wide range of needs created by developments in the field of endeavor, would have seen a benefit to selecting the combination claimed.

40. I understand that the test for obviousness is not whether the features of a secondary reference may be bodily incorporated into the structure of the primary reference. The test for obviousness, in other words, is not whether the references could be physically combined but whether the claimed inventions are rendered obvious by the teachings of the prior art as a whole.

IV. THE '954 PATENT

A. Overview of the Specification

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

additional content-matched audio-visual information.

42. The '954 patent describes its purpose at a very high level. It states as its “Technical Field” that it “relates to a method and a system for generating various and useful data from videos.” ('954, 1:18-19.)² And the summary of the invention confirms the broad brush the patent purports to take: “The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.” ('954, 1:36-38.)

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

44. The specification of the '954 patent provides primarily a series of high-level, functional explanations for how to implement the alleged invention, with scant information on how to carry out any of its steps. For example, Figure 1 of the '954

² All underlining added to quoted material unless noted.

patent, shown below, “illustrates an embodiment of present invention,” and specifically “[a]n embodiment of video-to-content engine operation”:

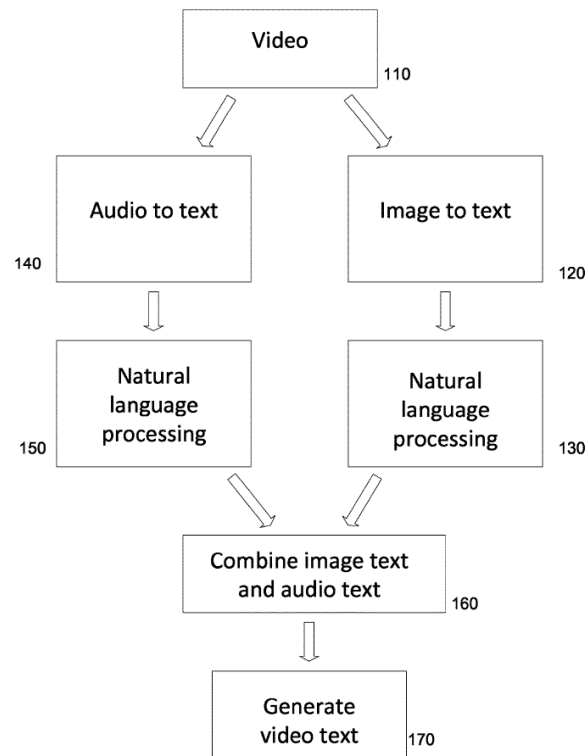


FIG. 1

(’954, Fig. 1, 2:24, 4:9-10.) At step **110**, “a video stream is presented.” (’954, 4:10.) “The engine can extract audio data and image data (e.g. images or frames forming the video) from the video stream.” (’954, 4:16-17.) At step **120**, “the video-to-content engine performs image data processing on the video stream,” such as symbol (or object) detection and identification, such as identifying a dog in the video image data. (’954, 4:22-23, 4:28-35.) Notably, the ’954 patent explains that this may be performed by an “image processing technique such as color edge detection,” but

does not explain how to carry out color edge detection. ('954, 4:29-31.) A POSA, however, would have recognized color edge detection as one of many well-known methods for image processing available in 2013. ('954, 1:23-32 (acknowledging image identification was known).) Then, “[b]ased on the identified symbol, a plurality of instances of the symbol can be compared to a topic database to identify a topic (such as an event)[,]” such as identifying the dog as running or jumping. ('954, 4:44-47.) Thus, “text describing a symbol of the video and topic relating to the symbol may be generated[.]” ('954, 4:50-52.) At step **130**, “natural language processing can be applied to the image text.” ('954, 6:31-32.) “For example, based on dictionary, grammar, and a knowledge database, the text extracted from video images can be modified as the video-to-content engine selects primary semantics from a plurality of possible semantics.” ('954, 6:32-36.) Applying dictionary definitions and grammatical rules to text was nothing new in 2013.

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

1:23-32 (acknowledging audio-to-text was known).)

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

47. At step **170**, “the video-to-content engine generates video text, such as text describing the content of the video, using the result of the combined texts and cross reference.” (’954, 7:44-47.) This is just summarizing the video, such as what a theater critic or museum curator would do for a play or a work of multimedia art. “For example, key words indicating topic and semantic that appear in both texts can be selected or emphasized.” (’954, 7:47-49.) An “advertisement (text, images, or animation)” can then be generated based on the video text and added to the video stream. (’954, 8:60-63.)

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

As one example, if the audio soundtrack of a video is an announcer

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

[T]he linear sequencing of frames from a video—as opposed to a stand-alone image—avails itself to a set images {images x-y} from which context can be derived. This contextual methodology can be viewed as systematic detection of probable image false positives by identifying an object from one video frame (or image) as an anomaly when compared to and associated with a series of image frames both prior and subsequent to the purported anomaly. According to the objects, faces, words, etc. of a given set of frames (however defined), a probability can be associated with an identified anomaly to determine whether an image may be a false positive and, if so, what other likely results should be.

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

49. I discuss additional aspects of the ’954 patent in my analysis of the claims below.

B. The Challenged Claims

50. This Declaration addresses claims 1-13 of the '954 patent. Independent claim 1 is representative and 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 audio files and the image files across a plurality of processors and processing the audio files and the image files in parallel;
 - [c] converting audio files associated with the video to text;
 - [d] identifying an object in the image files;
 - [e] determining a contextual topic from the image files;
 - [f] assigning a probability of accuracy to the identified object based on the contextual topic;
 - [g] converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic;
 - [h] cross-referencing the text and the video data with the video to determine contextual topics;
 - [i] generating a contextual text, an image, or an animation based on the determined contextual topics;

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

(’954, 13:66-14:18 (Claim 1 (bracketed notation (e.g., “[a]”) added).)

51. I address the claims further in my detailed analysis in **Part V** below.

V. APPLICATION OF THE PRIOR ART TO CHALLENGED CLAIMS

52. I have reviewed and analyzed the prior art references and materials listed in **Part I.B** above. In my opinion the claims of the ’954 patent would have been obvious to a person of ordinary skill in the art based on the following combinations of the prior art.

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

53. I am informed by counsel that each of the references cited in the grounds above qualifies as prior art to the challenged claims because each reference was filed and/or published before the earliest filing date to which the ’954 patent can claim priority (August 15, 2013).

A. Brief Summary and Overview of Prior Art

1. References Common to All Grounds

(a) Kritt (EX1003)

54. **Kritt**, U.S. Patent Application Publication No. 2014/0181668 A1, is entitled “Visual Summarization of Video for Quick Understanding.” I am informed that Kritt qualifies as prior art because it was both filed (on December 20, 2012) and published (on June 26, 2014) before the earliest priority date of the challenged claims. Kritt was not cited during prosecution of the ’954 patent.

55. Like the ’954 patent, Kritt describes a system for processing videos to generate data about the videos. For example:

According to various embodiments, a viewer selects a video, and human, nonhuman, and audio objects of the video are identified. In addition, key words that are spoken by human objects in the video are identified. Human, nonhuman, and audio objects may be used to classify a particular segment of a video as a scene. The objects and key words are then associated with the scenes of the video.

(Kritt, ¶0017.) As Kritt notes in the passage above, both video and audio are processed. (*See also* Kritt, ¶¶0043 (“visual objects in a shot may be identified by application of one or more known image recognition processes to the shot”), 0059 (“Segments of the video for which the audio feature values are similar to those that

are characteristic of speech may be classified as speech.”), 0065 (“In operation **612**, an audio transcript may be determined. An audio transcript may include all of the words spoken in the video...[S]poken words may be determined from audio features classified as speech using any known technique.”.) Kritt uses the results of its audio processing to identify “key words:” “A key word may be a word that provides significant information content about a scene in a video.” (Kritt, ¶0016.)

56. Kritt then uses the results of its video and audio processing to identify scenes within the video, and associates the identified objects and keywords with specific scenes: “Human, nonhuman, and audio objects may be used to classify a particular segment of a video as a scene. The objects and key words are then associated with the scenes of the video.” (Kritt, ¶0017.) Kritt further uses the results of its audio and video processing “to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.” (Kritt, ¶0017.) Kritt additionally teaches comparing the results of its analyses to verify that the objects and scenes identified are accurate. For example, “a visual tag may indicate that a particular human object appears in a shot and an audio tag identifies an audio signature of the human object is associated with the shot. In these examples, if the tags that are compared indicate the same object, the positive or consistent result of the comparison may be used in operation **310** to validate that the

human object was correctly identified. . . . Alternatively, if a probability or confidence parameter associated with the particular tag is above a threshold, it may be determined that the object was correctly identified and that the shot or scene includes multiple objects. In this circumstance, the modification may include adding an indication to the tag that it may be relied on to a particular extent.” (Kritt, ¶0033.)

57. Finally, aspects of the video and audio processing and subsequent analysis are displayed to the user. (See Kritt, ¶¶0017 (“In various embodiments, visual representations of key words, and human, nonhuman, and audio objects summarizing all of the scenes of a video are presented to the viewer in a single display screen.”), 0018 (“In one embodiment, a display screen may include a first window for playing the video and a second window for rendering text, symbols, and icons corresponding with human, nonhuman, audio, and emotion objects, and key words.”).)

58. Kritt describes modules for processing video and audio of a video file, and that those modules can be distributed on different computer systems:

The memory **104** may store all or a portion of the following: an audio visual file container **150** (shown in FIG. 2 as container **202**), a video processing module **152**, an audio processing module **154**, and a control module **156**. These modules are illustrated as being included within the memory **104** in the computer system **100**, however, in other

embodiments, some or all of them may be on different computer systems and may be accessed remotely, e.g., via a network.

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

59. Kritt further describes how its systems could be provided using “cloud-computing infrastructure.” (Kritt, ¶¶0078-0079.) Kritt explains, as was known to a person of ordinary skill in the art, that “[c]loud computing generally refers to the provision of scalable computing resources as a service over a network.” (Kritt, ¶0078.) “More formally,” Kritt explains, “cloud computing may be defined as a computing capability that provides an abstraction between the computing resource and its underlying technical architecture (e.g., servers, storage, networks), enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.” (Kritt, ¶0078.) Kritt further explains that one benefit of cloud computing is that “[t]ypically, cloud-computing resources are provided to a user on a pay-per-use basis, where users are charged only for the computing resources actually used (e.g., an amount of storage space used by a user

or a number of virtualized systems instantiated by the user).” (Kritt, ¶0079.) It also allows a user to “access any of the resources that reside in the cloud at any time, and from anywhere across the Internet.” (Kritt, ¶0079.)

60. I will provide more information about Kritt in my discussion of the claim limitations below.

2. Additional References

(a) Fontana (EX1004)

61. **Fontana**, U.S. Patent Application Publication No. 2012/0078712 A1 to James A. Fontana et al., is entitled “Systems and Methods for Processing and Delivery of Multimedia Content.” I am informed that Fontana qualifies as prior art because it was filed on September 27, 2010, and was published on March 29, 2012, both before the earliest priority date of the challenged claims. Fontana was not cited during prosecution of the ’954 patent.

62. Fontana is directed to “[m]ethods and systems for providing multimedia content[.]” (Fontana, p.1 (Abstract).) Like the ’954 patent, Fontana discloses a method that generates images and audio files from a video, performs speech-to-text recognition of the audio, detects objects within the images, and generates and searches within meta-data about the video. Fontana teaches:

[A] method for providing multimedia content includes receiving

multimedia content from a source, and applying a container to the multimedia content using one or more computing systems, the container defining an interface through which metadata external to the multimedia content is linked with one or more portions of the multimedia content. The method includes constructing a set of text metadata describing an audio portion of the multimedia content, generating a set of object metadata describing at least a portion of one or more objects appearing in the multimedia content, and returning the container and multimedia content to the source. The method further includes storing the text metadata and object metadata at the one or more computing systems, and upon notification of a request for playback of the multimedia content at the one or more computing systems, providing the text metadata and the object metadata associated with the container for synchronized use during playback of the multimedia content via the container.

(Fontana, ¶0010.)

63. An embodiment of this method is depicted, for example, in Figure 6, which shows “Audio Processing” at 606, “Video Processing” at 608, “Full-Text Search” at 630, and “Video Metadata” at 634.

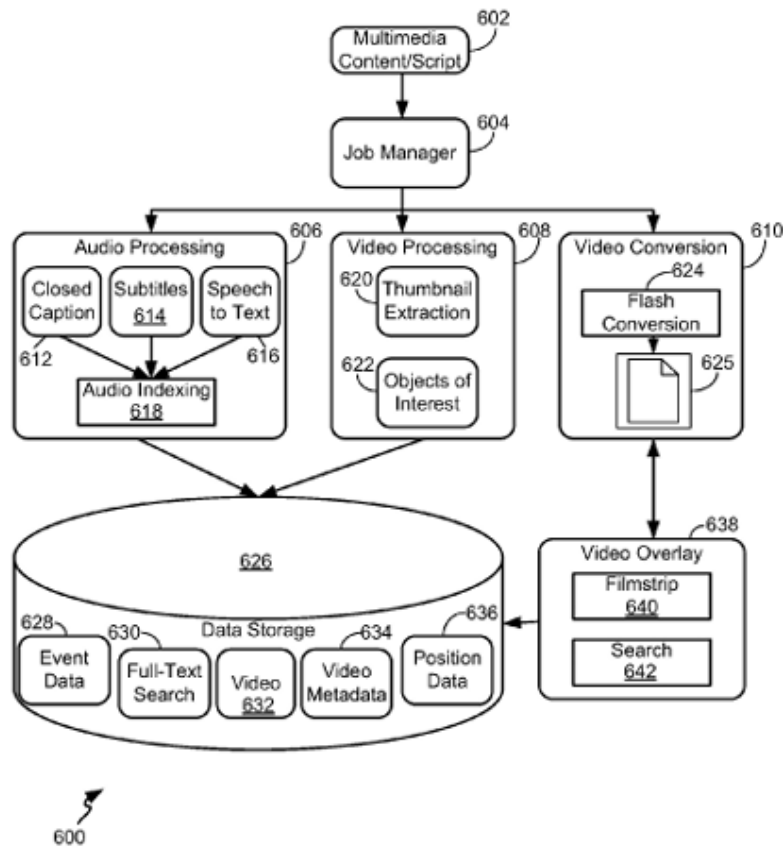


FIG. 6

(Fontana, Fig. 6.)

64. Fontana additionally teaches implementing its method in “a plurality of distributed computing systems.” (Fontana, ¶0045.) In one embodiment shown in Figure 3, “the distributed computing network **300** includes a plurality of computing systems, illustrated as servers **302 a-c**. The servers **302 a-c** are communicatively interconnected, and each includes a corresponding data storage system **304 a-c**. The servers **302 a-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 **304 a-c**. The servers **302 a-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.” (Fontana, ¶0052.) “[O]ne or more of the servers **302 a-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 **302 a-c** as well. The servers **302 a-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; Fig. 3.)

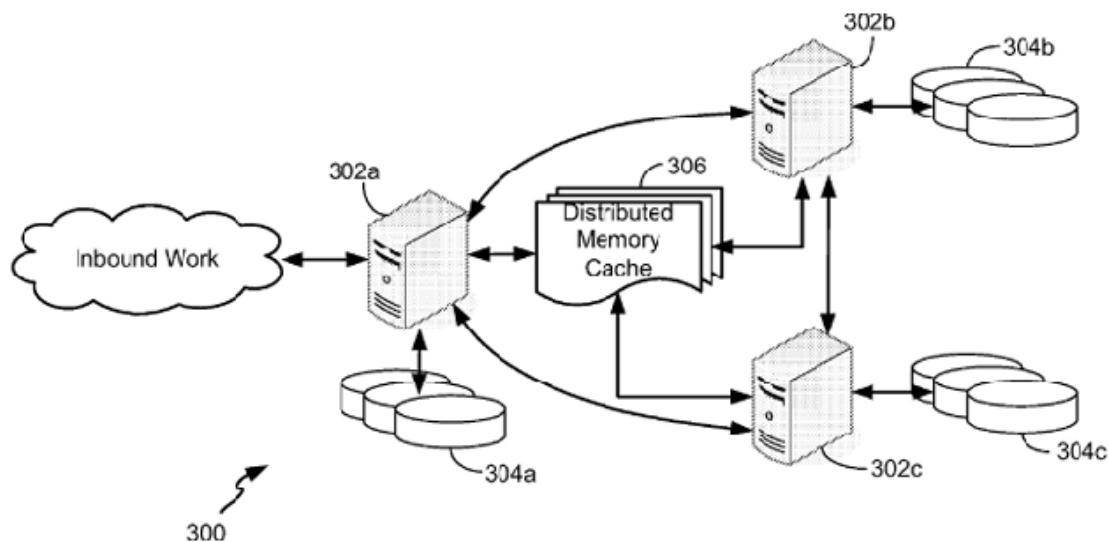


FIG. 3

(Fontana, Fig. 3.)

65. I will provide more information about Fontana in my discussion of the claim limitations below.

(b) Lau (EX1005)

66. **Lau**, U.S. Patent Application Publication No. 2007/0112630 A1 to Wai Kit Lau et al., is entitled “Techniques for Rendering Advertisements with Rich Media.” I am informed that Lau qualifies as prior art because it was filed on November 7, 2006 and published on May 17, 2007, which are both before the earliest priority date of the challenged claims. Lau was cited during prosecution and mentioned by the examiner as “cited to disclose rendering advertisements with rich media” but was not relied upon in any ground of rejection. I(EX1006, p.182 (May 11, 2017 Non-Final Rejection).)

67. Like Kritt, Lau relates to audio and image processing of multimedia content, such as videos. Lau is directed to a method by which “an advertisement is matched to subject matter in a portion of rich media content, such as a digital video[.]” (Lau, p. 1 (Abstract).) “In one embodiment, engine **102** correlates advertisements to subject matter associated with rich media content. Accordingly, an advertisement that correlates to the subject matter associated with the portion of rich media content may be served such that it can be rendered on user device **104** relative to the portion of rich media content. Different methods may be used to

correlate or match advertisements to portions of the rich media content.” (Lau, ¶0020.)

68. One method of matching advertisements to media content disclosed by Lau is based on keyword searching: “For example, an advertisement may be associated with a keyword. When that keyword is used in the rich media, correlation engine **202** correlates the advertisement to a portion of rich media content in which the keyword is used.” (Lau, ¶0037.) These keywords may be derived from analysis of audio and visual elements of the media:

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, ¶0040.)

69. Lau teaches that the advertisement may be generated by identifying “ad units” based on keywords, and then combining those ad units into an advertisement for display with the media content:

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.) The matched advertisement can then “be displayed in serial, parallel, or be injected into the rich media content.” (Lau, ¶0034.)

70. I will provide more information about Lau in my discussion of the claim limitations below.

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

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

71. 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 audio files and the image files across a plurality of processors and processing the audio files and the image files in parallel;
 - [c] converting audio files associated with the video to text;
 - [d] identifying an object in the image files;
 - [e] determining a contextual topic from the image files;
 - [f] assigning a probability of accuracy to the identified object based on the contextual topic;
 - [g] converting the image files associated with the video to video data, wherein the video data comprises the object, the probability, and the contextual topic;
 - [h] cross-referencing the text and the video data with the video to determine contextual topics;
 - [i] generating a contextual text, an image, or an animation based on the

determined contextual topics;

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

(’954, 13:66-14:18 (Claim 1 (bracketed notation (e.g., “[a]”) added).)

72. The preamble of claim 1 recites “[a] **method to generate video data from a video comprising.**” Assuming the preamble provides a claim limitation, Krittr discloses it. Krittr discloses **generat[ing] video data from a video**, such as tags and other meta-data describing the video generated from performing audio and/or video processing of the video.

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

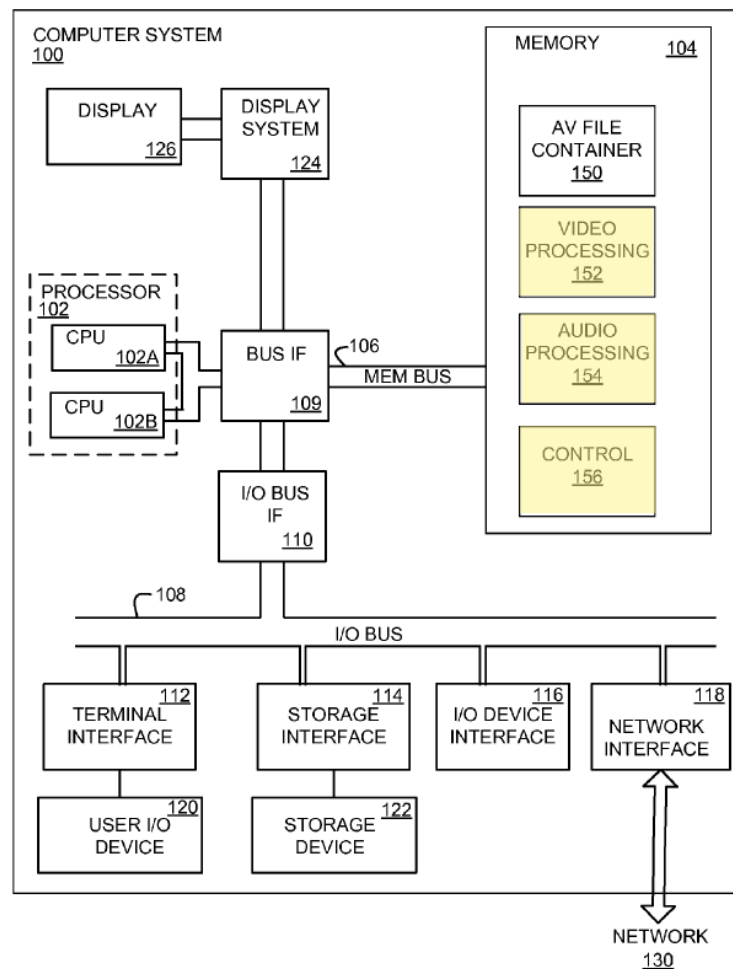


FIG. 1

(Kritt, ¶0023, Fig. 1) The video and audio processing modules are used to generate tags describing the content of the video: “The video processing module **152** may include various processes that generate visual tags according to one embodiment. The audio processing module **154** may include various processes for generating audio and key word tags according to one embodiment.” (Kritt, ¶0024.) Kritt’s “visual tags” and “audio and key word tags” are each data about a video, generated

from the video, making them “**video data from a video.**”

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

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

75. Claim 1[a] is disclosed by and obvious over Kritt in view of Fontana.

76. Kritt teaches that the video for processing may be stored in a container format, which contains the audio and video files in addition to any metadata or other files:

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

(Kritt, ¶0002.) Kritt’s Figure 2 illustrates “an exemplary audio-visual file container **202** that may contain a video file **204**, an audio file **206**, a subtitle file **208**, and a metadata file **210** according to one embodiment.”

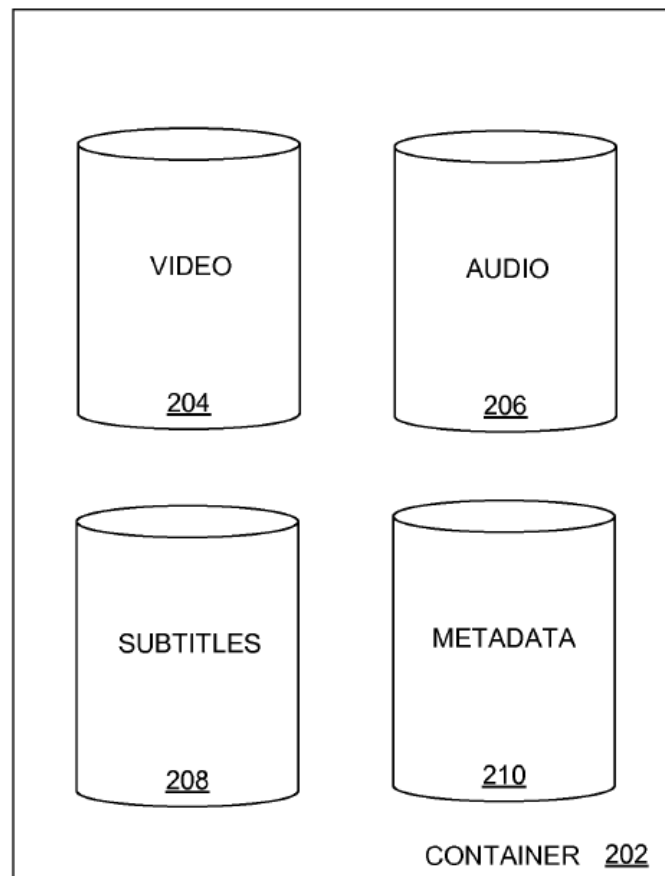


FIG. 2

(Kritt, ¶0031, Fig. 2.) 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.

77. Kritt discloses **generating image files from the video**, in the form of “shot files” and extracted video frames, each of which a POSA would have recognized as a form of image file. As to “shot files,” Kritt explains that “a video

file **204** may be parsed into shot files. As mentioned, a shot may be a continuous sequence of frames captured without interruption by a camera oriented in a single direction or camera angle.” (Kritt, ¶0039.) A POSA would recognize a frame as an image, and would further appreciate that the disclosed “shot files” are image files, in that they consist of a sequence of frames (that is, images). Generating such shot files from videos was well-known to a POSA, as Kritt acknowledges: “A video file **204** may be parsed into shot files according to any known method.” (Kritt, ¶0040.) Kritt describes, for example, a method of generating shot files from a video by computing a “histogram ... for each frame of the video file and the histograms for consecutive frames compared. If the histogram intersection of first and second consecutive frames is greater than a threshold, it may be inferred that the frames are similar, and consequently, that the two frames are part of the same shot. On the other hand, if the histogram intersection of first and second consecutive frames is less than the threshold, it may be inferred that the two frames form a shot boundary.” (Kritt, ¶0040.)

78. Kritt additionally discloses extracting video frames from the video, which as noted above a POSA would have recognized as image files. In the example above for parsing shot files, for example, Kritt analyzes “frames” to generate the shot files. (See Kritt, ¶0040.) Kritt further discloses extracting a key frame for a

shot file, which is a single image file to represent the entirety of a shot file. Specifically, “a key frame may be determined for a shot file. The key frame may be deemed to be representative of all frames in the shot, permitting descriptive data for the shot to be determined only for the key frame and not for every frame of the shot. In one embodiment, a key frame may be determined for each shot file.” (Kritt, ¶0041.)

79. Kritt also renders obvious **generating audio files from the video**. Kritt describes storing an audio file as part of the video container file, and separate modules for processing video and audio. Kritt further discloses that the audio and video processing modules can be on different computer systems.

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

(Kritt, ¶0022.) A POSA would have understood and found it obvious that, in Kritt’s system involving distributed video and audio processing, audio files and image files would be generated from the video in order to distribute them to the appropriate

modules for processing. Kritt further explains that “one or more audio features or audio signal descriptors may be extracted from an audio file[.]” (Kritt, ¶0056.) A POSA would have found it obvious that in order to extract audio features or audio signal descriptors from an audio file, the audio file must first be extracted from the container.

80. Furthermore, the combination of Kritt with Fontana renders obvious **“generating audio files and image files from the video.”**

81. Fontana discloses **“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.

82. 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 processing module 608 is configured to process the video portion(s) of multimedia content[.]” (Fontana, ¶¶0091, 0094.) Fontana further discloses that 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, like Kritt, 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**.”

83. Fontana specifically teaches that its audio processing 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. Fontana additionally teaches that “it is recognized that a number of sources provide speech to text conversion programs that approach the conversion differently.” (Fontana, ¶0169.) One such program disclosed in Fontana is “large vocabulary continuous speech recognition (LVCSR) engines” that “depend on a language model that includes a vocabulary/dictionary for

speech-to-text conversion of audio files.” (Fontana, ¶0169.) A POSA would thus have appreciated that Fontana’s disclosures encompass “**generating audio files ... from the video.**”

84. Fontana additionally teaches that the “video processing module” **generates** an additional set of “**image files**” 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, ¶0095.) A POSA would appreciate that a thumbnail is an image. Such images were commonly stored as image files in 2013, such as in gif, jpg, or tiff formats, for example, which are common formats for graphical image files. A POSA would thus have appreciated that by extracting thumbnail images from the video, Fontana’s disclosures satisfy “**generating ... image files from the video.**”

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

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

86. A POSA would have appreciated that Kritt’s method of generating video data from a video could incorporate aspects of Fontana’s audio and video

processing. In particular, as to generating image files, Kritt teaches that “a key frame may be determined for a shot file” and the “key frame may be deemed to be representative of all frames in the shot[.]” (Kritt, ¶0041.) However, Kritt discloses only one specific method of determining a key frame—“a key frame may be determined by selecting a middle frame of the shot file”—and otherwise explains that “[a]ny known method for determining a key frame may be employed” and “[o]ther known methods for determining a key frame may be employed[.]” (Kritt, ¶0041.) By contrast, Fontana discloses that a thumbnail can be generated to correspond to the location of an object of interest; this would make the thumbnail a better choice of a “key frame” to deem “representative of all frames in the shot” because it would include the relevant object of interest. A POSA would therefore have been motivated to combine Kritt with Fontana, to use Fontana’s disclosures of “generat[ing] thumbnails at possible locations the content provider would like to create an object of interest[.]” to identify the key frame based on likely locations of objects of interest and take advantage of Fontana’s benefits in object identification within thumbnails. (See Fontana, ¶0095; *see also, e.g.*, Fontana, ¶¶0139, 0141-0142, 0148 (describing identification of thumbnails corresponding to objects of interest).)

87. Likewise, Kritt teaches that “visual objects in a shot may be identified

by application of one or more known image recognition processes to the shot” and that “the object may be identified as an object of a particular type or instance using any known method[.]” (Kritt, ¶¶0043, ¶¶0045.) But Kritt provides few details on the substance of these known methods for object identification. Fontana provides multiple approaches to object identification. (See Fontana, ¶¶0139 (“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 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.”); see also, e.g., Fontana, ¶¶0084, 0087-0089, 0094-0095, 0139-0145, Figs. 9-10 (describing identification of objects of interest).) A POSA would also have been motivated to combine Kritt with Fontana, to use Fontana’s disclosures around object identification within thumbnail images in order to take advantage of Fontana’s improvements in object identification within the context of Kritt. And a POSA would have had a reasonable expectation of success in the combination in light of Kritt’s disclosure that “[o]ther known methods for determining a key frame may be employed[.]” and the well-known nature of the computer vision tools disclosed by Fontana. (See e.g., Kritt, ¶¶0041;

Fontana, ¶0139.)

88. As to audio files, Kritt discloses identifying “key words” from the audio files but does not explain how to perform speech-to-text recognition to identify these key words. (See, e.g., Kritt, ¶0016 (“A video may include spoken words.... The spoken words may include ‘key words.’”); Kritt, ¶0017 (“In addition, key words that are spoken by human objects in the video are identified.”).) Kritt further discloses creating “[a]n audio transcript [that] may include all of the words spoken in [a] video[,]” but does not disclose a specific algorithm for creating this transcript. (Kritt, ¶0065.) Instead, Kritt discloses that this transcript may be created “using any known technique.” (Kritt, ¶0065.) Compared to Kritt, Fontana provides additional details on well-known algorithms to perform speech-to-text recognition as part of audio processing. (See Fontana, ¶0169 (“it is recognized that a number of sources provide speech to text conversion programs that approach the conversion differently. ... LVCSR engines depend 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. 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, ¶0170 (listing

“Example sources of speech to text conversion programs performing LVCSR-based conversions”); Fontana, ¶0171 (listing examples of “Phonetic-based applications useable as one or more of the speech to text conversion programs”); Fontana, ¶0121 (“a transcript can be generated using a speech-to-text algorithm, such as that illustrated in FIGS. 11-12”); *see also* Fontana, Figs. 11-12, ¶¶00164-0174 (describing speech-to-text algorithms).) Therefore, a POSA would also have been motivated to combine Fontana’s disclosures on performing audio processing, including speech-to-text recognition, in order to implement Kritt. As above, a POSA would have had a reasonable expectation of success in the combination in light of the analogous nature of the references, Kritt’s instruction to use “any known technique” for speech recognition, and the well-known nature of the speech-to-text recognition algorithms disclosed in Fontana. (*See, e.g.*, Kritt, ¶0065; Fontana ¶¶0169-0171.)

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

89. Claim 1[b] is disclosed by and obvious over Kritt in view of Fontana.

90. As discussed above for claim 1[a], Kritt describes using separate modules for processing video and audio, and that the separate modules can be on different computer systems. Kritt explains:

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

(Kritt, ¶0022.) Kritt further teaches that “The video processing module **152** may include various processes that generate visual tags according to one embodiment. The audio processing module **154** may include various processes for generating audio and key word tags according to one embodiment.” (Kritt, ¶0024.) A POSA would have appreciated that each of the “video processing module” and “audio processing module” comprises a processor(s), such that Kritt’s distribution of the audio files and image files for processing by the video processing module and audio processing module “**distribut[es] the audio files and the image files across a plurality of processors.**”

91. That the audio and image files are distributed across a plurality of processors is confirmed by Kritt’s disclosure that its embodiments can be provided using cloud-computing infrastructure:

Embodiments according to this disclosure may be provided to end-users through a cloud-computing infrastructure. Cloud computing

generally refers to the provision of scalable computing resources as a service over a network. More formally, cloud computing may be defined as a computing capability that provides an abstraction between the computing resource and its underlying technical architecture (e.g., servers, storage, networks), enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. Thus, cloud computing allows a user to access virtual computing resources (e.g., storage, data, applications, and even complete virtualized computing systems) in “the cloud,” without regard for the underlying physical systems (or locations of those systems) used to provide the computing resources.

Typically, cloud-computing resources are provided to a user on a pay-per-use basis, where users are charged only for the computing resources actually used (e.g., an amount of storage space used by a user or a number of virtualized systems instantiated by the user). A user can access any of the resources that reside in the cloud at any time, and from anywhere across the Internet. In context of the present disclosure, a user may access applications or related data available in the cloud. For example, the nodes used to create a stream computing application may be virtual machines hosted by a cloud service provider. Doing so allows a user to access this information from any computing system attached to a network connected to the cloud (e.g., the Internet).

(Kritt, ¶¶0078-0079.) A POSA would have known that a cloud computing

infrastructure with “a shared pool of configurable computing resources” and “virtual machines hosted by a cloud service provider[,]” such as that taught by Kritt, distributes files for processing across **a plurality of processors**. (*See id.*)

92. Kritt further acknowledges that its disclosed processes may be executed in parallel:

The flowchart and block diagrams in the figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of code, which may include one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be **executed substantially concurrently**, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved.

(Kritt, ¶0080.) A POSA would have appreciated that executing two “blocks” of functionality concurrently is **parallel processing** of those functions. A POSA would further have recognized that Kritt’s video processing module and audio processing module—blocks 152 and 154, respectively, in Kritt’s Figure 1, which is reproduced

in my analysis for 1[pre] above—were two such blocks that could be executed in parallel, such that the audio processing and video processing occurred concurrently. A POSA would have appreciated that this results in **processing the audio files and the image files in parallel**.

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

94. Like Kritt, Fontana discloses implementing its system 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.) As with Kritt’s cloud

computing infrastructure, a POSA would have understood that the distributed computer system of Fontana is a form of “**parallel**” processing as required by the ’954 patent—by distributing tasks across a plurality of processors, each processor is able to handle a portion of the tasks, resulting in parallel processing of the set of requests. Indeed, one of the exemplary distributed computing environments taught by Fontana is “a cloud-based” environment, like that taught by Kritt. (*See* Fontana, ¶0048.) 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 and audio-processing tasks, resulting in processing multiple image files or audio files in parallel. Therefore, Fontana discloses that its distributed computing system distributes image and audio processing requests, such that it “**distribut[es] the audio files and image files across a plurality of processors and process[es] the audio files and the image files in parallel.**”

95. 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.)

96. 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 processing request to the integration framework **204** and the storage network **206**.”

(Fontana, ¶0050.) The 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.*) In other words, the

workflow server divides computationally intensive jobs, such as audio and 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.

97. Fontana further teaches that this distributed and parallel processing of audio and image files can be carried out by a plurality of servers. Fontana teaches that its system 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.” (Fontana, ¶0052.) 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. ... 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.) 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 this system of servers, which share processing jobs and workloads by, for example, separating and performing the jobs in parallel, carries out parallel processing, including for processing audio files and 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. Likewise, it would have been obvious that Fontana’s distributed computing system could **process the audio files in parallel** by likewise distributing the audio files across multiple servers.

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

98. Kritt discloses **“converting audio files associated with the video to text”** by creating an “audio transcript [that] may include all of the words spoken in [a] video” using “any known technique” for speech recognition:

In operation **612**, an audio transcript may be determined. An audio transcript may include all of the words spoken in the video. In one embodiment, an audio transcript may be provided with the video in the form of a closed caption file included in the AV file container. In another embodiment, spoken words may be determined from audio features classified as speech using any known technique. In yet another,

embodiment, spoken words may be manually determined.

(Kritt, ¶0065.) Kritt’s “audio transcript” that “may include all of the words spoken in the video” “**convert[s] audio files associated with the video to text.**”

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

99. Kritt discloses “**identifying an object in the image files.**”

100. As discussed for claim 1[a], Kritt teaches that “key frames” for a shot (i.e., image files) can be representative of the shots: “a key frame may be determined for a shot file. The key frame may be deemed to be representative of all frames in the shot, permitting descriptive data for the shot to be determined only for the key frame and not for every frame of the shot. In one embodiment, a key frame may be determined for each shot file.” (Kritt, ¶0041.) Kritt thus teaches that its analysis of shots may be performed on the key frame image frames.

101. Kritt further discloses identifying “visual objects,” such as faces or facial expressions, in the shots and frames (i.e., “**in the image files**”). For example, Kritt discloses identifying visual objects, and specifically human objects such as faces of specific individuals:

In operation **510**, visual objects in a shot may be identified and tagged.

In one embodiment, visual objects in a shot may be identified by application of one or more known image recognition processes to the shot. The operation **510** may operate on one or more key frames of the

shot. A shot may include the human and nonhuman visual objects. Both human and nonhuman visual objects may be identified in operation **510**. With respect to human visual objects, in one embodiment, a human visual object may be identified by identifying a face (“human facial object”) in a frame. The operation **510** may include determining whether or not a particular visual object is present in a shot and, if present, to identify its location in the frame. The operation **510** may include extracting an identified object for further processing. For example, an extracted human facial object may be further processed to determine the identity of the viewer or to determine a facial expression of the viewer.

(Kritt, ¶0043.) Kritt further discloses identifying the human objects (e.g., faces) with particular individuals: “a scene may include a first shot that is a long shot of a particular person and a second shot that is a medium shot of the same person” or “a sequence of four consecutive shots in which the first and third shots have a first human object and the second and fourth shots have a second human object[.]” (Kritt, ¶0051.) A human object (a face, or its associated person) is **an object in the image files**, such that Kritt discloses **identifying an object in the image files**.

102. Kritt further discloses processing the “human facial object ... to determine a facial expression of the viewer.” (Kritt, ¶0043.) Kritt alternately describes this as determining an “emotion”: “a human facial object may be processed

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

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

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

emotion for a shot or scene by evaluating facial expressions across a sequence of image files.

104. **Scenes/Shots:** Kritt discloses **determining a contextual topic** in the form of a “scene” or a “shot.” Shots and scenes are each a sequence of related images, and they are determined by evaluating a series of frames, which are **image files**.

105. A shot, as the term is used in Kritt, “may be a continuous sequence of frames captured without interruption by a camera oriented in a single direction or camera angle.” (Kritt, ¶0039.) Kritt explains that the “video file **204** may be parsed into shot files according to any known method[,]” and provides as an example determining a shot by comparing histograms for consecutive frames of the video:

The video file **204** may be parsed into shot files according to any known method. For example, in one embodiment, a histogram may be computed for each frame of the video file and the histograms for consecutive frames compared. If the histogram intersection of first and second consecutive frames is greater than a threshold, it may be inferred that the frames are similar, and consequently that the two frames are part of the same shot. On the other hand, if the histogram intersection of first and second consecutive frames is less than the threshold, it may be inferred that the two frames form a shot boundary. In addition, it may be inferred that the first consecutive frame is the last frame of a

preceding shot and the second consecutive frame is the first frame of a succeeding shot. In one alternative, the histograms of two or more consecutive first frames may be compared with the histograms of two or more consecutive second frames (the group of first and second frames being consecutive), and a shot boundary may be defined by more consecutive frames than merely two frames. For example, the shot transition between shots may be a “fade” rather than a “cut.” A time code and type of shot transition (fade or cut) may be recorded as metadata for use in content analysis described below. Other known methods for parsing a video file into shot files may be employed in operation 504. In addition, operation 504 may include parsing the video file so that sequential frames between determined shot boundaries are grouped togetherI [sic] or otherwise identified or tagged as being associated with a particular shot. Sequential frames associated with a particular shot may be referred to herein as a shot file.

(Kritt, ¶0040.) By determining which frames belong to a given shot, and tagging those frames as associated with a particular shot, Kritt determines a **contextual topic** (the shot) **from the image files** (the analyzed frames).

106. Similarly, Kritt also discloses identifying scenes, which are “semantic segments” comprised of two or more frames (and one or more shots) and which may be “intended to convey an element of a story[:]” “a segment of two or more video frames that includes common objects, that is intended to convey common emotional

content, that is intended to convey an element of a story, that is accompanied by a common audio segment, or some combination of the foregoing may be classified as a scene. A scene may also be referred to in this description and the claims as a ‘semantic segment.’” (Kritt, ¶0050.) “A visual scene may include one or more camera shots and one or more human and nonhuman objects.” (Kritt, ¶0051.) Kritt’s determination of a scene is a **contextual topic**.

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

108. **Emotions**: Additionally, Kritt discloses **determining a contextual topic** in the form of an emotion, mood, or theme for a shot or scene that is determined

by evaluating facial expressions across a sequence of image files: “the objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.” (Kritt, ¶0017.) An “emotion, mood, or theme” is a “**contextual topic**” because it describes the context (in this case, the general feeling) associated with the shot or scene.

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

110. This is analogous to Kritt’s determination of an emotion. Kritt teaches that it identifies “an emotional category of a human facial object” by analyzing the actions performed by the face—or as Kritt explains, “by comparing motion vectors” for the facial expression across “a sequence of frames” (i.e., image files):

In one embodiment, a human facial object may be processed to determine the emotion expressed on the facial object. To determine the emotion of a facial object, a process may, in one embodiment, employ a Gabor filter to determine facial features and their orientation, and a support vector machine to determine an emotion corresponding with detected facial features. In one embodiment, a sequence of frames in

which a facial expression morphs from one emotion to another may be analyzed to determine an emotional category of a human facial object. The sequence of frames need not include every consecutive frame, e.g., two or more key frames may be analyzed. The sequence of frames may be analyzed using a Tree-Augmented-Naive Bayes classifier. In addition, a category of emotion may be determined by comparing motion vectors with a template. The motion vectors may be based on deformation of facial features as reflected in an optical flow that occurs in a sequence of frames. Optical flow may be determined using differential, matching, energy-, or phase-based techniques. In various embodiments, motions that may be determined may include amusement, joy, anger, disgust, embarrassment, fear, sadness, surprise, and a neutral state. Other emotions or moods may be determined in alternative embodiments. The operation **510** may include associating a determined emotion with a human object. In addition, the operation **510** may include generating an emotion tag that is associated with the scene of the video in which the facial emotion was detected. In other embodiments, the emotion of a facial object may be determined in operation 510 using any known method

(Kritt, ¶0047.) Kritt thus teaches identifying the actions (motions) performed by a face from a series of frames (image files) to determine an emotional category or mood (e.g., “amusement, joy, anger, disgust, embarrassment, fear, sadness, surprise”), from which an emotion tag is generated. This corresponds to the ’954

patent's **determining a contextual topic from the image files.**

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

111. Kritt discloses and renders obvious **assigning a probability of accuracy to the identified object based on the contextual topic.** As discussed for claim 1[d], the identified objects include human objects (faces and/or people) as well as facial features and facial expressions. Kritt **“assign[s] a probability of accuracy”** to these **“identified object[s].”** That probability of accuracy, as explained below, is **“based on [a] contextual topic,”** which, as explained for claim 1[e], include shots, scenes, and emotions.

112. **Human objects (faces/people):** Kritt discloses this limitation by assigning a probability that a human object (**“identified object”**) was correctly identified by using other information about the shot or scene (**“the contextual topic”**) in which the human object was identified. The probability is therefore assigned **based on the contextual topic**, as detailed below.

113. Specifically, Kritt discloses “a visual tag” associated with a shot or scene to “indicate that a particular human object appears in a shot[.]” (Kritt, ¶0033.) Kritt teaches that this “visual tag” can be compared with other tags associated with the shot or scene “to validate that the human object was correctly identified.” (Kritt,

¶0033.) Kritt discloses that this validation process may involve assessing “if a **probability or confidence parameter** associated with the particular tag is above a threshold,” in which case “it may be determined that the object was correctly identified[.]” (Kritt, ¶0033.) As Kritt explains:

The operation **310** may include comparing a tag with one or more other tags associated with the same shot or scene for consistency. A shot may be a continuous sequence of frames captured without interruption by a camera oriented in a single direction or camera angle. As one example, a visual tag may indicate that a particular human object appears in a shot and a key word tag identifying the name of the human object is associated with the shot. As another example, a visual tag may indicate that a particular human object appears in a shot and an audio tag identifies an audio signature of the human object is associated with the shot. In these examples, if the tags that are compared indicate the same object, the positive or consistent result of the comparison may be used in operation 310 to validate that the human object was correctly identified. If there are no tags that are consistent with a particular tag, it may be determined that the object associated with the particular tag was misidentified. The operation **310** may include modifying a tag determined to be inconsistent with other tags associated with the same shot. The modification may include adding an indication to the tag that it should not be used in other processes. Alternatively, if a probability or confidence parameter associated with the particular tag is above a

threshold, it may be determined that the object was correctly identified and that the shot or scene includes multiple objects. In this circumstance, the modification may include adding an indication to the tag that it may be relied on to a particular extent.

(Kritt, ¶0033.) A POSA would have appreciated that assessing “if a probability or confidence parameter associated with the particular tag is above a threshold” requires **assigning a probability of accuracy to the identified object**. In the context of Kritt, this probability is assessed based on the above-described comparison of the visual tag for the human object to other tags associated with the same shot or scene—which are **contextual topics**. Thus, the probability of accuracy is assigned **based on the contextual topic**.

114. **Facial expressions**: Kritt renders obvious this limitation for the facial expression objects by describing a “confidence level” that a perceived emotion (a **contextual topic**) “accurately represents a ground truth emotion.” (Kritt, ¶0034.)

115. As described in claim 1[e], Kritt uses the identified facial expressions (*see* claim 1[d]) to detect emotions and generate an emotion tag associated with a scene:

In one embodiment, a human facial object may be processed to determine the emotion expressed on the facial object The operation **510** may include associating a determined emotion with a human

object. In addition, the operation 510 may include generating an emotion tag that is associated with the scene of the video in which the facial emotion was detected.

(Kritt, ¶0047.)

116. Kritt then discloses assigning to that emotion “a confidence level that the perceived emotion accurately represents a ground truth emotion:”

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

(Kritt, ¶0034.) A POSA would have appreciated that a confidence level corresponds to a **probability of accuracy**. A POSA further would have found it obvious that, because Kritt identifies the emotion based on the facial expression, a probability of accuracy could be assigned to the underlying facial expression based on the probability of accuracy (ie, confidence level) assigned to the emotion (the

contextual topic).

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

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

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

119. I discuss above Kritt’s identification of an **object** (claim 1[d]), determination of a **contextual topic** (claim 1[e]), and assignment of a **probability** (claim 1[f]). Each of these is generated from analysis of the **image files** (i.e., frames and shots), as I previously explained. Below I explain how for each of these, Kritt converts the object, the probability, and the contextual topic into **video data** in the

form of tags, tag files, or other metadata.

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

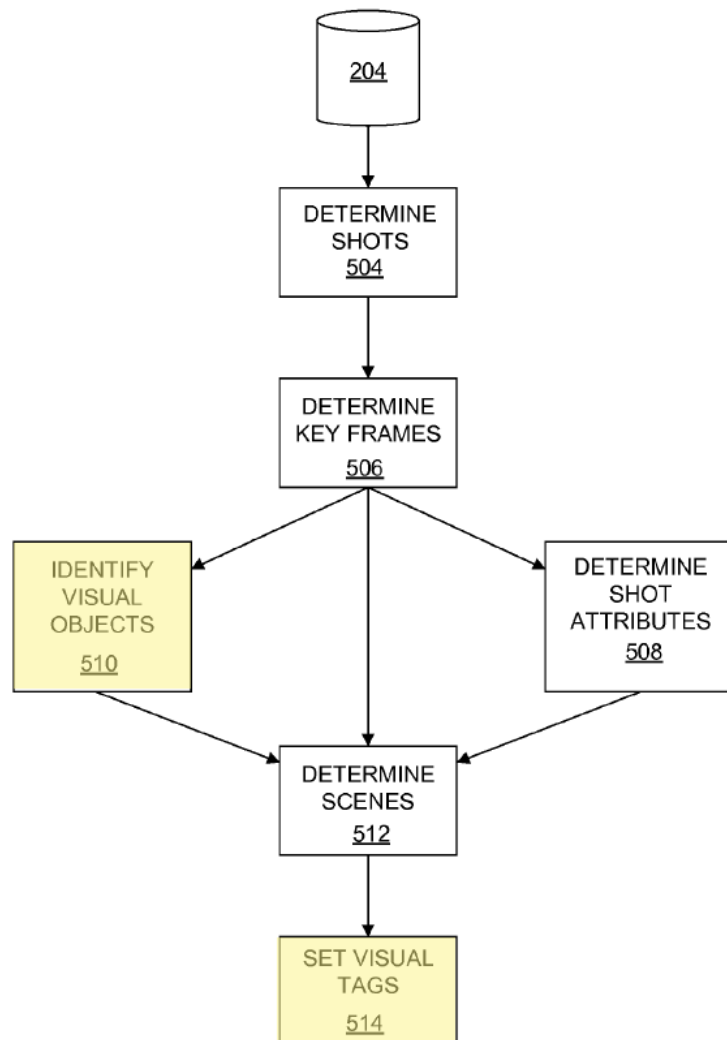


FIG. 5

(Kritt, Fig. 5 (annotated).) A POSA would further appreciate that these visual tags could include facial expressions. Kritt explains that a tag may “identif[y] a human object displaying a particular emotion.” (Kritt, ¶0034.) A POSA would have

appreciated that a tag for “a human object displaying a particular emotion” in the context of Kritt is a tag for a facial expression, because Kritt teaches determining emotions by analyzing facial expressions. This would be a particularly obvious implementation in light of Kritt’s disclosure of analyzing changes in facial features/facial expression across a series of frames:

In one embodiment, a human facial object may be processed to determine the emotion expressed on the facial object. To determine the emotion of a facial object, a process may, in one embodiment, employ a Gabor filter to determine facial features and their orientation, and a support vector machine to determine an emotion corresponding with detected facial features. In one embodiment, a sequence of frames in which a facial expression morphs from one emotion to another may be analyzed to determine an emotional category of a human facial object. The sequence of frames need not include every consecutive frame, e.g., two or more key frames may be analyzed. The sequence of frames may be analyzed using a Tree-Augmented-Naive Bayes classifier. In addition, a category of emotion may be determined by comparing motion vectors with a template. The motion vectors may be based on deformation of facial features as reflected in an optical flow that occurs in a sequence of frames. Optical flow may be determined using differential, matching, energy-, or phase-based techniques. In various embodiments, motions that may be determined may include amusement, joy, anger, disgust, embarrassment, fear, sadness, surprise,

and a neutral state. Other emotions or moods may be determined in alternative embodiments.

(Kritt, ¶0047.)

121. As for the **contextual topic**, both the contextual topics of shots/scenes and emotions are stored as **video data** in the form of tags. Kritt teaches that determining a shot in “operation **504** may include parsing the video file so that sequential frames between determined shot boundaries are grouped togetherI [sic] or otherwise identified or tagged as being associated with a particular shot.” (Kritt, ¶0040.) This is illustrated in Figure 5:

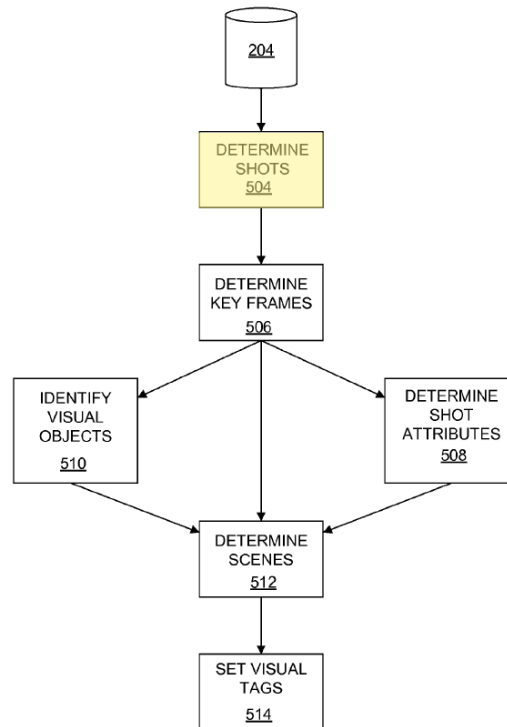


FIG. 5

(Kritt, Fig. 5 (annotated).) A POSA would have found it obvious that the frames or shots belonging to a determined scene could likewise be tagged, in operation **512** of the same Figure **5**. This is consistent with Kritt’s use of tags to identify a frame as belonging to a particular shot (Kritt, ¶0040) and is further consistent with Kritt’s use of tags in connection with scenes: “One or more of the various tags described herein may be associated with a particular scene or semantic segment if the particular tag is determined from content in the scene.” (Kritt, ¶0050.)

122. Kritt additionally teaches generating tags for the identified emotion,

which is another contextual topic. Kritt discloses an “emotion tag file **314**” that “includes tags that are associated with emotion objects.” (Kritt, ¶0034.) In particular, “a tag may be generated for an emotion object.” (Kritt, ¶0034.) Kritt explains:

In operation **312**, an emotion tag file **314** may be created from the attribute tag file **308** and the consistency-corrected visual tag **302**, audio tag **304**, key word tag **306**, and metadata **210** files. The emotion tag file **314** includes tags that are associated with emotion objects. In one embodiment, an emotion object may be associated with an emotion, mood, or theme that a typical viewer might be expected to perceive or that the creators of a video intended the audience to perceive. Each emotion object may be of a predefined type and associated with a time stamp. An emotion object may include parameters corresponding with intensity of the perceived emotion or a confidence level that the perceived emotion accurately represents a ground truth emotion. ... In one embodiment, a tag may be generated for an emotion object. An emotion tag may include an intensity level of the emotion, mood, or theme.

(Kritt, ¶0034.)

123. Kritt likewise discloses including the **probability** as part of the video data of tags. As discussed in claim 1 [f], Kritt discloses or renders obvious assigning a probability of accuracy to the human objects (e.g., faces) and facial expressions

(determined via emotions). Kritt teaches use of “a probability or confidence parameter associated with the particular tag” for visual objects, including human objects (like faces or people), to determine whether the object was correctly identified:

The operation 310 may include comparing a tag with one or more other tags associated with the same shot or scene for consistency. ... As one example, a visual tag may indicate that a particular human object appears in a shot and a key word tag identifying the name of the human object is associated with the shot. As another example, a visual tag may indicate that a particular human object appears in a shot and an audio tag identifies an audio signature of the human object is associated with the shot. In these examples, if the tags that are compared indicate the same object, the positive or consistent result of the comparison may be used in operation 310 to validate that the human object was correctly identified. If there are no tags that are consistent with a particular tag, it may be determined that the object associated with the particular tag was misidentified. The operation 310 may include modifying a tag determined to be inconsistent with other tags associated with the same shot. The modification may include adding an indication to the tag that it should not be used in other processes. Alternatively, if a probability or confidence parameter associated with the particular tag is above a threshold, it may be determined that the object was correctly identified and that the shot or scene includes multiple objects. In this

circumstance, the modification may include adding an indication to the tag that it may be relied on to a particular extent.

(Kritt, ¶0033.) As the above quotation shows, Kritt further discloses “adding an indication to the tag that it may be relied on to a particular extent.” (Kritt, ¶0033.) A POSA would have appreciated that the confidence level (**probability**) associated with a human object is one such indication of the extent to which an object identification can be relied upon, and that the probability could accordingly be added to the tag.

124. As to facial expressions and emotions, Kritt teaches that the emotion tag file (**video data**) can include “a confidence level that the perceived emotion accurately represents a ground truth emotion”:

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

(Kritt, ¶0034.) As discussed for claim 1[f], it would have been obvious to a POSA that the confidence level (that is, **probability**) assigned to an emotion could likewise be assigned to the facial expression/facial features used to determine that emotion. (*See supra.*)

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

125. Kritt discloses claim 1[h].

126. As discussed above for claim 1[e], Kritt **determines contextual topics** in the form of emotions/moods and shots/scenes. Claim 1[e] explains how this determination is based on information derived from the image files, which Kritt stores as visual tags (**video data**, *see* claim 1[g]). Kritt further discloses that the contextual topics are determined by **cross-referencing the text** in the form of the audio transcript and resulting key words **and the video data with the video**.

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

objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.”

(Kritt, ¶0017.) The objects include visual objects, which are **video data** (*see* claim 1[g]): “visual objects in a shot may be identified and tagged...A shot may include the human and nonhuman visual objects.” (Kritt, ¶0043.) The visual objects can be represented by a visual tag: “The visual tag file **302** includes tags that correspond with visually perceivable objects, such as human and nonhuman objects.” (Kritt ¶0032.) The key words are derived from the **text**: “key words may be determined from the audio transcript...a key word may be determined from the audio transcript and counting the frequency of occurrences of words, the most frequently occurring verbs and nouns being determined to be key words” (Kritt, ¶0066; Kritt ¶0032 (“The key word tag file **306** includes tags that correspond with key word objects.”).) And, as discussed further below, the other data includes information such as metadata files, time stamps, and scene information that is derived from **the video**. For example, Kritt teaches that “[e]ach tag may be associated with a time stamp that indicates the start and stop time in which the object or attribute is rendered or otherwise associated.” (Kritt, ¶0032.) As noted below, these time stamps are used as part of determining an emotion.

128. Kritt teaches how each of the visual objects (**video data**), key words

from the text (**text**), and other information from the **video** (e.g., metadata files, time stamps) are **cross-referenced** to determine an emotion:

In operation **312**, an emotion tag file **314** may be created from the attribute tag file **308** and the consistency-corrected visual tag **302**, audio tag **304**, key word tag **306**, and metadata **210** files. The emotion tag file **314** includes tags that are associated with emotion objects. In one embodiment, an emotion object may be associated with an emotion, mood, or theme that a typical viewer might be expected to perceive or that the creators of a video intended the audience to perceive. Each emotion object may be of a predefined type and **associated with a time stamp**. An emotion object may include parameters corresponding with intensity of the perceived emotion or a confidence level that the perceived emotion accurately represents a ground truth emotion. An emotion object may be generated directly from the attribute file **308**, such as where the attribute file identifies an association or correlation of an attribute with a perceived emotion. In addition, an emotion object may be generated directly from the visual tag **302**, such as where the tag identifies a human object displaying a particular emotion. Further, an emotion object may be generated directly from the audio tag **304** or key word tag **306** files, such as where an audio tag identifies a segment of sound associated or correlated with an emotion, mood, or theme, or a key word is associated with an emotion, mood, or theme. Moreover, **an emotion object may be generated in operation 312 by identifying patterns of visual, audio, key word, and attribute tags that**

correspond or correlate with an emotion object.

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

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

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

(Kritt, ¶0017.)

130. Kritt therefore discloses **cross-referencing the text** (key words) **and**

the video data (visual tags for, e.g., faces and other human objects) **with the video** (metadata, time stamps, and scenes) **to determine contextual topics** (emotions).

131. **Scenes**: Kritt also **determines a contextual topic** in the form of a shot or scene, as discussed for claim 1[e]. As with emotions, Kritt discloses that the scenes are determined by **cross-referencing the text and the video data with the video**. As with determinations of emotion, the determination of scenes derives from cross-referencing multiple types of information about the video. Kritt explains:

According to an aspect, a segment of two or more video frames that includes common objects, that is intended to convey common emotional content, that is intended to convey an element of a story, that is accompanied by a common audio segment, or some combination of the foregoing may be classified as a scene. A scene may also be referred to in this description and the claims as a “semantic segment.” ...

In operation **512**, a visual scene may be determined according to any known method. A visual scene may include one or more camera shots and one or more human and nonhuman objects. In one embodiment, scenes may be determined by grouping together consecutive shots having visual or audio objects corresponding with the same ground truth.

(Kritt, ¶¶0050-0051.) As before, the objects include visual objects, which are **video data** (*see* claim 1[g]): “visual objects in a shot may be identified and tagged...A shot may include the human and nonhuman visual objects.” (Kritt, ¶0043.) A POSA

would have appreciated that the “common audio segment” and “audio objects” include **the text** (*see* claim 1[c]), such as spoken words identified in an audio transcript. (*See* Kritt, ¶0016 (“A video may include spoken words, music, and other sounds, which may be referred to herein as audio objects.”); Kritt, ¶0017 (“Human, nonhuman, and audio objects may be used to classify a particular segment of a video as a scene. The objects and key words are then associated with the scenes of the video.”).) Finally, the **video** frames “intended to convey common emotional content” and “intended to convey an element of a story” are from the **video**, as are the determined emotions. (*See* Kritt, ¶0050; *see also* claim 1[e] (describing determining an emotion).)

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

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

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

134. **Emotions:** Kritt teaches **generating a contextual text, an image, or**

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

135. Additionally, the emotion denoting icon is **an image, or an animation**. Kritt explains the emotion denoting icons as follows:

In various embodiments, as shown in FIG. 4, emotion, mood, or theme denoting icons **416** may be rendered in the second window **406**. An emotion denoting icon **416** may be associated with and representative of an emotion tag. An emotion denoting icon **416** may be rendered at horizontal locations corresponding with the temporal location of the particular emotion tag in the video. In one embodiment, an emotion or mood denoting icon **416** may be an “emoticon.” In other embodiments, an emotion or mood denoting icon **416** may be a colored or gray-scale icon. While depicted as circular, an icon **416** may be any shape. In various embodiments, the size, color, or shade of an icon **416** may correspond with an intensity of the associated emotion tag. For example, an icon **416** associated with amusement or a funny mood may be relatively large if the mood or emotion would be expected to be perceived intensely, but the same icon may be relatively small if the

mood or emotion would be expected to be perceived mildly.

(Kritt, ¶0038.) Exemplary emotion denoting icons are shown in Figure 4:

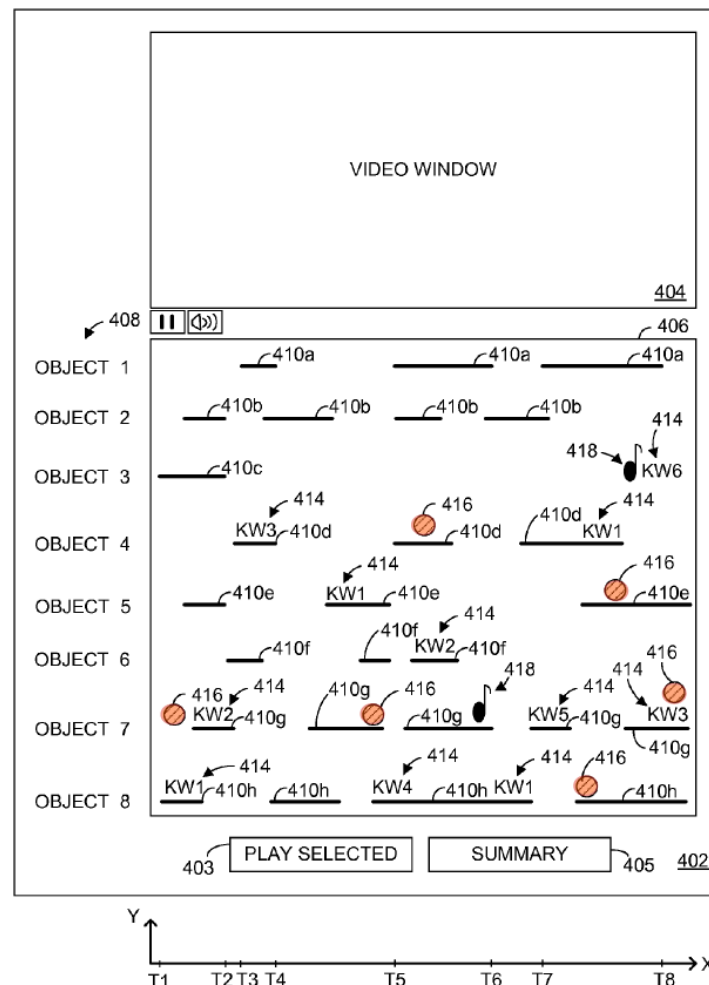


FIG. 4

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

136. Kritt additionally teaches **generating a contextual text, an image, or an animation based on the determined contextual topics** of scenes in the form of

“visual information summarizing all of the scenes of a video[.]” (Kritt, ¶0016.)

These visual summaries “describe the scenes of a video to a viewer.” (Kritt, ¶0016.)

Kritt explains:

In particular, according to various embodiments, visual representations of key words, and human, nonhuman, audio, and emotion objects may be used to describe the scenes of a video to a viewer. In addition, visual representations of the relationships between these objects and key words may be used to describe the scenes of a video to a viewer. By visually presenting this information to the viewer, he or she may be enabled to generally understand the scene. The information may enable the viewer to determine whether a particular scene is of interest or is objectionable. In various embodiments, visual information summarizing all of the scenes of a video may be presented to the viewer in a single display screen.

(Kritt, ¶0016; *see also* Kritt, ¶0017.)

137. The visual information is **contextual**, because it “describe[s] the scenes of a video[.]” (Kritt, ¶0016.) Furthermore, “visual information” includes **text, an image, or an animation**. (*See* Kritt, ¶0035 (“A variety of text, symbols, lines, and icons (‘indicia’) for summarizing the video may be rendered in the window[.]”).) And the visual information summary is **based on the determined contextual topics** because it describes a scene, which is a **contextual topic**. (*See* claim 1[e].)

Additionally, as described previously, one of the ways of summarizing the scenes is through the use of emotion denoting icons, which are additionally based on the contextual topic of emotion. (See, e.g., Kritt, ¶0038 (“It will be appreciated that the display of emotion denoting icons 416 facilitates a quick understanding by a viewer of the types and locations of various emotions, moods, or themes in the video. A viewer can determine in a single view the proportion of the video that is associated with a particular emotion, mood, or theme, such as action or comedy. In addition, a viewer can determine in a single view where emotion objects of a particular type are located, e.g., funny portions of the video.”).)

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

138. Kritt discloses **generating a content-rich video based on the generated text, image, or animation.**

139. Kritt teaches **generating a content-rich video** by displaying the video alongside images and text generated to summarize the scenes, including the emotion denoting icons. Kritt explains:

FIG. 4 depicts a display screen 402 of a display device, e.g., display 126 (FIG. 1), for displaying a user interface. In one embodiment, the user interface includes windows 404 and 406, which may be rendered on the display screen 402 along with a variety of textual information,

and control icons or buttons, e.g., buttons **403**, **405**, outside of the windows. The video may be played in the window **404**. A variety of text, symbols, lines, and icons (“indicia”) for summarizing the video may be rendered in the window **406**. The horizontal extent of the window **406** may correspond with the duration or total time of the video. The x axis shown in the figure represents the horizontal extent or time, while the y axis represents a vertical direction. While FIG. **4** depicts a user interface that includes a window **404** for playing a video, in other embodiments, a user interface may omit the window **404**, i.e., in other embodiments, a user interface may include only the window **406** for rendering text, symbols, and icons for summarizing the video (along with control icons or buttons **403**, **405** outside of the window **406**).

...

In various embodiments, as shown in FIG. **4**, emotion, mood, or theme denoting icons **416** may be rendered in the second window **406**. An emotion denoting icon **416** may be associated with and representative of an emotion tag. An emotion denoting icon **416** may be rendered at horizontal locations corresponding with the temporal location of the particular emotion tag in the video. In one embodiment, an emotion or mood denoting icon **416** may be an “emoticon.” In other embodiments, an emotion or mood denoting icon **416** may be a colored or gray-scale icon. While depicted as circular, an icon **416** may be any shape. In various embodiments, the size, color, or shade of an icon **416** may correspond with an intensity of the associated emotion tag. For

example, an icon **416** associated with amusement or a funny mood may be relatively large if the mood or emotion would be expected to be perceived intensely, but the same icon may be relatively small if the mood or emotion would be expected to be perceived mildly.

(Kritt, ¶¶0035, 0038.) This display of the generated **content-rich video** is depicted in Figure 4:

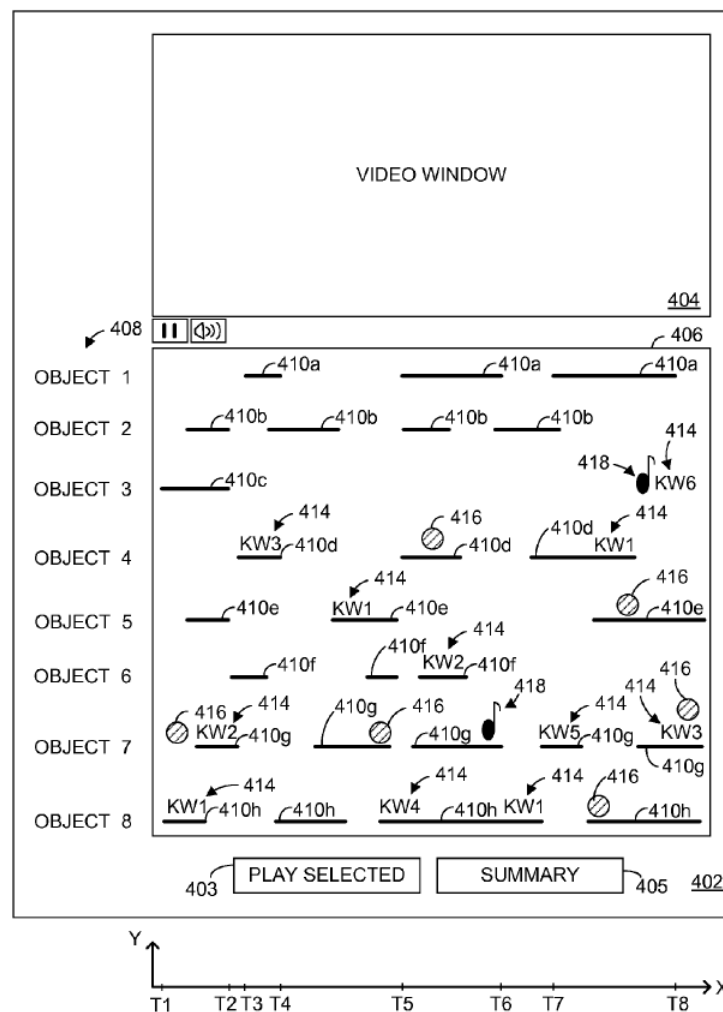


FIG. 4

(Kritt, Fig. 4.) This two-window display of the content-rich video depicts the **contextual text, image, or animation** generated in claim 1[i]—namely, the visual information summary, and the emotion denoting icons. (*See supra.*) This aligns with the '954 patent specification. The '954 patent gives as one example of content-rich video that an advertisement may be generated based on the video content, and the advertisement “can be placed at a specific time in the video” and “can also be placed at a specific time as part of the video player, e.g., side panel, and also may be placed on a second screen.” ('954, 2:5-10.) So too here, the content-rich video is generated by placing the generated contextual text, image, or animation in a side panel at a specific time in the video player.

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

140. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. The combination of Kritt with Fontana renders obvious **applying natural language processing to the text to determine context associated with the video.**

141. Kritt describes determining key words from an audio transcript (“**the text**”), including key words that provide “significant information content about a scene in a video” (“**context**”), such as “a concept or idea that is central to a plot or

story”:

In operation **614**, key words may be determined from the audio transcript. A key word may be a word that provides significant information content about a scene in a video. For example, a key word may be a name of an actor that appears in a scene. A key word may be a name of a concept or idea that is central to a plot or story. For example, the word “run” may be a key word for the movie Forrest Gump. A key word may be a name of a song. A key word may be a word that is predefined to be objectionable or liked by a viewer. For example, a vulgar word may be predefined as a key word. In one embodiment, a key word may be determined from the audio transcript by counting the frequency of occurrences of words, the most frequently occurring verbs and nouns being determined to be key words. The operation **614** may include generating key word objects for each determined key word.

(Kritt, ¶0066.)

142. As described in the above passage, Kritt discloses determining key words “by counting the frequency of occurrences of words, the most frequently occurring verbs and nouns being determined to be key words.” (Kritt, ¶0066.) A POSA would have recognized this as a rudimentary form of **natural language processing**. Such natural language processing techniques were well-known by the priority date of the ’954 patent, and a POSA would have found it obvious to build on Kritt’s disclosure by implementing one, such as term frequency/inverse document

frequency (TF-IDF), which is a form of natural language processing that identifies salient words by multiplying term frequency by inverse document frequency. (*See, e.g.,* EX1009, U.S. Pat. No. 8,706,664, 3:53–4:49 (describing TF-IDF); EX1008, U.S. Pat. No. 9,245,254, 12:4-15 (listing term frequency-inverse document frequency among known techniques for natural language processing).) Because a “key word may be a name of an actor that appears in a scene[,]” the key words determined by Kritt are a **context associated with the video**. (*See* Kritt, ¶0066.)

143. Furthermore, as discussed for claim 1[a], it would have been obvious to combine Kritt with the audio processing methods described in Fontana. One such audio processing technique discussed in Fontana is **natural language processing**: “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 id.* ¶0100 (“In alternative embodiments, additional search arrangements can be included as well, such as a natural language search...”).) A POSA combining Kritt with the audio processing techniques of Fontana thus would have found it obvious to additionally implement natural language processing, such as to identify key words.

144. For example, Fontana 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 in the multimedia content, analogous to the key words described by Krittr. Fontana explains, for 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, “[a] 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 is **determin[ing] context associated with the video**, and could be combined with Krittr to determine key words using **natural language processing**.

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

145. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. Additionally, for the reasons I explain for claim 2, the combination of Kritt with Fontana renders obvious **applying natural language processing to the text to extract a topic**. A POSA would have appreciated that Kritt’s key words, and Fontana’s keywords, are **topics**.

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

146. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. Kritt additionally discloses **processing the image files to extract additional text** by identifying “nonhuman objects” such as text appearing in the video: “Yet another type of nonhuman object that may be shown in a video frame is text. For instance, words spoken by humans may be displayed as text in a particular area of the frames.” (Kritt, ¶0004.) Kritt teaches identifying these nonhuman objects, which include the **additional text** as noted above:

According to various embodiments, a viewer selects a video, and human, nonhuman, and audio objects of the video are identified. In

addition, key words that are spoken by human objects in the video are identified. Human, nonhuman, and audio objects may be used to classify a particular segment of a video as a scene. The objects and key words are then associated with the scenes of the video. Further, the objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.

(Kritt, ¶0017.)

147. Kritt further teaches **extracting** the identified nonhuman objects, such as text: “Both human and nonhuman visual objects may be identified in operation **510**. ... The operation **510** may include extracting an identified object for further processing.” (Kritt, ¶0043.)

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

148. A POSA would have found claim 4 obvious over the combination of Kritt with Fontana, as I discuss above. In particular, as discussed for claim **1[b]**, it would have been obvious to combine Kritt with Fontana to **process the image files in parallel**. Furthermore, Kritt discloses that **the additional text** of claim 4 is **generated by segmenting the image files before processing**.

149. As discussed for claim 4, Kritt discloses that the additional text may be

identified and extracted (**generated**) by “operation 510.” (Kritt, ¶0043.) Operation 510 occurs within shots: “In operation **510**, visual objects in a shot may be identified and tagged.” (Kritt, ¶0043.) Kritt illustrates operation 510’s placement in the logical flow in Figure 5, showing that operation 510 (“identify visual objects”) takes place after “determine shots” (operation 504):

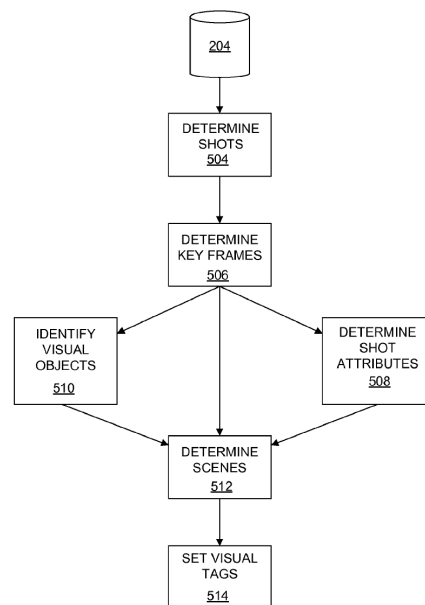


FIG. 5

(Kritt, Fig. 5.)

150. As I explained in claim 1[a], a shot is a series of consecutive image files, such that the process of generating the shot files **segments the image files** for the video. Kritt explains this process:

The video file 204 may be parsed into shot files according to any known

method. For example, in one embodiment, a histogram may be computed for each frame of the video file and the histograms for consecutive frames compared. If the histogram intersection of first and second consecutive frames is greater than a threshold, it may be inferred that the frames are similar, and consequently that the two frames are part of the same shot. On the other hand, if the histogram intersection of first and second consecutive frames is less than the threshold, it may be inferred that the two frames form a shot boundary.

(Kritt, ¶0040.) This segmentation to determine shots takes place **before processing the image files**, as shown in Figure 5 above. And as discussed for claim 1, a POSA would have found it obvious over the combination of Kritt with Fontana to process the image files **in parallel**.

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

152. As a general matter, as described for claim 1[b], Fontana teaches a distributed computing system for video, audio, and image processing that segments all tasks into discrete portions that can be run in parallel: “As such, 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.) As described for claim 1[b], Fontana’s 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 processing request to the integration framework **204** and the storage network **206**.” (Fontana, ¶0050.) The 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.*) A POSA would have appreciated that this use of a distributed computing system **segments** tasks (including image processing tasks), prior to processing the images in parallel.

153. Furthermore, Fontana teaches segmenting image files 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.) This “series of thumbnails representing scenes” **segments the image files** into representative thumbnail images

for each scene. (*Id.*) 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.

154. Fontana also discloses 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. 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, 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, ¶0058.) Because Fontana implements image processing in a distributed computing system (*see* claim 1**[b]**), each of these algorithms results in the segmented image files being processed in

parallel.

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

155. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. Additionally, Kritt discloses **processing the image files to extract an object and determine a motion associated with the object.**

156. As discuss for claim 1[d], Kritt discloses identifying objects including human objects, such as faces and facial features. (*See supra* § V.B.1(d).) Kritt further discloses **extracting** these human objects: “Both human and nonhuman visual objects may be identified in operation 510. ... The operation 510 may include extracting an identified object for further processing.” (Kritt, ¶0043.)

157. Kritt additionally teaches **determining a motion associated with the object**. Specifically, Kritt detects facial feature motion in order to identify emotions:

In one embodiment, a sequence of frames in which a facial expression morphs from one emotion to another may be analyzed to determine an emotional category of a human facial object. The sequence of frames need not include every consecutive frame, e.g., two or more key frames may be analyzed. The sequence of frames may be analyzed using a Tree-Augmented-Naive Bayes classifier. In addition, a category of

emotion may be determined by comparing motion vectors with a template. The motion vectors may be based on deformation of facial features as reflected in an optical flow that occurs in a sequence of frames. Optical flow may be determined using differential, matching, energy-, or phase-based techniques.

(Kritt, ¶0047.)

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

158. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. Additionally, the combination of Kritt with Fontana renders obvious that **the text** (the audio transcript, *see* claim 1[c]) is **generated by segmenting the audio before processing the audio files in parallel.**

159. Kritt discloses **segmenting the audio before processing**. For example, Kritt describes segmenting audio by classifying whether a segment is silence, speech (spoken words), music or sound effect. If a segment is classified as “speech” then audio transcription is performed and key words are determined (i.e., “**the text is generated by segmenting the audio**”), including in parallel as described for claim 1[b]. Kritt explains segmenting and classifying the audio as follows:

FIG. 6 illustrates a process for generating audio and key word tags according to one embodiment. Referring to FIG. 6 in operation **602**, one

or more audio features or audio signal descriptors may be extracted from an audio file 206. An audio feature may be a time domain feature, such as zero crossing rate, energy contour, volume contour, or fundamental frequency, or a frequency domain feature, such as short term energy, bandwidth, entropy, spectral centroid, Mel-Frequency Cepstral Coefficients, or a Discrete Wavelet Transform. Many audio features are known in the art and any known audio feature or features that are suitable may be extracted in operation 602.

In operation 604, audio features or audio signal descriptors extracted from an audio file 206 may be classified. Each classification may be defined by a set of characteristic audio feature values. In one embodiment, audio features may be classified as silence, speech (spoken words), music, and a fourth category of other sounds that will be referred to herein as “sound effect.”

Segments of the video for which sound is not detectable may be classified as silent. In operation 605, an audio tag with a silent type attribute may be associated with a silent audio feature, the tag having a time stamp that indicates the start and stop time of the silent period.

Segments of the video for which the audio feature values are similar to those that are characteristic of speech may be classified as speech. An audio tag with a speech type attribute may be associated with the audio feature, the tag having a time stamp of the period of speech. Segments

of the video for which the audio feature values are similar to those that are characteristic of music may be classified as music. An audio tag with music type attribute may be associated with the audio feature, the tag having a time stamp of the period of music.

Segments of the video for which the audio feature values are not similar to those that are characteristic of speech or music (and are not silent) may be classified as a sound effect. An audio tag with sound effect type attribute may be associated with a time stamp of the period of music. The sound effect category may include sounds conventionally understood to be movie or television sound effects, such as an explosion, a door being slammed, a motor vehicle engine, a scream, laughter, applause, wind, and rain. The sound effect category may include any sound that may not be classified as speech, music, or silence, even if the sound may not be conventionally understood to be a theatrical sound effect.

(Kritt, ¶¶0056-0060.) After the audio segments as classified, the segments corresponding to speech are processed to generate the audio transcript (**the text**): “In operation 612, an audio transcript may be determined. An audio transcript may include all of the words spoken in the video. ... In another embodiment, spoken words may be determined from audio features classified as speech using any known technique.” (Kritt, ¶0065.) This is illustrated in Figure 6, **segmenting the audio** and classifying the segments occurs in block 604 **before processing the audio files**

through the audio transcription process in block 612:

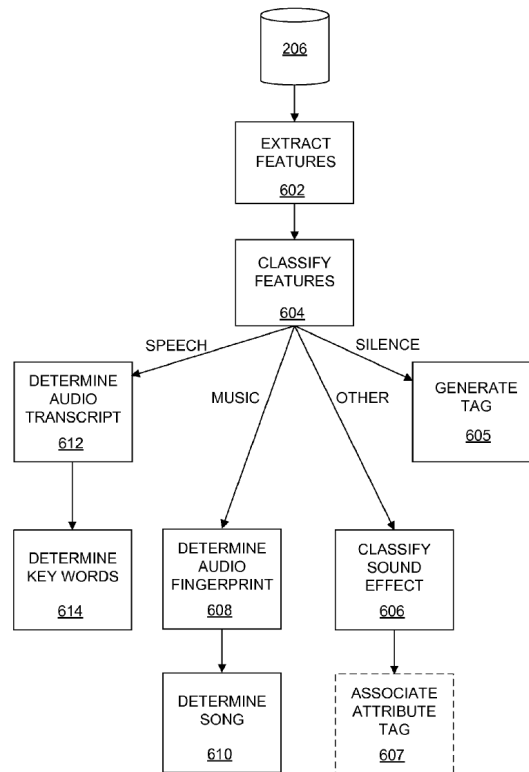


FIG. 6

(Kritt, Fig. 6.)

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

161. Additionally, a POSA would have found it obvious in light of Fontana's parallel processing disclosures to **"segment[] the audio before processing the audio files in parallel"** in order to take advantage of Fontana's teaching that

multiple servers can be used to efficiently carry out tasks like audio processing.

162. 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:

In certain embodiments, the servers **302 a-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 **302 a-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 **302 a-c** as well. The servers **302 a-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.) The “servers **302 a-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.)

163. 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 “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.)

164. Fontana thus discloses parallel processing of audio files across multiple servers, and segmenting data to facilitate parallel processing. A POSA would thus have found it obvious that to implement audio processing in parallel, 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 the combination of Kritt with Fontana to **segment[] the audio before processing the audio files in parallel.**

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

165. A POSA would have found claim 6 obvious over the combination of

Kritt with Fontana, as I discuss above. Likewise, for the reasons discussed for claims 5 and 7, a POSA would have found it obvious over the combination of Kritt with Fontana renders obvious that **the audio files and the image files are segmented**.

166. 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. The '954 patent, for example, gives the example that “the audio data can be processed and converted into a spectrum. Locations where the spectrum volatility is below a threshold can be detected and segmented. Such locations can represent silence or low audio activities in the audio data.” ('954, 7:11-15.) Kritt further discloses segmenting audio and image files **at spectrum thresholds** in order to, for example, segment the audio and image files based on scene or content changes, and separate spoken word audio segments from other segments such as silence. For example, Kritt discloses segmenting a video by creating shot files (“**the image files**”) that are segmented by comparing the histogram for consecutive frames, and segmenting the files into separate shots if the difference between the histograms exceeds a given threshold:

The video file 204 may be parsed into shot files according to any known method. For example, in one embodiment, a histogram may be

computed for each frame of the video file and the histograms for consecutive frames compared. If the histogram intersection of first and second consecutive frames is greater than a threshold, it may be inferred that the frames are similar, and consequently that the two frames are part of the same shot. On the other hand, if the histogram intersection of first and second consecutive frames is less than the threshold, it may be inferred that the two frames form a shot boundary.

(Kritt, ¶0040.) This comparison of the intersection between histograms representing video frames, and using a threshold to determine whether the frames belong to the same shot, **segments the image files at spectrum thresholds.**

167. Similarly, Kritt discloses segmenting **the audio files** by comparing one or more audio features, such as energy contour, volume contour, or frequency domain feature to distinguish, for example, speech segments from silent segments. A POSA would have understood and found it obvious that such segmenting would involve **spectrum thresholds**, such as a threshold energy or volume, or other information in a frequency domain (such as spectral centroid), that differentiates speech from silence. (*See, e.g., supra* claim 7.) Kritt explains:

Referring to FIG. 6 in operation 602, one or more audio features or audio signal descriptors may be extracted from an audio file 206. An audio feature may be a time domain feature, such as zero crossing rate, energy contour, volume contour, or fundamental frequency, or a

frequency domain feature, such as short term energy, bandwidth, entropy, spectral centroid, Mel-Frequency Cepstral Coefficients, or a Discreet Wavelet Transform. Many audio features are known in the art and any known audio feature or features that are suitable may be extracted in operation **602**.

In operation **604**, audio features or audio signal descriptors extracted from an audio file **206** may be classified. Each classification may be defined by a set of characteristic audio feature values. In one embodiment, audio features may be classified as silence, speech (spoken words), music, and a fourth category of other sounds that will be referred to herein as “sound effect.”

Segments of the video for which sound is not detectable may be classified as silent. In operation **605**, an audio tag with a silent type attribute may be associated with a silent audio feature, the tag having a time stamp that indicates the start and stop time of the silent period.

Segments of the video for which the audio feature values are similar to those that are characteristic of speech may be classified as speech. An audio tag with a speech type attribute may be associated with the audio feature, the tag having a time stamp of the period of speech. ...

(Kritt, ¶¶0056-0059.) This use of silence to segment audio aligns with the '954 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 data 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 audio activities in the audio data. The quiet periods in the audio data can be ignored, and the processing requirements thereof can be reduced.

(’954, 7:8-17.)

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

168. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. This combination **generates metadata**, including tags, tag files, and other metadata about a video. For example, Kritt teaches “various shot attributes may be determined and recorded as metadata. ... Metadata for a shot may include mood, emotion, or theme where another shot attribute is associated with a known cinematic technique for evoking the mood, emotion, or theme.” (Kritt, ¶0042.) A POSA would further appreciate that each of the tags generated by Kritt—including identification of shots and scenes; tags identifying human and nonhuman objects, audio objects, and emotions; and

probabilities associated with those tags—is **metadata**. (*See, e.g.*, claims 1[b], 1[f], 1[g], 1[h].)

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

169. A POSA would have found claim 11 obvious over the combination of Kritt with Fontana, as I discussed above. Furthermore, Kritt discloses including **timestamp references** among the generated **metadata**. For example, as noted for claim 11, each tag generated in Kritt is metadata. Kritt teaches that, for tags for identified objects and attributes, “[e]ach tag may be associated with a time stamp that indicates the start and stop time in which the object or attribute is rendered or otherwise associated.” (Kritt, ¶0032.) Furthermore, “[e]ach emotion object may be of a predefined type and associated with a time stamp.” (Kritt, ¶0034.) Likewise, audio tags are time stamped: “An audio tag with a speech type attribute may be associated with the audio feature, the tag having a time stamp of the period of speech. ... An audio tag with music type attribute may be associated with the audio feature, the tag having a time stamp of the period of music.” (Kritt, ¶0059.)

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

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

170. As I discuss above for Ground 1, a POSA would have found claim 1 obvious over the combination of Kritt with Fontana. Additionally, to the extent the combination of Kritt with Fontana does not render claim 2 obvious, a POSA would have found it obvious to combine Kritt and Fontana with Lau to **apply natural language processing to the text to determine context associated with the video.**

171. Like Fontana, Lau discloses using **natural language processing** to determine context associated with a video. For Lau, natural language processing is performed for purposes of matching the video to an advertisement. Lau teaches that “based on one or more keywords, ad units from the ad matrix are determined.” (Lau, ¶0038.) Lau additionally teaches that this process can build off 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 natural language processing engine.” (Lau, ¶0041.) A POSA would have appreciated that using recognition detection techniques, such as natural language processing, to recognize the content

is **applying natural language processing to the text to determine context associated with the video.**

172. **Rationale and Motivation to Combine (Kritt and Fontana with Lau):** It would have been obvious to combine Kritt and Fontana with Lau in order to use Lau’s disclosures for how to place advertisements within a video to carry out Fontana’s disclosures that advertisements can be linked to video content based on the results of the audio and image processing, and a POSA would have had a reasonable expectation of success in the combination. Like Kritt and Fontana, Lau is an analogous reference to the ’954 patent, addressed to the field of processing and generating multimedia content, including video, and specifically how to search within such multimedia content. (’954, 1:36-38 (“The present invention is generally directed to a method to generate data from video content, such as text and/or image-related information.”), *id.* at p.1 (Abstract) (“The text and the video content can be cross-referenced with the video.”); Lau, ¶0002 (Background) (“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.”); *see* claim 1[a] (analogous analysis for Kritt and Fontana).) Furthermore, both Fontana and Lau disclose linking advertisements to the multimedia content. (’954, 2:5-8 (“An advertisement and/or product or service recommendations can be placed at a specific time in the video based on the video content and/or section symbol of a video image.”); Fontana, ¶0117 (“FIG. 7M illustrates example advertisement data **716** that can be used in association with multimedia content, to link one or more advertisements with multimedia content during playback.”); Lau, ¶0002 (“Embodiments of the present invention generally relate to digital media and more specifically to displaying advertisements with rich media content.”).)

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

advertisement, and would have been motivated to further combine with an advertisement-matching method, such as that disclosed by Lau, in order to practice Fontana's disclosures of linking advertisement information with the multimedia content analyzed in Kritt and Fontana.

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

175. Likewise, a POSA would have appreciated that the audio and language processing algorithms taught in Lau—including, in particular, natural language processing—could be used in the context of Kritt with Fontana. Kritt discloses

performing its audio processing for speech-to-text recognition “using any known technique,” and a POSA would have appreciated that the natural language processing described in Lau was one such well-known technique. (Kritt, ¶0065.)

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

176. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana and Lau, as I discuss above. Additionally, for the reasons I explain for claim 2, the combination of Kritt with Fontana and Lau renders obvious **applying natural language processing to the text to extract a topic**. A POSA would have appreciated that Kritt’s key words, Fontana’s keywords, and Lau’s keywords are **topics** that could be extracted using **natural language processing**, as taught by Lau (*see supra* claim 2.).

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

177. A POSA would have found claim 1 obvious over the combination of Kritt with Fontana, as I discuss above. Additionally, as I noted in my discussion of the combination of Kritt with Fontana and Lau for claim 2, the combination of Kritt with Fontana and Lau **generat[es] an advertisement based on the text and the video data**. Specifically, the advertisement generated by Lau is created by

compiling ad units, each of which is selected by matching the ad units to the video based on the text (ie, the audio transcript and keywords) and the video data (ie, the metadata generated by Kritt with Fontana).

178. Lau discloses performing a search to determine advertisements whose content and/or topics match the content and/or topics for portions of video content. The search involves counting the number of keyword and/or concept matches for a search term near a particular time in the video content. As previously noted, a POSA would have found it obvious to combine Kritt and Fontana with Lau to use **the text and the video data** generated by Kritt with Fontana (e.g., the audio transcript, keywords, and metadata from audio and image processing) as keywords and concepts searched by Lau for purposes of identifying ads or ad units to match to the video based on context or subject matter.

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

keywords, categories, and concepts associated with the ad to information in the content.” (Lau, ¶0056.)

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

181. Lau For each ad, Lau 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....

(Lau, ¶0057.)

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

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

183. POSA would have found claim 9 obvious over the combination of Kritt with Fontana and Lau, as I discuss above. Additionally, the combination with Lau **plac[es] the advertisement in the video at a preferred time.**

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

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

186. 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.**

D. Ground 3: Claim 13 Is Obvious Over Kritt

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

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

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

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

The computer system **100** may contain one or more general-purpose

programmable central processing units (CPUs) **102A** and **102B**, herein generically referred to as the processor **102**. In an embodiment, the computer system **100** may contain multiple processors typical of a relatively large system; however, in another embodiment, the computer system **100** may alternatively be a single CPU system. Each processor **102** executes instructions stored in the memory **104** and may include one or more levels of on-board cache.

(Kritt, ¶0020.)

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

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

Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable

combination thereof. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination thereof. In the context of this disclosure, a computer readable storage medium may be any tangible medium that can contain, or store, a program for use by or in connection with an instruction execution system, apparatus, or device.

(Kritt, ¶¶0070-71.)

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

190. For the same reasons I explained in claim 1[c] in Ground 1, Kritt discloses **converting audio associated with the video to text.**

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

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

explains Kritt's disclosures around converting images associated with the video to video data, wherein the video data comprises objects previously identified in claim 1[d].

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

192. Kritt discloses **extracting topics from the video data** for the same reasons I explain in claim 1[e]. The topics extracted from the video data are shots/scenes, and emotions, moods, or themes.

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

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

195. Kritt additionally teaches analyzing the text to determine emotions. For example, Kritt discloses that “the objects, key words, and other data may be used to determine an emotion, mood, or theme for one or more of the scenes, and to generate corresponding emotion objects.” (Kritt, ¶0017.) The objects include audio objects, such as spoken words in the audio transcript (**the text**). (See, e.g., Kritt, ¶0016 (“A video may include spoken words, music, and other sounds, which may be referred to herein as audio objects.”); Kritt, ¶0065 (“An audio transcript may include all of the words spoken in the video.”).) Additionally, the key words are derived from **the text** (audio transcript). (See, e.g., Kritt, ¶0016 (“The spoken words may include ‘key words.’”); Kritt, ¶0017 (“In addition, key words that are spoken by human objects in the video are identified.”); Kritt, ¶0066 (“a key word may be determined from the audio transcript by counting the frequency of occurrences of words, the most frequently occurring verbs and nouns being determined to be key words.”).)

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

196. For the same reasons I discuss above in Ground 1 for claim 1[f], Kritt discloses and renders obvious **assigning probabilities of accuracy for each of the identified objects, wherein the probabilities are based on the extracted topics.**

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

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

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

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

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

199. For the reasons I explained in Ground 1 for claim 1[j], Kritt discloses

limitation 13[i].

V. NO SECONDARY CONSIDERATIONS OF NONOBVIOUSNESS

200. As explained in **Part III.B**, I understand that “secondary considerations” may be used as indicia of nonobviousness, provided that any such consideration has a nexus to the invention claimed. I understand from Petitioner’s counsel that Patent Owner in the related district court litigation has not yet identified any evidence related to secondary considerations with a nexus to the claimed invention. Nor am I aware of information, such as commercial success, unexplained results, long felt but unsolved need, industry acclaim, simultaneous invention, copying by others, skepticism by experts in the field, and failure of others, suggesting that the claims addressed in this Declaration are not obvious. To the extent Patent Owner later provides information it claims relates to secondary considerations, I reserve the right to supplement my analysis and opinions to comment on it.

VI. CONCLUSION

201. In signing this Declaration, I recognize that the Declaration will be filed as evidence in a contested case before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I also recognize that I may be subject to cross-examination in this proceeding. If required, I will appear for cross-

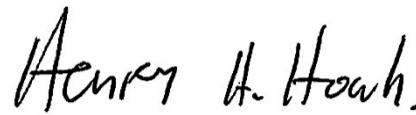
Declaration of Henry Houh, Ph.D. in Support of
Petition for *Inter Partes* Review of
U.S. Patent No. 10,218,954 B2

examination at the appropriate time. I reserve the right to offer opinions relevant to the invalidity of the challenged claims at issue and/or offer testimony in support of this Declaration.

202. I hereby declare that all statements made herein of my own knowledge are true and that all statements are made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001.

Dated: May 9, 2025

Respectfully submitted,

A handwritten signature in black ink that reads "Henry H. Houh." The signature is written in a cursive, slightly slanted style.

Henry Houh, Ph.D.
Lexington, MA

APPENDIX A

Henry H. Houh

15 Ballard Ter.
Lexington, MA 02420
henry.houh@gmail.com
cell: 781-325-8273

Education

Massachusetts Institute of Technology, Cambridge, MA

- PhD in Electrical Engineering and Computer Science, February 1998. “Designing Networks for Tomorrow's Traffic,” thesis supervised by Professor David Tennenhouse and Professor John Guttag. GPA 4.7/5.0
- Master of Science in Electrical Engineering and Computer Science, February 1991. “Demonstration of a laser repetition rate multiplier,” thesis. GPA 4.5/5.0
- Bachelor of Science in Electrical Engineering and Computer Science, June 1989. “Boundary element analysis of arbitrarily shaped dielectric structures,” thesis. GPA 4.7/5.0
- Bachelor of Science in Physics, February 1990. GPA 4.7/5.0

MITx, Cambridge, MA

- September 2020 – present: Enrolled in Statistics and Data Science MicroMasters program, a sequence of five MIT-offered graduate-level courses in statistics and machine learning, culminating in a MicroMasters certificate. Have completed four of five required courses.

Experience

Houh Consulting Inc. / Independent Consultant

- 2009 - present: Technical consultant specializing in Social Networking, Web 2.0, Web Site Development, Data Networking, Optical Networking, Telecommunications, Media Streaming and Voice Over IP. Clients include: Akin Gump Strauss Hauer & Feld LLP, Arnold & Porter Kaye Scholer LLP, BBN, Baker Botts LLP, Barnes & Thornburg LLP, BBN Technologies, Beus Gilbert McGroder PLLC, Bracewell LLP, Christensen O'Connor Johnson Kindness PLLC, Cooley LLP, Covington & Burling LLP, Cozen O'Connor P.C., DLA Piper, Davis Polk & Wardwell LLP, Desmarais LLP, Duane Morris LLP, Erise IP, P.A., Faegre Baker Daniels LLP, Finnegan, Henderson, Farabow, Garrett & Dunner, LLP, Fish & Richardson P.C., Goldman Ismail Tomaselli Brennan & Baum LLP, Haynes and Boone, LLP, Kaye Scholer, Kellogg Huber Hanson Todd Evans & Figel PLLC, Kirkland & Ellis LLP, Klarquist Sparkman, LLP, McCarthy Tetrault LLP, McGuireWoods LLP, Morgan, Lewis & Bockius LLP, Munger Tolles & Olson LLP, Orrick, Herrington & Sutcliffe LLP, O'Melveny & Myers LLP, Orrick, Herrington & Sutcliffe LLP, Paul Hastings LLP, Paul, Weiss, Rifkind, Wharton & Garrison LLP, Perkins Coie LLP, Quinn Emanuel Urquhart & Sullivan, LLP, Sidley Austin LLP, Smith Baluch LLP, Sterne, Kessler, Goldstein & Fox PLLC, Troutman Sanders LLP, Venable LLP, Weil, Gotshal &

Manges LLP, Wiley Rein LLP, Wilmer Cutler Pickering Hale and Dorr LLP, and Winston & Strawn LLP.

BlocksCAD Inc. (H3XL spin-out)

- 2014 - present: Founder and member of Board of Directors. BlocksCAD is a visual Computer-Aided Design (CAD) program, designed to teach children mathematics and geometric visualization. BlocksCAD is being used in U.S. schools and worldwide in maker spaces, with 350,000+ users. BlocksCAD was selected to be part of the LearnLaunch 2017 Boost Accelerator, the MIT Play Labs 2018 Summer Accelerator, and the MassChallenge 2018 Boston Cohort. Recipient of DARPA DSO grant in 2014, and SBIR grant from the USDA in 2019.

H3XL Inc. d/b/a Einstein's Workshop (formerly Lexington Robotics)

- 2009 - 2022: Founder and President. Started a science, technology, engineering and math enrichment program and creative/maker space, in 7,000 square feet of space. Serve 2,000+ kids and families annually. As of end of 2022, have delivered an estimated 250,000 student-instruction-hours of STEM courses. Principle Investigator for 2-year DARPA grant to improve 3D Computer Aided Design tool developed in-house for the purposes of teaching 8 year olds and up 3D CAD.

Eons

- 2008 - 2009: Chief Technology Officer. Created product that Eons acquired from BBN Technologies. Integrated BBN product with Eons social networking platform and significantly increased the Eons group creation rate. Helped evaluate advertisement platform offerings and rolled out the "Boom Network" advertising network. Eons raised \$32 million from General Catalyst Partners, Charles River Ventures, Sequoia Capital, and Intel Capital.

BBN Technologies

- 2007 - 2008: Delta Division, Vice President of Technology. Grew "Boomerang" counter-sniper project engineering team and significantly de-risked \$10 million worth of product deliveries. Identified sales team leading to \$10 million dollars of new sales within 6 months and \$100+ million in additional orders in the following two years. Boomerang was a significant asset leading to the acquisition of BBN by Raytheon in 2009. Created new business plan and grew team; launched new fully-featured social networking web site in 5 months. Served as lead expert witness in patent infringement lawsuit, resulting in \$58 million jury award to client; verdict for patents I testified on were upheld on appeal which resulted in a \$120 million settlement.
- 2004 - 2007: Delta Division, Director of Technology, responsible for commercializing IP and creating new businesses. Hired and grew division's initial engineering team. Wrote three business plans, two of which were funded. For call center business plan, acted as general manager, hiring and managing engineering team, inside sales team, and identifying and recruiting a new general manager. Identified and recruited other key employees to Delta Division, including senior members of team leading to successful internal sales growth and spin-outs of projects. Contracted by BBN to BBN spin-out PodZinger as VP of Operations and Technology. Additional areas studied include mesh networking, network security, satellite communications, and semantic web.

PodZinger Inc. (BBN spin-out, also known as EveryZing and now RAMP)

- 2006: Vice President of Operations and Technology. Significantly upgraded capability of consumer-facing search site and redeployed web site from company to co-location facility. Identified key portions of infrastructure for upgrading and cost reduction. Helped write business plan, evaluating advertisement models of revenue. Hired in operations replacement and phased back to BBN.

Commonwealth Capital

- 2004: Entrepreneur-in-residence (informally), performed technical due diligence on business plans, brainstormed ideas for new businesses with venture partner. With venture partner, left for portfolio company BBN to form core of commercialization team.

Empirix, Inc./Teradyne, Inc.

- 2001 - 2004: Chief Technologist, Engineering Manager, Web Application Test Group. Researched potential new product areas; developed product plan and prototype. Responsible for three new and existing products. Managed off-shore development team. Chief architect for all web testing products. Re-architected core testing product, helped write javascript interpreter. Provided technical vision for core product.
- 2000 - 2001: Chief Technologist, Communication Infrastructure Test Group. Responsible for incorporating new technology internally, tracking new technologies, technical evaluation of partnerships and potential acquisitions. Helped develop division strategy. Developed plans which formed core capabilities for successful new products introduced in 2004-5.
- 2000 - 2001: Engineering Manager, Communications Infrastructure Test Group. Execution of new product plan developed in prior role. Grew team from four existing engineers to team of over 30 on immediate team and over 40 on project. Delivered new platform in one year. Platform and derivatives accounted for large portion of booked products for the division within 2 years and is currently (2008) a key portion of new product offerings.
- 2000: Empirix was formed as a spin-out of Teradyne in January 2000. Reported to CEO in carve-out of Empirix from Teradyne.
- 1999: (Teradyne) Director of Business Development, Software Test Units. Reported directly to Chairman of the Board/Founder and then to general manager of software test unit (6 divisions of Teradyne). Evaluated and researched acquisition and partnership candidates. Internally assessed technology position in market and gaps in product lines. Worked with senior division staff to develop new product strategies. Attended internal Teradyne divisional board meetings. Chairman served as my mentor.

3Com Corporation/NBX Corporation

- 1999: Software Engineer 5. Continued work after 3Com acquisition of NBX. Built cross-division relationships for new products and research directions. NBX was acquired by 3Com in March 1999.
- 1997 - 1999: (NBX) Senior Scientist and Engineer. Work in IP Telephony. Architected next-generation product. Protocol design and validation for core protocol now used tens of millions of times daily. Led team in integration of IP protocols into current product. Designed audio reconstruction algorithms. Developed applications for bug analysis and diagnosis of system

problems. Implementation of network simulator. Work on collaborative projects with external partners. Worked to identify gaps in product. Representative at numerous trade shows. Innovated novel methods of using product.

MIT Department of Electrical Engineering and Computer Science, Cambridge, MA

- 1991 - 1998: Research assistant, Telemedia Network Systems Group. Design, development, and implementation of Gigabit ATM network for distributed multimedia system. Studied host interface design issues. ATM network simulation.
- Spring 1989, Fall 1990, Spring 1995: Teaching assistant, Computation Structures digital systems course. (Spring 1995 Head TA)
- 1988 - 1989: Head laboratory teaching assistant for Computation Structures. Responsibilities included writing and revising lab assignments, and maintaining the lab.
- 1987: Laboratory teaching assistant for Computation Structures.
- 1987: Design, construction, and programming of 16-bit computer.

Agora Technology Group, Incorporated

- 1994 - 1996: Founder and CEO. Conceived and oversaw development of targeted advertising-supported Web sites. Responsible for company's vision and direction. Attended the first two WWW Conferences, presenting a workshop and paper at the first, and appearing on the "Commercialization and Economics of the Web" panel and chairing the "Where Commercial Services and the Web Are Headed" panel at the second. Sold company intact; is currently an operating stand-alone company.

AT&T Bell Laboratories, Holmdel, NJ

- 1989 - 1990: Implementation of cascadable all-optical fiber logic gate. Modelocking of all-fiber erbium laser. Construction of modelocked laser repetition rate booster. Strong optics laboratory and fiber optic experience.
- Summer 1988, 1987: Research in integrated optics. Analysis of rectangular waveguides using microwave modeling. Fabrication of integrated optical components.

Honors

- MIT Alumni Association Great Dome Award, 2010, Baker 60th Anniversary Reunion Co-Chair (highest group award given by MIT Alumni Association)
- MIT Alumni Association Presidential Citation Award (now known as Great Dome), 2008, Member of MIT Chairman's Salon committee
- MIT Alumni Association Bronze Beaver Award, 2007 (highest individual award given by MIT Alumni Association)
- MIT Alumni Association Volunteer Honor Roll, February 2004
- MIT Alumni Association Lobdell Award, 1999
- Boston Museum of Science Gold Pin for 1000 hours of Volunteer Service, April 1999
- MIT Alumni Association Presidential Citation Award (now known as Great Dome), 1997, Member of Alumni Online Communications Committee

Patents and Patent Publications

- U.S. Patent #9,697,231, H. Houh, J. Stern, R. Spina, M. Meteer, “Methods and apparatus for providing virtual media channels based on media search,” July 4, 2017.
- U.S. Patent #9,697,230, H. Houh, J. Stern, “Methods and apparatus for dynamic presentation of advertising, factual, and informational content using enhanced metadata in search-driven media applications,” July 4, 2017. See also WO2007056485.
- U.S. Patent #7,975,296, L. Apfelbaum, H. Houh, T. Mayberry and G. Friedman, “Automated security threat testing of web pages,” July 5, 2011. See also US20030159063, WO2003067405.
- U.S. Patent #7,877,736, H. Houh and J. N. Stern, “Computer language interpretation and optimization for server testing,” January 25, 2011. See also US20050138104, WO2005043300.
- U.S. Patent #7,801,910, H. Houh and J. N. Stern, “Method and apparatus for timed tagging of media content,” September 21, 2010. See also US20070112837, US20090222442, WO2007056535.
- U.S. Patent #7,590,542, D. C. Williams, W. C. Hand, H. Houh, A. R. Seeley, “Method of Generating Test Scripts Using a Voice-Capable Markup Language,” September 15, 2009. See also EP1530869, US20030212561, WO2003096663.
- U.S. Patent #6,967,963, H. H. Houh, P. Anderson, C. Gadda, “Telecommunication method for ensuring on-time delivery of packets containing time-sensitive data,” November 22, 2005. See also EP1060400, WO2000033092, CA2318774.
- U.S. Patent #5,144,375, M. C. Gabriel, H. H. Houh, N. A. Whitaker, “Sagnac Optical Logic Gate,” September 1, 1992. Also issued as European Patent # EP0456422, July 23, 1997, German Patent #DE69126913, August 28, 1997
- U.S. Patent Application Publication No. 20020015387, “Voice Traffic Packet Capture and Analysis Tool for a Data Network” (Abandoned in 2007)
- U.S. Patent Application Publication No. 20020016708, “Method and Apparatus for Utilizing a Network Processor as Part of a Test System” (Abandoned in 2007)
- U.S. Patent Application Publication No. 20020016937, “Method and Apparatus for Utilizing a Network Processor as Part of a Test System.” (Abandoned in 2007) See also WO2002011413.
- U.S. Patent Application Publication No. 20070106646, “User-directed navigation of multimedia search results” (Abandoned in 2009)
- U.S. Patent Application Publication No. 20070106660, “Method and apparatus for using confidence scores of enhanced metadata in search-driven media applications” (Abandoned in 2009)
- U.S. Patent Application Publication No. 20070106685, “Method and apparatus for updating speech recognition databases and reindexing audio and video content using the same” (Abandoned in 2015)
- U.S. Patent Application Publication No. 20070118873, “Methods and apparatus for merging media content” (Abandoned in 2015)
- U.S. Patent Application Publication No. 20090222442, “User-directed navigation of multimedia search results” (Abandoned in 2016)
- U.S. Patent Application 11/395,732, “Search snippet creation for audio and video data” (Abandoned in 2009)
- U.S. Patent Application 11/774,931, “Methods and apparatus for managing a social networking web site”
- U.S. Patent Application 11/774,947, “Methods and apparatus for organizing media files”
- U.S. Patent Application 11/774,956, “Methods and apparatus for managing an online event”

- U.S. Provisional Patent Application 61/086,909, “Measuring and ranking relationship activity”
- U.S. Provisional Patent Application 61/086,914, “Detecting media object commonality”
- U.S. Provisional Patent Application 61/086,904, “Message categorization based on message characteristics”
- U.S. Provisional Patent Application 61/086,905, “Photo tagging to request action”

Trials, Testimony and Depositions

- Case No. 1:06CV682 (CMH/BRP), Verizon vs. Vonage, U.S. District Court for the Eastern District of Virginia, filed expert report, was deposed and testified at trial.
- Case No. 1:08CV157 (CMH/TRJ), Verizon vs. Cox, U.S. District Court for the Eastern District of Virginia, was deposed and testified at trial.
- Case No. 5:09-cv-476, Two-Way Media vs. AT&T, U.S. District Court for the Western District of Texas, filed expert report, testified at trial.
- Case No. 2:10-cv-248 (RAJ/FBS), ActiveVideo Networks vs. Verizon, U.S. District Court for the Eastern District of Virginia, filed expert report and was deposed as an expert witness.
- Case No. 1:11-cv-00880-TSE-JFA, Bear Creek Technologies, Inc. vs. Verizon Services Corp., et al, U.S. District Court for the Eastern District of Virginia, was deposed as an expert witness.
- Case No. 3:10-CV-298-BBC, AlmondNet, Inc. vs. Microsoft Corp., U.S. District Court for the Western District of Wisconsin, filed expert report.
- Case No. 6:10-cv-00597, Guardian Media Technologies, Ltd. Vs. AT&T Operations, Inc. et al., U.S. District Court for the Eastern District of Texas, Tyler Division, filed expert report.
- Case No. ESCV2010-02282C, The Octopus Solution LLC v. Gary Brown et al., Essex, MA Superior Court, testified at trial.
- Investigation No. 337-TA-882, In the matter of Certain digital media devices, including televisions, Blu-ray disc players, home theater systems, tablets and mobile phones, components thereof and associated software, U.S. International Trade Commission, filed expert reports, was deposed and testified at hearing.
- Investigation No. 337-TA-995, In the matter of Certain communications or computing devices, and components thereof, U.S. International Trade Commission, filed expert reports, and was deposed.
- Case No. 8:12-cv-122-LES-TDT, Prism Technologies LLC v. AT&T Mobility LLC, U.S. District Court for the District of Nebraska, filed expert report, was deposed and testified at trial.
- Case No. 11-2684 (JWL), Sprint Communications Co., L.P., v. Comcast Cable Communications, LLC, et al., U.S. District Court for the District of Kansas, filed expert report and was deposed.
- Case No. 11-2686 (JWL), Sprint Communications Co., L.P., v. Time Warner Cable, Inc., et al., LLC, et al., U.S. District Court for the District of Kansas, filed expert report, was deposed, and testified at trial.
- Case No. 12-487 (SLR), Cox Communications, Inc., et al., v. Sprint Communications Company L.P., et al., U.S. District Court for the District of Delaware, filed expert report and was deposed.
- Case IPR 2014-00039, Microsoft Corporation v. B.E. Technology LLC, U.S. Patent No. 6,628,314. Submitted declaration and cross-examined in deposition.
- Case IPR 2014-00086, Apple, Inc. v. Evolutionary Intelligence, LLC, U.S. Patent No. 7,010,536. Submitted declaration and cross-examined in deposition.
- Case No. 6:11-CV-421, Stragent, LLC v. Intel Corporation, U.S. District Court for the Eastern District of Texas, Tyler Division, deposed and testified at trial as a fact witness.

- IPR2014-01366, IPR2014-01367, IPR2014-01368, Samsung et al. v. Straight Path, submitted IPR declarations and was deposed.
- IPR2015-1006, IPR2015-1007, Cisco v. Straight Path, submitted IPR declarations.
- IPR2014-01457, IPR2014-01459, Microsoft v. Biscotti, submitted IPR declaration and was deposed.
- IPR2016-00302, Apple v. Nonend, submitted IPR declaration.
- IPR2014-00812, Yelp and Twitter v. Evolutionary Intelligence, submitted IPR declaration.
- IPR2015-00307, Cisco v. AIP, submitted IPR declaration.
- Microsoft v. Acacia, submitted IPR declaration.
- IPR2015-00342, AT&T Mobility LLC and Cellco Partnership d/b/a Verizon Wireless v. Solocron, submitted IPR declaration.
- Mediatube v. Bell Canada, Canadian Federal Court, submitted expert report and testified at trial.
- CBM2016-00036, Google v. At Home Bondholder's Liquidating Trust, submitted CBM declaration.
- IPR2016-01198, IPR2016-01201, Apple, Inc. v. VoIP-Pal.com, Inc., submitted IPR declaration and was deposed.
- Case No. 3:14-cv-2824-VC, Trend Micro Incorporated v. Rpost Holdings, Inc., et al, submitted expert report and was deposed
- Akamai v. Limelight Networks, IPR2016-01631, IPR2016-01711, IPR2017-00249, IPR2017-00349, submitted IPR declarations and was deposed
- Hyperlync Technologies v. Verizon Sourcing, No. 650151/2015 New York Supreme Court, submitted declaration and was deposed
- Cascades Streaming Technologies v. Big Ten Networks, submitted expert reports and was deposed
- Klaustech, Inc. v. Google LLC., Case No. 4:10-CV-05899-JSW (N.D. Cal.), submitted expert report and was deposed
- Cisco v. Uniloc, IPR2017-00198, IPR2017-00058, submitted IPR declarations and was deposed.
- Cisco v. Uniloc, IPR2017-00597, submitted IPR declaration and was deposed
- Twitter v. Youtoo Technologies, LLC, IPR2017-01133, IPR2017-01131, submitted IPR declarations and was deposed
- Taser International Inc. (now Axon Enterprise, Inc.) v. Digital Ally Inc., IPR2017-00375, submitted IPR declaration and was deposed
- Taser International Inc. (now Axon Enterprise, Inc.) v. Digital Ally Inc., IPR2017-00376, IPR2017-00515, IPR2017-00775, submitted IPR declarations
- HTC Corporation v. IPA Technologies, IPR2018-00306, IPR2018-00307, submitted IPR declaration
- Hulu v. Soundview, IPR2018-00864, IPR2018-01023, submitted IPR declarations and was deposed
- XpertUniverse v. Cisco, 3:17-cv-3848, U.S. District Court for Northern District of California, submitted declaration in support of claim construction and was deposed
- Improved Search v. Microsoft, C.A. No. 16-cv-650-JFB-SRF (Del.), submitted declaration in support of claim construction
- Apple v. Uniloc, IPR2018-00361, IPR2018-00394, IPR2018-00395, IPR2019-00700, IPR2019-00701, IPR2019-01667, submitted IPR declarations
- LG Electronics v. Uniloc, submitted IPR declaration
- SRC Labs and Saint Regis Mohawk Tribe v. Microsoft, 2:18-cv-00321-JLR, U.S. District Court for Western District of Washington at Seattle, submitted declaration in support of claim construction and was deposed

- Verizon Wireless v. Barkan Wireless IP Holdings LP, submitted IPR declarations
- Verizon Wireless and Samsung Electronics America Inc. v. Barkan Wireless IP Holdings LP, IPR2019-199, IPR2019-200, IPR2019-234, IPR2019-631, IPR2019-632, submitted IPR declarations
- Samsung Electronics America Inc. v. Barkan Wireless IP Holdings LP, IPR2019-100, submitted IPR declaration
- Cisco Systems Inc. v. Meetrix IP LLC, IPR2019-539, IPR2019-540, IPR2019-541, IPR2019-542, IPR2019-543, IPR2019-544, submitted IPR declarations
- GMG Products LLC v. Traeger Pellet Grills LLC, PGR2019-24, PGR2019-34, PGR2019-35, PGR2019-36, submitted PGR declaration and was deposed
- Uniloc 2017 LLC v. Google LLC, 2:18-cv-00502-JRG-RSP, U.S. District Court for the Eastern District of Texas, Marshall Division, submitted declaration in support of claim construction
- Microsoft Corporation v. Uniloc 2017 LLC, IPR2020-00102, IPR2020-00103, submitted IPR declarations
- Microsoft Corporation and HP Inc. v. Syncloud Technologies, LLC, IPR2020-00316, IPR2020-01031, IPR2020-01032, IPR2020-01269, IPR2020-01270, IPR2020-01271, submitted IPR declarations
- Seven Networks, LLC v. Apple Inc, 2:19-cv-115-JRG, U.S. District Court for the Eastern District of Texas, Marshall Division, submitted declaration in support of claim construction
- Unified Patents Inc. v. Ortiz & Associates Consulting, LLC, IPR2019-00743, submitted IPR declaration
- Uniloc 2017 LLC v. Apple Inc., 3:19-cv-01905-JD, U.S. District Court for the Northern District of California, San Francisco Division, submitted declaration in support of claim construction and was deposed
- ClearPlay Inc. v. Dish Network LLC et al.; 2:14-cv-00191, U.S. District Court for the District of Utah, Central Division, submitted expert report and re-examination declarations, was deposed, and testified at trial
- AMC Networks, Inc. et al v. Sound View Innovations, LLC, IPR2020-00482, submitted IPR declaration
- Case No. 1:17-cv-01734-RGA, Sprint Communications Co., L.P., v. Charter Communications, Inc., et al., U.S. District Court for the District of Delaware, filed expert report and was deposed.
- CommScope Technologies LLC et al. v. Barkan Wireless IP Holdings LP et al., IPR2020-00827, IPR2020-00829, IPR2020-00831, IPR2020-00833, IPR2020-00835, IPR2020-00838, submitted IPR declarations
- Cisco Systems Inc. v. Monarch Networking Solutions LLC, IPR2020-01226, submitted IPR declaration
- CoolTV Network.com v. Blackboard, Inc., Facebook, Inc., International Business Machines Corporation, Kaltura, Inc., Limelight Networks, Inc., Microsoft Corporation, Ooyala, Inc., Snap, Inc., and Trapelo Corp., submitted declaration in support of claim construction and was deposed
- Facebook, Inc. v. Onstream Media Corporation, IPR2020-01507, IPR2020-01508, IPR2020-01525, IPR2020-01527, IPR2020-01528, submitted IPR declarations
- Hisense Co LTD. v. Maxell, LTD, IPR2020-01598, IPR2020-01600, submitted IPR declarations
- Microsoft Corp. v. O'Brien, IPR2021-00015, submitted IPR declaration and was deposed
- Sable Networks Inc. v. Cisco Systems, Inc., WA:20-CV-00288-ADA, U.S. District Court for the Western District of Texas, Waco Division, submitted declaration in support of claim construction
- Peloton Interactive, Inc. v. Icon Health and Fitness Inc et al., IPR2021-00342, submitted IPR declaration

- Hulu, LLC v. SITO Mobile R&D IP, LLC et al., CBM2020-00028, IPR2020-00158, IPR2020-00206 , IPR2020-00219 , IPR2020-00265 , IPR2020-00298 , IPR2020-00304 , IPR2020-00308 , submitted CMB and IPR declarations and was deposed
- Twitter, Inc. and Google LLC v. B.E. Technology, LLC, IPR2021-00482, IPR2021-00483, IPR2021-00484, IPR2021-00485, submitted IPR declarations
- MG Freesites LTD v. ScorpCast, LLC, IPR2021-00511, submitted IPR declaration
- Avionte, LLC v. Sandeep Acharya, Samar Basnet, and Aqore LLC, Arbitration, Hon. Jeffrey Keyes (Ret.)
- Microsoft Corporation v. WSOU Investments, LLC d/b/a Brazos Licensing and Development, IPR2021-00751, submitted IPR declaration
- SAP America, Inc. v. Express Mobile, Inc., IPR2021-01144, IPR2021-01145, IPR2021-01146, submitted IPR declarations
- StratosAudio Inc. v. Hyundai Motor America, 6:20-cv-01125-ADA, U.S. District Court for the Western District of Texas, Waco Division, submitted claim construction declaration and was deposed
- Peloton Interactive, Inc. v. Ifit Health and Fitness Inc, IPR2022-00029, IPR2022-00030, submitted IPR declarations and was deposed
- Investigation No. 337-TA-1263, In the matter of Certain televisions, remote controls, and components thereof, U.S. International Trade Commission, filed expert reports, was deposed, and testified at hearing
- Alexander Stross v. Netease, Inc., et al., U.S. District Court (Central District of California), Case 2:20-cv-00861-AB(JC), filed expert reports, and was deposed
- Ralph Ledergerber v. Netease, Inc., et al., U.S. District Court (Central District of California), Case 2:20-cv-00862-AB(JC), filed expert reports and was deposed
- John Walmsley v. Netease, Inc., et al., U.S. District Court (Central District of California), Case 2:20-cv-00863-AB(JC), filed expert reports and was deposed
- Bernhard Kuhmstedt v. Netease, Inc., et al., U.S. District Court (Central District of California), Case 2:20-cv-02044-AB(JC), filed expert reports and was deposed
- Netflix, Inc. v. CA, Inc., IPR2021-01319, submitted IPR declaration and was deposed
- F45 Training Pty Ltd. v. Body Fit Training USA Inc., C.A. No. 20-1194 (LPS) (D. Del.), filed expert report
- Apple v. Fintiv, IPR2022-00976, IPR2022-01149, IPR2022-01150, IPR2023-00399, submitted IPR declarations and was deposed
- Estech Systems, Inc. v. Howard Midstream Energy Partners d/b/a Howard Energy Partners, C.A. No. 6:20-cv-777-ADA (Western District of Texas, Waco Division), filed expert reports and was deposed
- Echelon Fitness Multimedia, LLC. v. Peloton Interactive, Inc., IPR2020-01187, submitted IPR declaration and was deposed
- VIZIO. V. Maxell Ltd., IPR2022-01458, submitted IPR declaration
- Maxell, Ltd. et el. v. VIZIO, Inc., Case No. 2-21-cv-06758 (C.D. Cal.). filed declaration and was deposed
- Microsoft Corporation v. SurfCast, Inc., submitted IPR declarations
- The Walt Disney Company et al. v. WAG Acquisition LLC, IPR2022-01227, IPR2022-01228, IPR2022-01346, submitted IPR declarations and was deposed
- Rakuten Rewards et al. v. IBM, submitted IPR declarations and was deposed
- Fubotv Media Inc. v. SITO Mobile R&D IP, LLC et al., submitted IPR declarations
- Unified Patents, LLC v. Noblewood IP LLC, IPR2022-01111, submitted IPR declaration
- Unified Patents, LLC. v. Pedersen, Peter Henrik, IPR2023-00029, submitted IPR declaration
- Ericsson Inc v. Dali Wireless Inc., submitted IPR declarations

- Universal Electronics, Inc. et al. v. Roku, Inc., IPR2022-00818, submitted declarations and was deposed
- Google LLC v. Wildseed Mobile LLC et al., IPR2023-00244, IPR2023-00245, IPR2023-00246, IPR2023-00247, IPR2023-00248, submitted declarations
- Lululemon Athletica Canada Inc. et al. v. Nike, Inc., IPR2023-00180, IPR2023-00424, submitted declarations
- Zoho Corporation, et al. v. Meetrix IP, LLC, IPR2023-00371, IPR2023-00377, IPR2023-00379, IPR2023-00380, IPR2023-00382, submitted declarations
- WSOU Investments, LLC d/b/a/ Brazos Licensing and Development v. Google LLC, 6:20-cv-572, U.S. District Court for the Western District of Texas, Waco Division, submitted reports and was deposed
- Google LLC v. Safecast Limited LLC, IPR2023-00652, submitted declaration
- Investigation No. 337-TA-1338, In the matter of Certain smart televisions, U.S. International Trade Commission, filed reports and was deposed
- Cisco Systems, Inc. v. Orckit IP LLC, IPR2023-00714, submitted declaration
- SupplyPro, Inc. v. Sandvik Machining Solutions AB, et al., AAA Case No. 01-22-000-7897. submitted declaration
- TrackThings LLC v. NETGEAR, Inc., Case No. 22-00981-RGA-JLH, U.S. District Court for the District of Delaware, submitted declaration
- Correct Transmission, LLC v. Nokia of America Corporation, Case No. 2:22-CV-00343-JRG-RSP, U.S. District Court for the Eastern District of Texas, Marshall Division, submitted export reports, was deposed, and testified at trial
- Walt Disney Parks & Resorts US Inc. v. Agile Journeys LLC; IPR2024-00029, submitted declaration
- Aylo Freesites Ltd v. Dish Technologies LLC; IPR2024-00043, IPR2024-00044, IPR2024-00045, IPR2024-00046, IPR2024-00047, IPR2024-00048, IPR2024-00146, IPR2024-00147; submitted declarations and was deposed
- Arm Inc. et al. v. Icpillar LLC; IPR2024-00476, IPR2024-00566; Retained by Arm; submitted declarations
- Microsoft Corp et al. v. Litl LLC, IPR2024-00454, IPR2024-00455, IPR2024-00456, IPR2024-00457, IPR2024-00458; submitted declarations
- Cisco Systems Inc et al. v. InfoExpress Inc, IPR2024-00539, IPR2024-00540, IPR2024-00677, IPR2024-00678; submitted declarations and was deposed
- Vizio Inc v. Multimedia Technologies Pte Ltd; IPR2024-00722, IPR2024-00723; Retained by Vizio; submitted declarations and was deposed
- Sap America Inc v. Isix IP LLC; IPR2024-00615; Retained by SAP; submitted declarations
- Cisco Systems Inc et al. v. Portsmouth Network Corp; IPR2024-00505, IPR2024-00506; submitted declarations and was deposed
- Aylo Freesites Ltd v. Dish Technologies LLC; IPR2024-00512, IPR2024-00513, IPR2024-00514, IPR2024-00515, IPR2024-00516, IPR2024-00517, IPR2024-00518, IPR2024-00519, IPR2024-00940, IPR2024-00941; submitted declarations and was deposed
- Aylo Freesites Ltd v. WellcomeMat LLC; IPR2024-00710, IPR2024-01101; submitted declarations and was deposed
- Verizon Corporate Service Group et al. v. Headwater Partners I LLC; IPR2024-00809, IPR2024-00945; Retained by Verizon; Technology: Mobile device management; Submitted declarations
- Microchip Technology Inc v. Aptiv Technologies Ltd et al.; IPR2024-00228; Retained by Aptiv; Technology: Universal Serial Bus; submitted declaration and was deposed

- Vusion Group SA v. Hanshow Technology Co Ltd; IPR2024-00963; Retained by VusionGroup; Technology: Video analysis; submitted declaration
- Maxell, Ltd. v. TCL Electronic Holdings Ltd. et al.; Case No. 5:23-cv-00108-RWS-JBB; U.S. District Court for the Eastern District of Texas, Texarkana Division; Retained by TCL; submitted declaration and was deposed
- Cerence Operating Company v. Samsung Electronics Co., LTD. et al., Case No. 2:23-cv-00482-JRG-RSP, U.S. District Court for the Eastern District of Texas, Marshall Division, submitted export report on claim construction and was deposed
- MasterClass and FuboTV v. Dish Technologies LLC; IPR2024-00917, IPR2024-00918, IPR2024-00919; submitted declarations and was deposed
- Microsoft Corp v. Dialect LLC; IPR2025-00655, IPR2025-00656, IPR2025-00657, IPR2025-00658, IPR2025-00659; submitted declarations

Publications

- “IP switching: server driven flow classification,” H. H. Houh, *Proceedings of the Washington University Workshop on Integration of IP and ATM*, November 1996.
- “Aurora at MIT,” D. D. Clark, H. H. Houh, and D. L. Tennenhouse, Editors, *MIT Laboratory for Computer Science Technical Report 673*, December 1995.
- “ViewStation Applications: Implications for Network Traffic,” C. J. Lindblad, D. Wetherall, W. Stasior, J. F. Adam, H. H. Houh, M. Ismert, D. Bacher, B. Phillips, and D. L. Tennenhouse, *IEEE Journal of Selected Areas in Communications*, 1995.
- “The VuNet Desk Area Network: Architecture, Implementation, and Experience,” H. H. Houh, J. F. Adam, M. Ismert, C. J. Lindblad, and D. L. Tennenhouse, *IEEE Journal of Selected Areas in Communications*, 13 (4), May, 1995.
- “Reducing the Complexity of ATM Host Interfaces,” H. H. Houh and D. L. Tennenhouse, Hot Interconnects II Symposium Proceedings, Stanford, August 11-12, 1994.
- “Media-intensive data communications in a ‘desk-area’ network,” J. F. Adam, H. H. Houh, M. Ismert, and D. L. Tennenhouse, *IEEE Communications*, August 1994.
- “ViewStation Applications: Intelligent Video Processing Over A Broadband Local Area Network,” C. J. Lindblad, D. J. Wetherall, W. Stasior, B. Phillips, D. Bacher, J. Adam, H. Houh, M. Ismert, and D. L. Tennenhouse, Proceedings of the 1994 USENIX Symposium on High-Speed Networking, Oakland, CA, August 1994.
- “The Media Gateway: Live Video on the World Wide Web,” H. H. Houh, C. J. Lindblad, J. Soo, D. L. Tennenhouse, and D. J. Wetherall, Workshop at the 1994 World Wide Web Conference, Geneva, Switzerland, May 1994.
- “Active Pages: Intelligent Nodes on the World Wide Web,” H. H. Houh, C. J. Lindblad, and D. J. Wetherall, Proceedings of the 1994 World Wide Web Conference, Geneva, Switzerland, May 1994.
- “Wavelength Division vs. Code Division Access Methods for Optical Networks,” H. H. Houh, Area Exam Paper, May 1993.
- “Experience with the VuNet: A Network Architecture for a Distributed Multimedia System,” J. F. Adam, H. H. Houh, D. L. Tennenhouse, Proceedings of the 18th Conference on Local Computer Networks, pp. 70-76, September 1993
- “The VudBoard: A Simple DMA Interface,” H. H. Houh, Proceedings of the 4th Gigabit Minijam, January 1994.

- “A Software-Oriented Approach to the Design of Media Processing Environments,” D. L. Tennenhouse, J. Adam, D. Carver, H. Houh, M. Ismert, C. Lindblad, W. Stasior, D. Weatherall, D. Bacher, and T. Chang., submitted to the International Conference on Multimedia Computing and Systems, May 1994.
- “A Network Architecture for Distributed Multimedia Systems,” J. F. Adam, H. H. Houh, M. Ismert, and D. L. Tennenhouse, submitted to the International Conference on Multimedia Computing and Systems, May 1994.
- “The Viewstation Collected Papers,” D. L. Tennenhouse, J. Adam, C. Compton, A. Duda, D. Gifford, H. Houh, M. Ismert, C. Lindblad, W. Stasior, R. Weiss, D. Wetherall, D. Bacher, D. Carver, and T. Chang, MIT Laboratory for Computer Science Technical Report, MIT/LCS/TR-590, November 1993.
- “The Viewstation Collected Papers II,” W. F. Stasior, D. L. Tennenhouse, J. F. Adam, D. R. Bacher, H. H. Houh, M. Ismert, C. J. Lindblad, B. M. Phillips, D. J. Wetherall, MIT Laboratory for Computer Science Technical Report, MIT/LCS/TR-696, April 1996.
- “A System's Perspective of the Sagnac Fiber Logic Gates and Their Possible Applications,” A. Huang, N. Whitaker, C. Gabriel, H. Avramopoulos, P. M. W. French, H. H. Houh, and I. Chuang, Applied Optics, September 10, 1994
- “Complete Switching in a Three-Terminal Sagnac Switch,” H. Avramopoulos, P. M. W. French, M. C. Gabriel, H. H. Houh, N. A. Whitaker, T. Morse, IEEE Phot. Tech. Lett. **3** (3), 235
- “Complete Switching in a Three-Terminal Sagnac Switch,” H. Avramopoulos, P. M. W. French, M. C. Gabriel, H. H. Houh, N. A. Whitaker, IEEE/LEOS Annual Meeting, Paper PDP-13, November 1990
- “All-optical phase-locked oscillator,” N. A. Whitaker, Jr., H. H. Houh, H. Avramopoulos, T. F. Morse, IEEE/LEOS Annual Meeting, Paper ELT2.4/MOO3, November 1990
- “Passive modelocking of an all-fiber erbium laser,” H. Avramopoulos, H. H. Houh, N. A. Whitaker, M. C. Gabriel, T. F. Morse, IEEE/LEOS Conference on Optical Amplifiers and their Applications, Paper PDP-8, August 1990
- “Transverse modes, waveguide dispersion, and 30ps recovery in submicron GaAs/AlAs microresonators,” J. L. Jewell, S. L. McCall, A. Scherer, H. H. Houh, N. A. Whitaker, A. C. Gossard, and J. H. English, Appl. Phys. Lett. **55** (1), July 3, 1989

Leadership, Activities and Interests

- Leadership
 - Reading Symphony Orchestra
 - Board of Directors (Treasurer), 2022 – present
 - Discovery Museum (Acton, MA)
 - Board of Directors, 2023 – present
 - Board of Directors, Development Committee, 2023 – present
 - Science and Technology Advisory Council, 2012 – present
 - The Discover Museum received the 2024 National Medal for Museum and Library Service, the nation’s highest honor given to museums and libraries that make significant and exceptional contributions to their communities.
 - Tau Beta Pi National Engineering Honor Society
 - Executive Councilor and Secretary, 2024 – 2026
 - Advisor, MA B Chapter at MIT, 2003 – present
 - District Director (National Officer), Tau Beta Pi, New England Area, 1991 – 2003

- President, MA B Chapter at MIT, Fall 1988 – Spring 1989
- Laureate award, 1989
- MIT Alumni Association Board of Directors
 - MIT Annual Giving Board, 3 year term, 2024 – 2027
 - National Selection Committee, 3 year term, 2020 – 2023
 - K-12 STEM Initiatives Co-chair, 2013 – 2017
 - Awards Committee Chair, 2012 – 2014
 - Awards Committee, 3 year term, 2011 – 2014
 - Vice President, 2 year term, 2004 – 2006
 - Alumni Ad-Hoc Committee, Chair, 2005-2006
 - Board Member, 2 year term, 1997 – 1999
- MIT Club of Boston
 - Board of Directors, 2006 – 2011
 - K-12 Initiatives Chair, 2009 – 2012
 - VP of Communications, MIT Club of Boston, 2003 – 2006
 - Past-President, MIT Club of Boston, 2002 – 2003
 - President, MIT Club of Boston, 2001 – 2002
 - President-Elect, MIT Club of Boston, 2000 – 2001
 - VP of Programs, MIT Club of Boston, 1999 – 2000
 - Activities Super-Chair, MIT Club of Boston, 1998 – 1999
- MIT Enterprise Forum of Cambridge, Inc.
 - Past Chair, 2009 – 2011
 - In-NOW-vation Co-chair, 2010
 - Chair, 2007 – 2009
 - Vice Chair, 2005 – 2007
 - Executive Board Member, 2002 – 2011
 - Winter Workshop Co-Chair, February 2007 – conceived idea for conference, which sold-out and produced largest attendance numbers in recent memory
 - Spring Workshop Co-Chair, Spring 2004
 - Membership Committee Chair, Fall 2003 - 2006
 - 25th Anniversary Dinner Chair, Fall 2003
 - As Membership Chair and Board Member, started Special Interest Groups in 2004; a SIG won the MIT Presidential Citation award, the MIT Alumni Association's highest award for organizations, in 2006
- Estabrook Elementary School PTA
 - Advisory committee to the superintendent on PCB issue, 2010-2011
 - 4th Grade after-school science program co-organizer, 2010-2012
 - 4th and 5th Grade before-school Math Olympiad co-organizer, 2009-2013
 - 5th Grade BBQ and Yearbook Committee, 2011, 2013
 - Family Math Night volunteer, 2008-2012
- MIT Class of 1989
 - 35-year Reunion Gift Committee, Co-chair, 2024
 - Secretary, five consecutive 5 year terms, 1989 - 2014
 - 30-year Reunion Committee, 2019
 - 25-year Reunion Committee, 2014
 - 20-year Reunion Committee and Gift Committee, 2009
 - 15-year Reunion Committee and Gift Committee, 2004
 - 10-year Reunion Committee and Gift Committee, 1999

- 5-year Reunion Committee, 1994
 - Interim Treasurer, 1993 - 1994
 - Instituted annual senior class career fair, now raising over \$100,000 annually for senior class activities, Fall 1988
 - Strong, consistent record of leadership dating to high school
- Acting
 - '21' (Sony Pictures), credited as "Chinatown Dealer," 2007, Kevin Spacey's movie about the MIT Blackjack Team inspired by "Bringing Down the House" by Ben Mezrich, opened nationwide on March 28, 2008. 21 was the number one movie in US for two weeks and number one globally for one week. 21 also topped the DVD sales, Blu-ray sales and DVD rental charts.
 - Spring Lake Theater Company, first New York-area off-broadway production of "A Chorus Line," played role of Mark, Summer 1990
- Former member of the MIT Blackjack team
- Producer for 10,000 Maniacs' 2013 album "Music from the Motion Picture"
- Executive Producer for 10,000 Maniacs' 2015 album "Twice Told Tales"
- Violist, violist, harpist, guitarist, singer, actor: played in many amateur/semi-professional groups including Firebird Pops Orchestra, Reading Symphony Orchestra, MIT Summer Philharmonic Orchestra, Carlisle Chamber Orchestra, Rivers Symphony Orchestra, Opera51 Orchestra, Freisinger Chamber Orchestra, Merrimack Valley Philharmonic, Longwood Symphony, MIT Symphony, and Somerville Community Chorus
- Violist, "Coming Home" album by Karen Phillips, 2014
- Google Scholar citations: 2224, h-index: 22, i10-index: 27 (since 2019: 539/12/12)
- Erdos-Bacon number of 7
- Erdos-Bacon-Sabbath number of 11